ENHANCING UNCERTAINTY ESTIMATION IN SEMANTIC SEGMENTATION VIA MONTE-CARLO FREQUENCY DROPOUT

Tal Zeevi[1], *Lawrence H. Staib*[1,2,3,*], *John A. Onofrey*[1,2,4,*]

Departments of [1]Biomedical Engineering, [2]Radiology & Biomedical Imaging, [3]Electrical Engineering, [4]Urology, Yale University, New Haven, CT, USA

ABSTRACT

Monte-Carlo (MC) Dropout provides a practical solution for estimating predictive distributions in deterministic neural networks. Traditional dropout, applied within the signal space, may fail to account for frequency-related noise common in medical imaging, leading to biased predictive estimates. A novel approach extends Dropout to the frequency domain, allowing stochastic attenuation of signal frequencies during inference. This creates diverse global textural variations in feature maps while preserving structural integrity – a factor we hypothesize and empirically show is contributing to accurately estimating uncertainties in semantic segmentation. We evaluated traditional MC-Dropout and the MC-frequency Dropout in three segmentation tasks involving different imaging modalities: (i) prostate zones in biparametric MRI, (ii) liver tumors in contrast-enhanced CT, and (iii) lungs in chest X-ray scans. Our results show that MC-Frequency Dropout improves calibration, convergence, and semantic uncertainty, thereby improving prediction scrutiny, boundary delineation, and has the potential to enhance medical decision-making.

Index Terms— Segmentation Uncertainty, Monte-Carlo Dropout, Frequency Dropout, Selective Prediction

1. INTRODUCTION

Estimating prediction uncertainties in deterministic deep learning models often involves the strategic introduction of controlled artificial noise into the data [1]. This can occur either before [2, 3] or during [4, 5, 6] neural network processing, with subsequent measurement of variations in model performance to assess robustness. Techniques such as DropConnect [7] and Dropout [8], which randomly omit network edges or nodes during processing, have been foundational in this respect, effectively injecting random patterns of noise into the network's operation allowing the simulation of a predictive distribution approximating Bayesian inference [9].

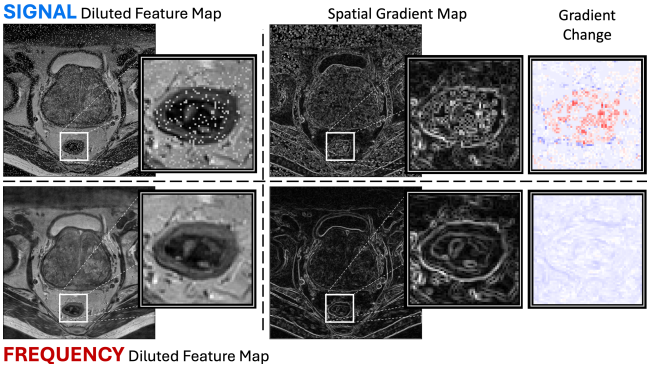

Fig. 1: Impact of Signal and Frequency Dropout on Structural Integrity. Left subplots show feature maps with traditional (Signal) Dropout (p = 0.1, top) and Frequency Dropout (p = 0.1, bottom). The middle subplots display Sobel gradient maps, while the right subplots illustrate structural changes by comparing pre- and post-Dropout gradients (red indicates intensified gradients, suggesting the emergence of new structures, and blue reflects diminished gradients, revealing blurring and disruption of existing structures). Signal Dropout introduces non-uniform changes in the gradient map, creating new edges while disrupting existing ones. In contrast, Frequency Dropout maintains more uniform changes across imaging features, preserving spatial dependencies. For example, gradient changes follow edges, rather than appearing irregular at the voxel level.

