Weight Entropy-Maximised Evidential Metamodel for Post Hoc Uncertainty Quantification

Gouranga Bala, Dhruvi Ganatra, Amit Sethi

Indian Institute of Technology Bombay, Mumbai 400076, India
gouranga.bala23@gmail.com, 21d070027@iitb.ac.in, asethi@iitb.ac.in

Abstract

Reliable uncertainty quantification (UQ) is crucial for deploying deep learning models in safety-critical domains such as medical imaging. Existing post hoc UQ methods either rely on multi-pass inference or suffer from limited expressiveness due to their dependence on final-layer embeddings. In this work, we propose evidential meta model, a lightweight post-hoc framework that enhances Dirichlet evidential modeling by extracting features from multiple layers of a frozen classifier. This multilayer strategy enriches the metamodel input with both low-level textures and high-level semantics, enabling more accurate modeling of aleatoric and epistemic uncertainty. To further boost epistemic fidelity, we incorporate *Max-WEnt* regularization, which maximizes the entropy of learnable scaling weights applied within the meta-model. This promotes internal hypothesis diversity without modifying the base network or incurring test-time overhead.

Across seven benchmarks including medical datasets (BACH, DIV2K, HAM10000, BreakHIS) and natural image tasks (SVHN, Fashion-MNIST, ImageNet-C) our evidential metamodel consistently improves AUROC and calibration over both the base model and prior post-hoc UQ methods. Ablation studies confirm the complementary benefits of multilayer features and Max-WEnt. Our approach offers a robust and efficient solution for trustworthy AI in clinical and other high stakes settings.

Introduction

Deep learning has achieved remarkable success in visual recognition tasks, particularly in medical image analysis, where neural networks now rival or exceed expert-level performance. However, despite these advances, deep classifiers often fail to indicate when their predictions may be unreliable, a limitation that becomes critical in safety-critical settings such as clinical diagnosis. In such contexts, knowing *when* a model is uncertain is as important as knowing *what* it predicts.

Uncertainty quantification(UQ) techniques aim to address this challenge by producing confidence-aware predictions. Among them, post-hoc UQ methods are especially attractive for practical deployment because they preserve performance of the original classifier and require minimal computational overhead. These methods often include model ensembles (Lakshminarayanan, Pritzel, and Blundell 2017), Monte Carlo dropout (Gal and Ghahramani 2016), and temperature scaling (Guo et al. 2017). More recent techniques employ evidential deep learning (EDL) frameworks (Sensoy, Kaplan, and Kandemir 2018) that output Dirichlet distributions over class probabilities. Yet, a common limitation persists: most post hoc UQ methods rely solely on the final-layer representation of the base model, which may omit critical intermediate cues.

To address this, we introduce a lightweight evidential meta-model, trained to estimate uncertainty from multi-layer features of a frozen base classifier. By extracting representations from three distinct layers (i.e. early, middle, and deep) our method captures both low-level textures and high-level semantic cues. This fusion equips the metamodel with a richer hypothesis space, enhancing its ability to detect uncertainty in out-of-distribution (OOD) or ambiguous samples. Crucially, the base classifier remains unchanged, ensuring that classification accuracy is not compromised.

To further improve epistemic expressiveness, we adapt Max-WEnt regularization (de Mathelin et al. 2025), a recent strategy that encourages diversity in internal model hypotheses by maximizing the entropy of learnable scaling weights. Applied to the meta-model, Max-WEnt promotes a broader and more robust uncertainty landscape, enabling stronger performance under distribution shifts.

We evaluate our evidential metamodel across seven challenging benchmarks, spanning both medical and natural image datasets. Results show that our approach consistently outperforms previous post hoc UQ methods, including prior work of post hoc uncertainty quantification in medical imaging, BAY-MED (Bala, Chauhan, and Sethi 2025), in both AUROC and calibration error. Ablation studies confirm that multilayer feature fusion and Max-WEnt regularization are both indispensable to this improvement.

Our contributions are as follows:

- We propose a multi-layer evidential meta-model that leverages diverse representations from a frozen classifier to estimate uncertainty more reliably.
- We integrate Max-WEnt regularization into post hoc UQ to enhance hypothesis diversity and epistemic calibration.

