Laplace Outperforms Evidential Deep Learning for Out-of-Distribution Detection using Epistemic Uncertainty

Abstract

Evidential deep learning (EDL) is a lightweight alternative to Bayesian approaches for uncertainty quantification of neural net classifiers. Yet a recent critique shows that EDL's estimates of epistemic uncertainty (EU) problematically plateau rather than approach zero as dataset size increases. In light of this critique, we set out to show that using an efficient last-layer Laplace approximation of the Bayesian posterior over weights outperforms EDL for EU-based out-of-distribution (OOD) detection. However, the current literature quantifies "epistemic uncertainty" with three mathematically different metrics based on mutual information, variance, and subjective logic uncertainty mass. First, we show empirically that for EDL, all three metrics fail to monotonically decrease as dataset size increases, unlike Laplace methods which show desired behavior. Second, evaluating OOD detection for image classifiers, we show that the mutual information metric with Laplace methods convincingly outperforms other metrics and other methods like EDL. Our results support the growing consensus that EDL fails to reliably quantify EU, and instead suggest Bayesian neural networks as a more promising alternative.

1 INTRODUCTION

The deployment of deep learning in high-stakes domains such as medical diagnosis highlights the need for uncertainty quantification (UQ) [Lopez et al., 2025]. In practice, UQ could enable the detection of uncertain or unreliable predictions which could warrant human review to avoid costly mistakes.

A key step in UQ methods is separating two kinds of uncertainty. *Aleatoric* uncertainty (AU) is irreducible noise in the data generation and measurement process. *Epistemic* uncertainty (EU) is reducible uncertainty in the chosen model due to limited data. By definition, measures of EU on in-distribution data should decay to zero as dataset size increases. Bayesian neural networks (BNNs) [Neal, 1996] are well-known to obey these desired theoretic trends, yet can be expensive to train and deploy.

Evidential deep learning (EDL) [Sensoy et al., 2018] has recently emerged as an efficient approach to UQ. EDL estimates uncertainty for a deterministic neural network classifier by producing a Dirichlet distribution over class probabilities in one forward pass. EDL ideas have inspired a host of recent methods [Malinin and Gales, 2018, 2019, Charpentier et al., 2020]. However, Shen et al. [2024] argue that deterministic EDL-based EU estimates do not decay to zero as dataset size increases, and so cannot reliably estimate EU. Shen et al. [2024] conjecture that BNNs could be a practical solution. Yet a head-to-head comparison of EDL and tractable BNNs remains underexplored.

Motivated by these limitations, we set out to show that a particularly tractable BNN approach, the *Laplace approximation* (LA) [MacKay, 1992, Daxberger et al., 2021], can outperform EDL for EU-based out-of-distribution (OOD) detection. However, the current literature uses the term "epistemic uncertainty" for three mathematically different metrics: mutual information [Shen et al., 2024], variance of expectation [Yoon and Kim, 2025], and subjective logic uncertainty mass [Sensoy et al., 2018]. Therefore, we evaluate several UQ methods with all three EU metrics and contribute two main findings. First, we show empirically that for EDL all three metrics fail to monotonically decrease as dataset size increases, supporting the theory from Shen et al. [2024]. Second, for out-of-distribution detection in image classifiers, we show that using Laplace approximation BNNs with the mutual information metric outperforms all other metrics and methods. Our results suggest that EDL fails to estimate epistemic uncertainty properly, while BNNs are better suited for reliable UQ.

Submitted to the 9th Workshop on Tractable Probabilistic Modeling at UAI (TPM 2026). **To be used for reviewing only**.

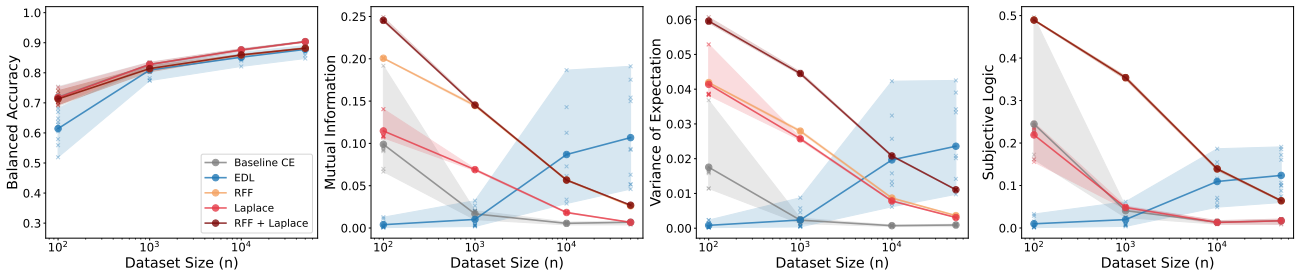

Figure 1: Classifier quality (panel 1) and Epistemic Uncertainty (EU) metrics (panels 2-4) on CIFAR-10 test set as training set size increases. In panels 2-3, Laplace y-values were rescaled by factor of 1/15 to fit on common plot (unscaled plots in Fig. D.1). In panel 4, Baseline CE is hidden under Laplace, similarly RFF is under RFF + Laplace.