In convolutional neural network (CNN) layers, commonly used in segmentation tasks, each convolution step corresponds to a node on the network graph, essentially turning Dropout into a random source of impulse noise within the CNN feature maps. This method, however, may not comprehensively capture the predictive distribution in medical imaging, where noise extends into the frequency domain – a range poorly addressed by impulse noise. Our recent findings [10] suggest that Frequency Dropout [11], which randomly removes frequency components from feature maps during Monte Carlo (MC) simulations, refines predictive uncertainty estimates in medical imaging classification over traditional Dropout.

In semantic segmentation, preserving precise structural integrity is essential, sometimes down to the pixel level. Impulse noise introduced by traditional Dropout can significantly impact spatial features, causing non-uniform and

* These authors jointly supervised this work.

abrupt changes in the gradient field. This can disrupt fine edges, distort subtle imaging markers, or introduce high variability in otherwise uniform regions, as exemplified in Fig. 1. Traditional Dropout operates independently on individual elements of the feature map, without accounting for dependencies across spatial features. In contrast, Frequency Dropout's attenuations in the frequency domain generate a global noise effect on the feature maps, which may better preserve structural relationships and dependencies between spatial features. This approach may more accurately capture uncertainty, particularly in regions where preserving feature correlations, such as object boundaries, is essential.

In this study, we explore MC-Frequency Dropout for semantic segmentation to assess its impact on uncertainty estimation. Our contributions are: (i) defining Signal and Frequency Dropout within CNN layers, highlighting their theoretical foundations and effects on structural integrity; (ii) comparing uncertainty estimates across diverse modalities, including X-ray, CT, and MRI; and (iii) examining how the placement of dropout layers within state-of-the-art medical imaging networks influences segmentation performance, demonstrating the benefits of strategic dropout positioning.

2. BACKGROUND

Without loss of generality, let $C_\theta : \mathbb{R}^{m\times n} \to \mathbb{R}^{m'\times n'}$ denote a convolution block within a CNN model M_Θ, comprising a single convolutional layer, where $\theta \in \Theta$ represents the set of trainable model parameters. The forward pass output of C_θ for a given input instance $X \in \mathbb{R}^{m\times n}$ is expressed as follows:

$$C_\theta(X) = \sigma(X * W + b)$$

where $W^{k\times k} \in \theta$ represents the convolution kernel, $b \in \theta$ denotes the bias term, $*$ signifies the convolution operation between X and W, and $\sigma(\cdot)$ denotes an element-wise nonlinearity (e.g., rectified linear unit activation function).

2.1. Signal Diluted Forward Pass

In a signal-diluted forward pass [8], the convolution kernel is multiplied by a binary mask in each convolution step:

$$(X * W)(i,j) = \sum_{u=1}^{k}\sum_{v=1}^{k} X(i+u-\lfloor k/2\rfloor, j+v-\lfloor k/2\rfloor)\cdot W(u,v)\cdot D_{ij}(u,v).$$

In this equation, i, j index the output feature map, u, v index the kernel, and $D_{ij}^{k\times k}$ acts as a binary dropout mask, where each element $D_{ij}(u,v)$ is independently sampled from a Bernoulli distribution with probability p.

Dropout is the operation of selectively disabling nodes from the network graph during the forward-pass. In convolutional layers, this process involves setting all kernel weights to zero during a particular convolution step, denoted by $D_{ij}(u,v) = D_{ij}(u',v'),\ \forall u,v,u',v'$. This action separates the dropout process from the convolution operation itself. As a result, the output from a CNN block that includes Signal Dropout, labeled as C_θ^D, is expressed as:

$$C_\theta^D(X) = \sigma([X * W] \odot D + b).$$

Here, $D^{m'\times n'} = [d(i,j)]$ acts as a binary dropout mask and the operator $\odot$ represents element-wise multiplication.

Applying the mask D via element-wise multiplication involves $m'n'$ operations, resulting in a computational complexity of $O(m'n')$ for Signal Dropout.