- We demonstrate significant improvements in AUROC across seven benchmarks, showing strong generalization to both medical and vision datasets.

Related Work

Uncertainty quantification (UQ) has emerged as a cornerstone of reliable AI systems, particularly in domains where overconfident false predictions can lead to critical failures such as medical imaging, autonomous driving, and scientific discovery. The broader UQ literature can be categorized into post hoc estimation methods, intrinsic Bayesian models, and ensemble-based techniques, each offering different trade-offs in accuracy, calibration, interpretability, and computational cost.

Post-Hoc Estimation Methods

Post-hoc UQ approaches are especially appealing for their compatibility with pretrained deterministic classifiers. These methods retain predictive capability of the original model while appending a secondary mechanism for uncertainty estimation. Monte Carlo Dropout (MC Dropout) (Gal and Ghahramani 2016) introduces stochasticity during inference by enabling dropout at test time, thereby approximating Bayesian model averaging. Deep Ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) train multiple independent networks and aggregate their predictions to capture epistemic uncertainty. Temperature scaling (Guo et al. 2017) applies a single scalar parameter to soften the output logits, improving confidence calibration without changing the decision boundaries of the classifier.

More recent techniques seek to improve UQ quality via auxiliary prediction heads or metamodels. Evidential Deep Learning (EDL) (Sensoy, Kaplan, and Kandemir 2018) reformulates softmax classification into a Dirichlet distribution prediction task, enabling both predictive mean and uncertainty estimation. Direct Epistemic Uncertainty Prediction (DEUP) (Jain et al. 2021) introduces a separate model trained to predict generalization error directly from data embeddings. Shen et al. (Shen et al. 2023) extend this with evidential meta-models that learn to produce Dirichlet parameters over logits of a frozen base model. However, these approaches often rely on shallow representations or introduce limited representational diversity, restricting their generalization under distribution shift.

Bayesian and Variational Methods

Bayesian Neural Networks (BNNs) (Blundell et al. 2015) and variational inference models (Neal 2012) offer a principled framework for modeling uncertainty through posterior distributions over network weights. While theoretically appealing, BNNs suffer from high computational complexity and poor scalability in large-scale or real-time scenarios. To address this, techniques such as low-rank variational approximations, stochastic gradient MCMC, and sparse Bayesian priors have been proposed. Evidential methods like EDL avoid full posterior inference by estimating belief mass over categorical distributions, but may still require strong priors and sensitive regularization.

Ensemble-Based Methods

Ensemble approaches remain one of the most effective empirical strategies for UQ, consistently demonstrating strong calibration and robustness under domain shift. However, training and storing multiple models poses scalability challenges. Lightweight variants such as snapshot ensembles and self-distilled ensembles offer partial remedies by reducing training redundancy while maintaining prediction diversity. Nevertheless, ensemble methods still incur significantly higher training and inference costs compared to single-model techniques.

Uncertainty Quantification in Medical Imaging

In medical imaging, UQ is essential for responsible deployment. Numerous methods target this domain specifically, such as test-time augmentation (TTA) for estimating prediction variance (et al. 2023), uncertainty-aware segmentation masks for interpretability (Jungo and Reyes 2020), and Bayesian approximations tailored for volumetric imaging (Kwon et al. 2020). While these methods often yield impressive results, many are domain-specific and computationally expensive.

Meta-modeling for UQ in medical imaging has gained traction due to its simplicity and flexibility. Shen et al. (Shen et al. 2023) proposed an evidential meta-model trained on base classifier embeddings to predict Dirichlet distributions, facilitating epistemic and aleatoric decomposition. DEUP (Jain et al. 2021) generalized this approach to predict uncertainty under low-data and OOD regimes. Earlier work(Bala, Chauhan, and Sethi 2025) in medical imaging proposed lightweight Dirichlet meta-models for post-hoc uncertainty estimation, yet these approaches lacked architectural enhancements designed to capture diverse feature representations.

