Weight Entropy-Maximised Evidential Metamodel for Uncertainty Quantification (Student Abstract)

Gouranga Bala, Abhimanyu Chauhan, Amit Sethi

Indian Institute of Technology Bombay, Mumbai 400076, India
gouranga.bala23@gmail.com, chauhanabhimanyu15@gmail.com, asethi@iitb.ac.in

Abstract

Reliable uncertainty quantification (UQ) is crucial for deploying deep learning models in safety-critical domains. Existing UQ methods often either rely on multi-pass inference, which increases computational cost, or restrict expressiveness by using only final-layer embeddings. In this work, we propose a lightweight evidential metamodel that leverages multi-layer feature fusion from a pretrained backbone, capturing both low-level features and high-level semantics to better estimate uncertainty. To further enhance epistemic fidelity, we integrate maximum weight-entropy (Max-WEnt) regularization, which encourages hypothesis diversity without altering the base network or adding test-time overhead. Experiments across two benchmark settings, medical (BACH, HAM10000, BreakHIS, DIV2K) and natural (ImageNet, SVHN, Fashion-MNIST, ImageNet-C) datasets, demonstrate consistent improvements in AUROC of out-of-distribution detection compared to prior post-hoc UQ methods. Our findings show that combining multi-layer evidential modeling with Max-WEnt provides a robust, efficient, and practical framework for trustworthy AI in high-stakes applications. The metamodel adds only 0.8M parameters and trains in under four hours on a single 48GB GPU, making it practical for real-world deployment.

Introduction

Deep neural networks achieve expert-level accuracy in visual recognition tasks, but remain overconfident in erroneous predictions. In safety-critical settings, this lack of calibrated uncertainty is a critical limitation, as knowing *when* to defer is as important as predicting *what*.

Most post-hoc UQ methods rely solely on the final-layer embedding of the classifier. However, our analysis reveals that important uncertainty cues are distributed across the activations of the network: early layers capture texture-level ambiguities, mid-level layers detect structural inconsistencies, and deeper layers encode semantic confidence. Ignoring these hierarchical features limits robustness under distribution shift.

Post-hoc UQ is especially valuable in deployment scenarios, since it preserves the original accuracy of the pretrained classifier by avoiding any change to backbone weights, and it also requires single-pass inference. This makes post-hoc UQ methods preferable in medical imaging, autonomous driving, and other safety-critical workflows where retraining or architectural changes are not feasible, for instance, due to a prior regulatory approval given to model that need to be frozen.

Post-hoc approaches such as ensembles (Lakshminarayanan, Pritzel, and Blundell 2017), Monte Carlo dropout (Gal and Ghahramani 2016), and temperature scaling (Guo et al. 2017) improve reliability, but these methods incur high inference cost. Our work improves upon the line of research on single-pass UQ models, such as Evidential deep learning (EDL)(Sensoy, Kaplan, and Kandemir 2018), Post-hoc Uncertainty Learning(Shen et al. 2023), and our prior work BAY-MED (Bala, Chauhan, and Sethi 2025) which predict Dirichlet parameters to capture both aleatoric and epistemic uncertainty.

Motivation. We hypothesize that reliable post-hoc UQ requires (i) leveraging multi-layer features of a frozen backbone to capture both texture and semantic signals, and (ii) encouraging diverse contributions from these features. To this end, we propose a lightweight evidential metamodel with Max-WEnt as a regularization. Our goal was to develop a practical, single-pass framework that improves UQ performance, while remaining lightweight for real-world deployment.

Proposed Approach

Multi-layer evidential metamodel. Figure 1 illustrates our framework. The pretrained backbone (left) is kept frozen. Instead of relying only on its final-layer embedding, we extract intermediate features from multiple depths: early layers capture fine textures, while deeper layers capture semantic information. These features are resized and aligned before being passed into a lightweight evidential metamodel (center), implemented as a small MLP. The metamodel is trained using only in-distribution data. No out-of-distribution (OOD) samples are required; instead, noisy in-distribution(ID) variants are used to encourage the model to express uncertainty. The metamodel outputs parameters of a Dirichlet distribution (right), enabling estimation of both aleatoric (data) and epistemic (model) uncertainty.

Max-WEnt regularization. Within the metamodel, we adapt Max-WEnt regularization (de Mathelin et al. 2025).