2.2. Frequency Diluted Forward Pass

The frequency-diluted forward pass [10] removes signal frequencies within feature maps. Instead of applying the dropout mask D directly to the signal, we transform the signal to the Fourier space, apply the mask to remove frequency components, and reconstruct the signal using the inverse Fourier transform. The output of a forward pass through a CNN block with frequency dilution, denoted as $C_\theta^{\mathcal{F}}$, is expressed as:

$$C_\theta^{\mathcal{F}}(X) = \sigma(\mathcal{F}^{-1}(\mathcal{F}[X * W] \odot D) + b).$$

Here, $\mathcal{F}$ and $\mathcal{F}^{-1}$ represent the Fourier transform and its inverse, respectively.

Using the Fast Fourier Transform (FFT), each transformation has complexity of $O(N \log N)$, where $N = m'n'$ is the number of elements in the feature map. Thus, the overall computational complexity of Frequency Dropout is $O(m'n' \log(m'n') + m'n')$, accounting for FFT operations and element-wise multiplication with the dropout mask.

3. EMPIRICAL EVALUATION

We compared uncertainty estimates from Monte Carlo (MC) simulations with Signal and Frequency Dropout during inference to identify segmentation errors across three public semantic segmentation tasks involving MRI, CT, and X-ray modalities. The CT and MRI datasets were sourced from the Medical Segmentation Decathlon [12]. The source code for our proposed method is publicly available[1].

3.1. Segmentation Datasets

Prostate zones on biparametric MRI (bpMRI) scans: 36 transverse T2-weighted and apparent diffusion coefficient (ADC) MRI scans of the prostate, each annotated to delineate two adjoining prostate zones: the peripheral zone (PZ) and the transitional zone (TZ) [12].

[1] https://github.com/talze/frequency-dropout.git

Liver and tumors on contrast enhanced CT scans: 37 contrast-enhanced liver CT scans with liver and tumor annotations from patients with metastatic liver disease. [12].

Lungs on Chest X-ray scans: 40 annotated chest X-rays from the National Library of Medicine's digital image database for Tuberculosis (TB) [13, 14], featuring TB and normal cases, with the lung area behind the heart excluded.

3.2. Segmentation Models

For the prostate MRI and liver CT segmentation tasks, we used the top-performing pre-trained models available in the Medical Segmentation Decathlon [12]. In both tasks, these models were developed using the nnU-Net approach, which implements a self-configuring segmentation pipeline based on the U-Net architecture [15]. For the lung X-ray task, we utilized the MedSAM transformer model [16], adapted from the Segment Anything (SAM) model [17]. This model is designed for broad medical image segmentation and is applicable to a range of modalities, including X-ray.

3.3. Dropout Variants in Segmentation Networks

We devised three model dropout variants, each with distinct dropout layer placements: (i) Encoder Dilution: Dropout layers are introduced after each encoder step, with no dilution in the decoder section. (ii) Decoder Dilution: Dropout layers are introduced after every decoding step, except for the output step, with no dilution in the encoder section. (iii) Global Dilution: Dropout layers are introduced after every encoding or decoding step, except for the output step.

3.4. Monte-Carlo (MC) Dropout Simulation

For each input instance, we performed $R = 30$ forward-pass repetitions, each with a new random diluted realization of the pre-trained U-Net or SAM models, using two dropout strategies: (i) Signal Dropout and (ii) Frequency Dropout. Voxel-level prediction estimates were obtained from the arithmetic mean of SoftMax values across Monte Carlo repetitions, with predictive uncertainty estimated from their standard deviation [9]. This procedure was repeated across various dropout rates (p=0.01, 0.02, 0.04, 0.08, 0.16, and 0.32) to evaluate the sensitivity of the dropout approaches to changes in dropout rate.

3.5. Evaluation Metrics

Expected Uncertainty Calibration Error (UCE): Uncertainty estimates can be used to reject uncertain predictions, acting as scores to classify predictions into binary outcomes: reject or do-not-reject, a process known as selective prediction [18, 19]. We evaluated the alignment of MC-Dropout uncertainty estimates with voxel-level segmentation errors from the full (no-dropout) model using UCE [20]. UCE, similar to Expected Calibration Error (ECE) [21], measures the calibration of uncertainty scores rather than predicted probabilities.