Our Contributions

In this work we extend our prior work with two key innovations. First, we enrich the meta-model's input space by aggregating features from three intermediate layers of a frozen base classifier, thereby combining early texture information with high-level semantics. This multi-layer design enhances the meta-model's sensitivity to both aleatoric and epistemic cues. Second, we introduce a novel regularization strategy *Max-WEnt* which maximizes the entropy of learnable weight scalings to encourage internal diversity without additional test-time cost. Inspired by recent advances in entropy-based regularization (de Mathelin et al. 2025), Max-WEnt allows the meta-model to explore a wider hypothesis space.

Together, these contributions lead to consistent improvements across seven benchmarks spanning medical and vision datasets. Our results confirm that multilayer evidential modeling and weight entropy maximization are both critical for high-quality, computationally efficient post-hoc uncertainty estimation.

Methodology

Dirichlet Evidential Meta-Model

Our metamodel is built on the evidential deep learning framework of Sensoy et al.(Sensoy, Kaplan, and Kandemir 2018) and the post hoc uncertainty quantification approach of Shen et al.(Shen et al. 2023). Our goal is to efficiently estimate predictive uncertainty by training a meta-model that outputs a Dirichlet distribution over class probabilities, without modifying the base classifier or relying on Monte Carlo sampling.Given a pre-trained backbone classifier f_θ, we extract the penultimate embedding $\mathbf{h} = f_\theta^{\text{penult}}(\mathbf{x}) \in \mathbf{R}^C$ for each input $\mathbf{x}$. This vector captures high-level, class-discriminative features. We pass $\mathbf{h}$ through a lightweight metamodel g_ϕ, implemented as a two-layer MLP with `ReLU` activations, which outputs an evidence vector $\mathbf{e} = g_\phi(\mathbf{h})$. A `soft-plus` activation ensures that all components of $\mathbf{e}$ are non-negative, and we obtain the Dirichlet concentration parameters as $\boldsymbol{\alpha} = \mathbf{e} + 1$.

The resulting Dirichlet distribution models the uncertainty over categorical predictions. The predictive mean provides class probabilities, while the total concentration $S = \sum_k \alpha_k$ and the differential entropy reflect the aleatoric and epistemic uncertainty, respectively. This formulation allows for fast, single-pass uncertainty estimation that avoids the complexity and inefficiency of ensembles or sampling-based Bayesian methods.

We empirically observed that early and mid-level convolutional features capture complementary uncertainty cues compared to deep semantic embeddings. Therefore, the meta-model input concatenates features from early–mid–deep layers (e.g., layer2–layer4 in ResNet-18), as identified through ablation in Results and Analysis section.

Max-WEnt Regularization

To enhance the epistemic fidelity of the metamodel, we incorporated *Maximum Weight Entropy (Max-WEnt)* regularization, a strategy originally proposed by Li et al. (de Mathelin et al. 2025) to improve model robustness under distribution shift. We adapt this approach to post hoc UQ by applying it to the weights of the evidential meta-model.

Specifically, we introduce learnable layer-wise scaling factors w_ℓ across all layers of g_ϕ. These weights are normalized to form a probability distribution $p_\ell = w_\ell / \sum_j w_j$, from which we compute the entropy:

$$H(\mathbf{w}) = -\sum_\ell p_\ell \log p_\ell. \tag{1}$$

This entropy is maximized during training to encourage a balanced use of all layers increasing hypothesis diversity improving epistemic uncertainty estimation. The overall loss becomes:

$$\mathcal{L} = \mathcal{L}\text{ELBO} - \lambda H(\mathbf{w}), \tag{2}$$

where $\mathcal{L}$ELBO is the Evidence Lower Bound Loss (comprised of expected log-likelihood and KL divergence with a uniform Dirichlet prior), and λ controls the strength of the regularization. This formulation allows the meta-model to explore a wider hypothesis space without altering the backbone f_θ or incurring test-time overhead.