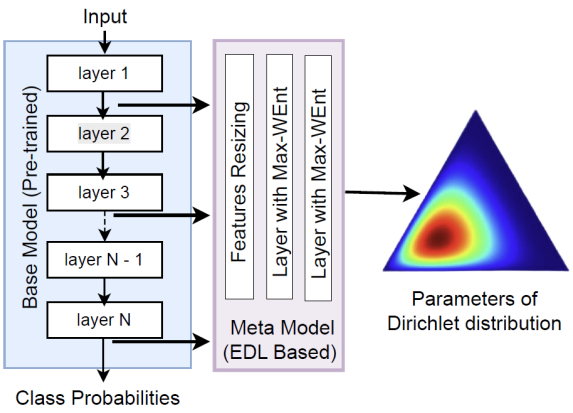

Figure 1: Proposed pipeline: Features from multiple layers of a frozen base classifier are resized and fused in the metamodel. Max-WEnt regularization balances feature contributions, yielding improved uncertainty estimation.

Each feature stream is assigned a learnable scaling weight, normalized into probabilities p_ℓ. To prevent the metamodel from over-relying on a single layer, we maximize the entropy

$$H(p) = -\sum_{\ell} p_\ell \log p_\ell,$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ELBO}}(\alpha; y) \; - \; \lambda \, H(p), \tag{1}$$

encouraging balanced contributions from all layers. This simple addition promotes hypothesis diversity and improves epistemic calibration.

Results

We evaluated our method on medical datasets (BACH as ID; HAM10000, BreakHIS, DIV2K as OOD) and natural datasets (ImageNet as ID; SVHN, Fashion-MNIST, ImageNet-C as OOD) using different metrics. We found that our method consistently improves post-hoc UQ:

- Higher AUROC of OOD detection across all benchmarks, often by wide margins over both the base model and prior methods (e.g., BAY-MED).
- Lightweight and practical: adds $\sim$0.8M parameters, trains in $<$4h on a single GPU, in a single pass.

Tables 1 and 2 summarize these improvements across the medical and natural-image benchmarks.

Ablations. On HAM10000, removing multi-layer fusion reduced OOD detection by up to 8 points. Additionally, comparing against BAY-MED (our prior work without Max-WEnt) shows that Max-WEnt regularization provides further improvements across all metrics.

Discussion

Our framework demonstrates that combining multi-layer feature with Max-WEnt in the metamodel substantially improves uncertainty estimation while maintaining efficiency. Early layers capture texture variations linked to aleatoric noise, while deeper layers encode semantic cues tied to misclassifications. Max-WEnt further prevents the fusion process from collapsing onto a single dominant layer, typically the deepest one; thereby maintaining diversity among feature contributions and improving epistemic sensitivity under distribution shift. Since the backbone remains frozen, all these gains arise solely from the metamodel, making the overall approach lightweight, easy to deploy, and well-suited for trustworthy AI applications.

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(%) of OOD detection on medical-imaging benchmarks, using ResNet-18 trained on BACH.

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(%) of OOD detection on natural-image benchmarks, using ResNet-18 as base model on ImageNet.

Limitations & Future Work

Our method represents an incremental extension of prior evidential metamodels, and therefore we do not include comparisons with heavier UQ baselines (e.g., deep ensembles) or report additional calibration metrics such as ECE, NLL, Brier score, or reliability diagrams. The approach demonstrates comparatively lower performance on natural-image OOD detection, indicating an opportunity for further improvement in this setting. Additionally, the metamodel is trained using noisy ID data as a surrogate for uncertainty, as no OOD samples are used during training.

Future work includes developing a more principled ID-only training strategy to enhance generalization, examining the theoretical basis of Max-WEnt based feature fusion, analyzing metamodel training dynamics, creating more compact variants via better pretrained feature selection, and extending the framework beyond vision tasks.

References

Bala, G.; Chauhan, A.; and Sethi, A. 2025. BAY-MED: Bayesian Approximation for Post-Hoc Uncertainty in Medical Imaging. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 1–4.

de Mathelin, A.; Deheeger, F.; Mougeot, M.; and Vayatis, N. 2025. Deep Out-of-Distribution Uncertainty Quantification via Weight Entropy Maximization. *Journal of Machine Learning Research*, 26(4): 1–68.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Balcan, M. F.; and Weinberger, K. Q., eds., *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, 1050–1059. New York, New York, USA: PMLR.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. *CoRR*, abs/1706.04599.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6405–6416. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. arXiv:1806.01768.

Shen, M.; Bu, Y.; Sattigeri, P.; Ghosh, S.; Das, S.; and Wornell, G. 2023. Post-hoc uncertainty learning using a dirichlet meta-model. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*. AAAI Press. ISBN 978-1-57735-880-0.