Dice Similarity Coefficient (DSC): Diluting neural network models can impact performance. We measured the deviation in segmentation accuracy between the diluted models and the full (no-dropout) model using the Dice Similarity Coefficient [12].
Both UCE and DSC were calculated per instance and averaged across the cohort to assess overall model performance.

4. RESULTS

Across all segmentation tasks, the best-performing Frequency Dropout configuration converged to better-calibrated uncertainty estimates than Signal Dropout, with uncertainties more closely matching full model segmentation errors (Fig. 2). Similar performance was observed for liver tumor segmentation. Frequency Dropout achieved higher or comparable DSC scores relative to Signal Dropout and the full (no-dropout) model, demonstrating stable segmentation performance across tasks (Tab. 1).

Performance varied across different dropout layer placements (Fig. 3). In liver and most prostate segmentation tasks, Signal Dropout performed better in the encoder and global settings, while Frequency Dropout excelled in the decoder setting. For liver tumors, higher dropout rates were more effective with encoder dilution, while lower rates worked better with decoder dilution. In lung segmentation, Frequency Dropout maintained stable performance across all placements. Lower dropout rates generally improved the calibration of uncertainty estimates, though not always optimally, while higher rates introduced variability depending on placement but excelled in certain tasks. Specifically, higher rates improved liver segmentation, while lower rates were preferable for lung segmentation. In prostate segmentation, lower rates showed minimal variation, whereas higher rates led to significant variability in both dropout methods.

Visual examples of uncertainty maps for MC-Signal and Frequency Dropout are shown in (Fig. 4).

5. DISCUSSION AND CONCLUSION

This paper explores two Monte Carlo (MC) Dropout approaches for medical imaging segmentation: traditional (Signal) Dropout, applied directly to feature maps, and Frequency Dropout, operating in the frequency domain. MC simulations generated uncertainty estimates to identify segmentation errors in tasks challenging state-of-the-art models. Our results show that Frequency Dropout produced better-calibrated uncertainty estimates (Fig. 2) while maintaining stable segmentation performance compared to Signal Dropout (Tab. 1).

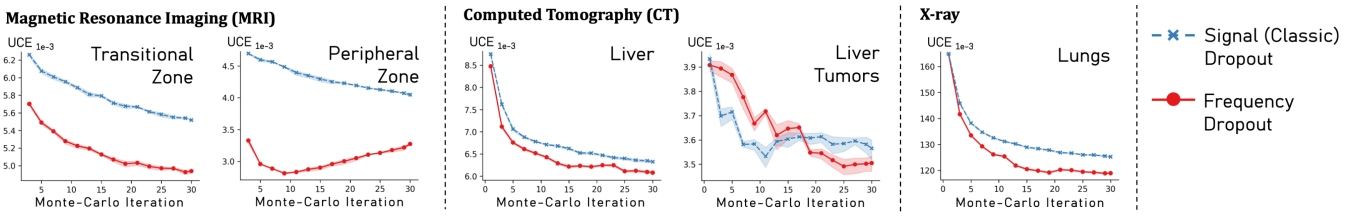

Fig. 2: Calibration of uncertainty estimates for the best dropout configurations (rates and layer placements) of frequency (red) and signal dropout (blue). Subplots show the discrepancy between MC-dropout uncertainties and full (no-dropout) model errors. Frequency dropout aligns uncertainty estimates more accurately across tasks, while preserving segmentation performance similar to the full model (Table 1).