Algorithm 1: Training of the meta model with Max-WEnt regularization

```
Input: Frozen classifier $f_\theta$, training dataset $\mathcal{D}$
Parameter: Meta-model $g_\phi$, scaling weights $\mathbf{w}$, regularisation strength $\lambda$
Output: Trained meta-model $g_\phi$
for mini-batch $(\mathbf{X}, \mathbf{y}) \in \mathcal{D}$ do
    Extract semantic features: $\mathbf{h} \leftarrow f_\theta^{\text{penult}}(\mathbf{X})$
    Apply Max-WEnt scaling: $\tilde{\mathbf{h}} \leftarrow \textsc{Scale}(\mathbf{h}, \mathbf{w})$
    Predict evidential vector: $\mathbf{e} \leftarrow g_\phi(\tilde{\mathbf{h}})$
    Compute Dirichlet parameters: $\boldsymbol{\alpha} \leftarrow \mathbf{e} + 1$
    Compute ELBO loss: $\mathcal{L}_{\text{ELBO}} \leftarrow \text{D-ELBO}(\boldsymbol{\alpha}, \mathbf{y})$
    Normalize scaling weights: $p_\ell \leftarrow w_\ell / \sum_j w_j$
    Compute entropy penalty:
    $\mathcal{L}_{\text{Max-WEnt}} \leftarrow -\lambda \sum_\ell p_\ell \log p_\ell$
    Total loss: $\mathcal{L} \leftarrow \mathcal{L}_{\text{ELBO}} + \mathcal{L}_{\text{Max-WEnt}}$
    Update $g_\phi$ and $\mathbf{w}$ via AdamW using gradient $\nabla \mathcal{L}$
end for
return trained meta-model $g_\phi$
```

Training Procedure

To encourage the metamodel to express high uncertainty in the absence of explicit out-of-distribution (OOD) samples during training, we utilize noisy in-distribution (ID) variants as a surrogate for uncertainty. This strategy forces the metamodel to learn a robust uncertainty landscape by treating perturbed ID samples as ambiguous cases

Experimental Setup

Datasets

- **Medical imaging**: BACH (histopathology), DIV2K (super-resolution proxy), HAM10000 (skin lesions), BreakHIS (breast-cancer histology).
- **Natural images**: SVHN and Fashion-MNIST.
- **Corruptions**: ImageNet-C (15 corruption types, 5 severities).

Evaluation Protocols

We follow the experimental setup of earlier work(Shen et al. 2023) for post-hoc uncertainty quantification using a Dirichlet meta-model. All meta-models are trained solely on ID training data and evaluated on held-out ID and OOD sets. We report OOD detection AUROC (positive = OOD). Uncertainty scores include predictive entropy (TotEnt), mutual information (MI), and Dirichlet entropy (D-Ent), consistent with earler work(Shen et al. 2023). All results are averaged over three random seeds; means (± std) are reported where space permits. For ImageNet-C experiments, we report MI-AUROC across corruption severities while ensuring no overlap between validation and test corruption levels.

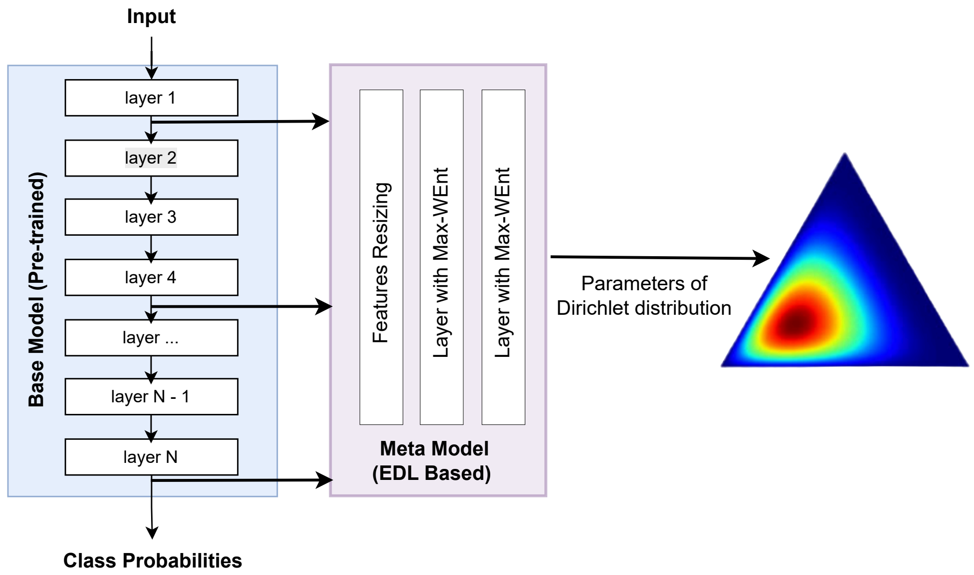

Figure 1: Overview of the training pipeline. Features from multiple layers of the frozen classifier are passed to the meta-model, which predicts the Dirichlet distribution parameters. The Max-WEnt regularization encourages diverse weight scaling for better epistemic uncertainty estimation.