Table 1: Impact of dilution on segmentation performance, reflected by DSC divergence between the dropout and full models

Task	**Modality**	**DSC at baseline**	% **Divergence from baseline DSC after** R **Monte-Carlo iterations (sd)**			
			R**=5**		R**=30**	
		No Dropout	**Signal** Dropout	**Frequency** Dropout	**Signal** Dropout	**Frequency** Dropout
Liver	CT	0.915	0.24 (0.31)	0.40 (0.41)	0.26 (0.36)	0.39 (0.38)
Liver Tumors	CT	0.606	*-14.53 (1.12)	1.23 (1.05)	*-11.72 (1.03)	*3.54 (0.99)
Lungs	X-ray	0.831	0.44 (0.14)	*0.91 (0.12)	0.63 (0.11)	*1.05 (0.10)
Transitional Zone	MRI	0.781	-0.05 (0.59)	-0.11 (0.50)	-0.03 (0.57)	-0.14 (0.42)
Peripheral Zone	MRI	0.627	-0.03 (0.58)	-0.98 (0.46)	-0.12 (0.58)	-1.02 (0.52)

* Statistically significantly different from the baseline model ($p < 0.05$, Bonferroni corrected).

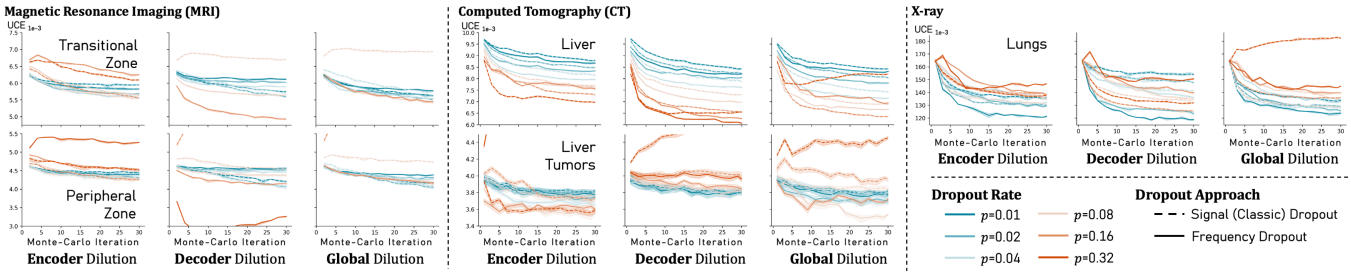

Fig. 3: Calibration of MC-dropout uncertainty estimates with model errors, measured by Expected Calibration Error (ECE) (↓) for different dropout configurations, including dropout rates and layer placements: Encoder Dilution, Decoder Dilution, and Global Dilution.

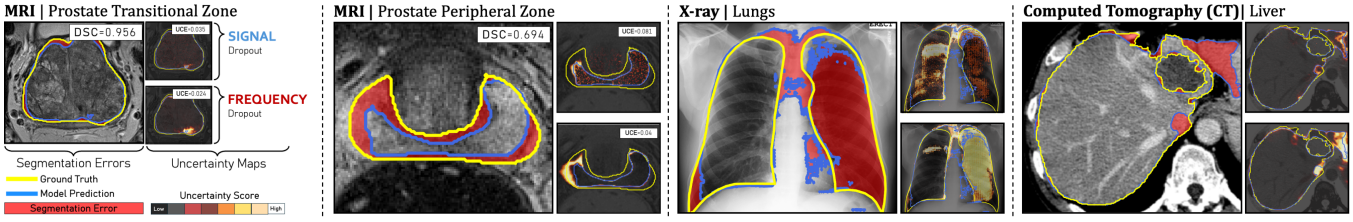

Fig. 4: MC-dropout uncertainty maps for segmentation tasks using Signal and Frequency Dropout. Large subplots show regions of ground truth (yellow line), full model prediction (blue line), and segmentation errors (red region). Small subplots present uncertainty maps: top for Signal Dropout, bottom for Frequency Dropout, with uncertainty measured as the standard deviation of voxel-level predictions across MC repetitions. Each plot uses the optimal dropout configuration for its approach.

Semantic segmentation requires processing spatial information while preserving dependencies between related elements. Traditional Dropout, applied independently to feature map elements, can disrupt local correlations. In contrast, Frequency Dropout operates globally, preserving structural coherence and generating multiple variations of structural information, improving structural uncertainty estimation.

Traditional Dropout's impulse noise effect varies with feature map intensity: near-zero intensities create subtler effects, while broader ranges amplify noise. Frequency Dropout provides consistent variation, unaffected by intensity scaling.

While both Signal and Frequency Dropout run efficiently on modern hardware, Frequency Dropout's extra complexity could slow inference for large feature maps. This trade-off

between processing time and enhanced uncertainty estimation should be weighed to determines suitability for specific tasks.

MC-Dropout has limitations; our study found that uncertainty estimates depend on layer placement, dropout rates, and MC repetitions (Fig. 3). Optimal parameters vary across modalities and tasks, reflecting the method's task-dependent nature. Despite these challenges, MC-Dropout remains popular for its simplicity. Optimizing strategies in state-of-the-art architectures like U-Net could provide valuable insights.

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by [12, 13, 14].

7. ACKNOWLEDGMENTS

No funding was received to conduct this study. The authors have no relevant financial or non-financial interests to disclose.

8. REFERENCES

[1] Ling Huang, Su Ruan, Yucheng Xing, and Mengling Feng, "A review of uncertainty quantification in medical image analysis: probabilistic and non-probabilistic methods," *Medical Image Analysis*, p. 103223, 2024.

[2] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, 2019.

[3] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren, "Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*. Springer, 2019, pp. 61–72.

[4] Davood Karimi, Qi Zeng, Prateek Mathur, Apeksha Avinash, Sara Mahdavi, Ingrid Spadinger, Purang Abolmaesumi, and Septimiu E Salcudean, "Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images," *Medical image analysis*, vol. 57, pp. 186–196, 2019.

[5] Yongkai Liu, Guang Yang, Melina Hosseiny, Afshin Azadikhah, Sohrab Afshari Mirak, Qi Miao, Steven S Raman, and Kyunghyun Sung, "Exploring uncertainty measures in bayesian deep attentive neural networks for prostate zonal segmentation," *Ieee Access*, vol. 8, pp. 151817–151828, 2020.

[6] Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Medical image analysis*, vol. 59, pp. 101557, 2020.

[7] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus, "Regularization of neural networks using dropconnect," in *International conference on machine learning*. PMLR, 2013, pp. 1058–1066.

[8] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[9] Yarin Gal and Zoubin Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[10] Tal Zeevi, Rajesh Venkataraman, Lawrence H Staib, and John A Onofrey, "Monte-carlo frequency dropout for predictive uncertainty estimation in deep learning," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024, pp. 1–5.

[11] Salman H Khan, Munawar Hayat, and Fatih Porikli, "Regularization of deep neural networks with spectral dropout," *Neural Networks*, vol. 110, pp. 82–90, 2019.

[12] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al., "The medical segmentation decathlon," *Nature communications*, vol. 13, no. 1, pp. 4128, 2022.

[13] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 577–590, 2013.

[14] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al., "Automatic tuberculosis screening using chest radiographs," *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 233–245, 2013.

[15] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[16] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, pp. 654, 2024.

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al., "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[18] Jishnu Mukhoti and Yarin Gal, "Evaluating bayesian deep learning methods for semantic segmentation," *arXiv preprint arXiv:1811.12709*, 2018.

[19] Aryan Mobiny, Pengyu Yuan, Supratik K Moulik, Naveen Garg, Carol C Wu, and Hien Van Nguyen, "Dropconnect is effective in modeling uncertainty of bayesian deep networks," *Scientific reports*, vol. 11, no. 1, pp. 5458, 2021.

[20] Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier, "Well-calibrated model uncertainty with temperature scaling for dropout variational inference," *arXiv preprint arXiv:1909.13550*, 2019.

[21] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.