Training Details

- **Dataset Splits:** 80/10/10 train/validation/test split. Out-of-distribution examples are sampled from ImageNet-C (severity level 3).
- **Optimizer:** AdamW with learning rate 5×10^{-4} and weight decay of 10^{-2}.
- **Batch Size:** 128 for medical datasets and 256 for natural images. Models are trained for 50 epochs using cosine learning rate decay.
- **Compute:** All experiments are run on a single NVIDIA A6000 (48GB) GPU with 2 CPU cores. The training time is less than 4 hours per dataset.
- **Random Seeds:** All results are averaged over 3 independent runs. We report the mean and standard deviation.

Model	Metric	OOD Datasets		
		DIV2K	HAM10000	BreakHIS
Base Model	Entropy	69	71	44
	Max-P	64	67	51
BAY-MED	D-Ent	73	78	36
	MI	79	87	53
	Entropy	62	58	38
	Max-P	62	58	43
Our Model	D-Ent	76	88	98
	MI	80	95	67
	Entropy	75	86	**99**
	Max-P	**81**	**96**	87

Table 1: AUROC (%) on medical-imaging benchmarks, base model is ResNet-18 and trained on BACH dataset.

Results and Analysis

Overall Performance

Our metamodel consistently outperforms the baselines on both medical and natural image datasets (Table 1 and 2).

Ablation and Layer-Selection Analysis

To identify which layers contribute most to reliable uncertainty estimation, we systematically varied the subset of ResNet-18 and VGG features provided to the meta-model. Table 3 summarizes the results for CIFAR-10 and SVHN OOD detection. Across both architectures, early convolutional layers yield the highest AUROC (up to 0.99 on VGG and 0.95 on ResNet-18), while very deep layers or indiscriminate fusion of all layers reduce performance. This trend persists across cross-domain settings (CIFAR → SVHN), confirming that low- and mid-level representations encode robust OOD cues, whereas late layers tend to over-fit class semantics.

These results justify our design choice of combining early–mid–deep features (layer2–layer4) rather than using all layers. Notably, the AUROC drops when all layers are concatenated, suggesting redundant or conflicting information. This motivates the use of Max-WEnt regularization to adaptively balance contributions across feature depths, enhancing epistemic diversity without manual pruning.

Discussion

The preceding layer-selection study demonstrates that uncertainty cues are distributed non-uniformly across depth: early layers encode texture-driven aleatoric signals, whereas deeper layers capture category-level semantics but exhibit lower generalization under shift. Max-WEnt regularization

Model	Metric	OOD Datasets		
		SVHN	Fashion-MNIST	ImageNet-C
Base Model	Entropy	88.2	84.1	95.3
	Max-P	85.6	82.5	91.1
Our Model	D-Ent	90.1	82.2	97.0
	MI	99.3	91.6	99.9
	Entropy	84.0	79.1	88.6
	Max-P	91.4	84.1	77.4

Table 2: AUROC (%) on natural-image benchmarks using ResNet-18 as base model.

acts as a principled mechanism to weight these heterogeneous features, encouraging balanced utilization across layers. This explains its consistent benefit in our ablations.

In our work, we extract features from three different layers of a pre-trained classification model, without altering its parameters. This preserves the original classification performance while enabling uncertainty quantification through a more expressive input space. Early convolutional layers contribute texture-level details, while deeper layers provide semantic abstractions. This diverse feature set equips the meta-model with richer information compared to conventional evidential deep learning (EDL) approaches that rely solely on the final layer's representation. As a result, our method achieves improved uncertainty estimation, particularly in settings involving distribution shifts or ambiguous samples.

To further enhance epistemic fidelity, we incorporate Max-WEnt regularization into meta-model training. By maximizing the entropy of learnable scaling weights across layers, Max-WEnt promotes hypothesis diversity within the metamodel. This strategy broadens the learned uncertainty landscape without altering the base classifier or increasing the test-time cost. Empirical study confirm that disabling Max-WEnt consistently reduces performance, underscoring its importance in robust uncertainty estimation.

Limitations. Limitations and Future Work While our evidential metamodel provides an efficient post-hoc solution, it has certain limitations: * Natural Image Performance: The approach demonstrates comparatively lower performance on natural-image OOD detection benchmarks compared to its high efficacy in medical imaging, suggesting a need for better domain-specific feature alignment. Training Surrogates: The reliance on noisy ID data as a surrogate for uncertainty may not fully capture the complexities of real-world distribution shifts. Complexity vs. Efficiency: Although the overhead is modest (0.8M parameters), future work will explore model compression techniques like pruning or distillation for deployment in edge-device clinical settings. * Theoretical Exploration: Further research is required to examine the theoretical basis of Max-WEnt-based feature fusion and analyze the specific training dynamics that lead to improved epistemic sensitivity

Conclusion

We introduced a post-hoc uncertainty quantification framework that combines multi-layer feature extraction with a lightweight Dirichlet evidential metamodel. Unlike conventional EDL methods that depend solely on final-layer features, our approach leverages spatial and semantic signals across the network to provide more nuanced uncertainty estimates. In addition, we integrate Max-WEnt regularization to improve hypothesis diversity and epistemic calibration without incurring additional test-time overhead.

Extensive experiments across seven benchmarks validate the effectiveness of our evidential meta-model, demonstrating consistent gains in AUROC and calibration metrics compared to existing post-hoc UQ methods. Ablations further confirm the essential contributions of both multilayer fusion and Max-WEnt, particularly in capturing texture-driven and semantic cues. Our findings establish this metamodel as a reliable and efficient UQ strategy for high-stakes, safety-critical applications such as medical imaging, where frozen backbones are often required due to regulatory constraints. Future directions include a theoretical exploration of Max-WEnt-based feature fusion and the development of more compact variants for real-world deployment on edge devices.

References

Bala, G.; Chauhan, A.; and Sethi, A. 2025. BAY-MED: Bayesian Approximation for Post-hoc Uncertainty in Medical Imaging. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight Uncertainty in Neural Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

de Mathelin, A.; Deheeger, F.; Mougeot, M.; and Vayatis, N. 2025. Deep Out-of-Distribution Uncertainty Quantification via Weight Entropy Maximization. volume 26, 1–68.

et al., H. 2023. Test Time Augmentation Meets Post-hoc Calibration: Uncertainty Quantification under Real-World Conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14856–14864.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Jain, S.; Vecerik, M.; Seso, A.; Shanmugam, K.; and Song, Y. 2021. DEUP: Direct Epistemic Uncertainty Prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jungo, A.; and Reyes, M. 2020. Analyzing the Effect of Model Uncertainty on Segmentation Quality in Biomedical Images. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

ID → OOD	Metric	All 1st Blks	1&3 Layers (1–2 Blks)	1&3 Layers (2–3 Blks)	1st Layer of All Blks	Last Layer of All Blks
ResNet18 Backbone						
CIFAR10→SVHN	MI-AUROC	0.571	0.623	0.685	**0.771**	0.700
	TotEnt-AUROC	0.570	0.624	0.682	**0.771**	0.700
SVHN→CIFAR10	MI-AUROC	0.245	0.233	**0.734**	0.208	0.727
	TotEnt-AUROC	0.246	0.234	**0.734**	0.208	0.729

Table 3: Study of backbone input selection for the evidential meta-model. Each entry shows the AUROC of OOD detection computed using Mutual Information (MI) and Total Predictive Entropy (TotEnt). Early or distributed features yield stronger uncertainty discrimination than deep features.

Kwon, Y.; Won, J.-H.; Kim, B. J.; and Paik, M. C. 2020. Uncertainty Quantification Using Bayesian Neural Networks in Classification: Application to Biomedical Image Segmentation. *Computational Statistics & Data Analysis*, 142: 106816.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Neal, R. 2012. *Bayesian Learning for Neural Networks*. Springer.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Shen, M.; Liang, K.; Datta, S.; and Chen, X. 2023. Post-hoc Uncertainty Learning Using a Dirichlet Meta-model. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.