2 RELATED WORKS

Critiques of EDL. Many works have questioned whether EDL captures meaningfully epistemic uncertainty [Kopetzki et al., 2021, Bengs et al., 2022, Shen et al., 2024]. Shen et al. specifically argue that EDL and the later inspirations [Yoon and Kim, 2025, Malinin and Gales, 2018, 2019, Charpentier et al., 2020, 2022] all conflate aleatoric and epistemic components and learn spurious distributional uncertainty. Building on this theory-based critique specific to mutual information EU, we corroborate its predicted failure mode across three metrics for estimated EU. For scope, we study only the original EDL formulation [Sensoy et al., 2018]; the critique, and likely our findings, extend to later variants, which we leave to future work.

UQ benchmarks. Mucsányi et al. [2024] benchmark 19 UQ methods using mutual information as the only EU metric and show that EDL does not disentangle AU and EU. We complement this benchmark by evaluating EDL across all three EU metrics (mutual information, variance of expectation, and subjective-logic uncertainty mass) and comparing its OOD detection performance to Bayesian alternatives.

Bayesian UQ. Bayesian methods quantify uncertainty through a distribution over weights [Neal, 1996], e.g. MC dropout [Gal and Ghahramani, 2016] and last-layer Laplace approximations [Daxberger et al., 2021]. Our experiments build on the spectrally normalized neural Gaussian process (SNGP) [Liu et al., 2020].

3 BACKGROUND

Problem setup. We consider classification with input $x \in \mathbb{R}^D$ and one-hot label $y \in \{0,1\}^C$ satisfying $\sum_{c=1}^C y_c = 1$. We aim to learn the conditional distribution $p(y|x)$ from a data set $\{(x_i, y_i)\}_{i=1}^N$ via deep neural network (DNN) with weights θ. Given an input x, the network produces a single vector π on the C-dimensional probability simplex, parameterizing the categorical likelihood $p(y|\pi) = \prod_{c=1}^C \pi_c^{y_c}$. The marginal of the c-th entry of vector y is $y_c|\pi_c \sim \text{Bern}(\pi_c)$.

Conventional DNNs produce a single vector π. Aleatoric and epistemic uncertainty cannot be disentangled from a single point. The spread of plausible π values, a component of epistemic uncertainty, remains unobservable.

Evidential deep learning (EDL). EDL introduces a *meta-distribution* $q(\pi|x,\theta)$ over the simplex, so that $p(y|x,\theta) = \int p(y|\pi)\, q(\pi|x,\theta)\, d\pi$. Generally, EDL methods define $q(\pi|x,\theta) = \text{Dir}\big(\pi|\alpha(x,\theta)\big)$, where learned network $\alpha(x,\theta)$ produces a C-dim. non-negative parameter vector for the Dirichlet. Our focus is on the canonical formulation [Sensoy et al., 2018], $\alpha(x,\theta) = e_\theta(x) + u$, where e is a DNN producing a non-negative C-dim. vector that depends on input x and u is a constant vector controlling the base frequency of classes, typically $u = 1$. Later EDL approaches have pursued other constructions and even more flexible extensions of the Dirichlet [Yoon and Kim, 2025, Malinin and Gales, 2018, 2019, Charpentier et al., 2020, 2022].

UQ metrics. Given the meta-distribution $q(\pi|x,\theta)$, we observed three common measures of uncertainty in the literature. First, a mutual information decomposition [Shen et al., 2024]:

$$\mathrm{H}_{p(y|x,\theta)}(y) = \underbrace{\mathbb{E}_{q(\pi|x,\theta)}\left[\mathrm{H}_{p(y|\pi)}(y)\right]}_{\text{aleatoric}} + \underbrace{I(y;\pi|x,\theta)}_{\text{EU}_{\text{MI}}:\text{epistemic}}. \quad (1)$$

Second, a decomposition of the sum of marginal variances [Yoon and Kim, 2025]:

$$\sum_c \mathrm{Var}_{p(y|x,\theta)}(y_c) = \underbrace{\textstyle\sum_c \mathbb{E}_{q(\pi|x,\theta)}\left[\mathrm{Var}_{p(y_c|\pi_c)}(y_c)\right]}_{\text{aleatoric}} + \underbrace{\textstyle\sum_c \mathrm{Var}_{q(\pi|x,\theta)}\left(\mathbb{E}_{p(y_c|\pi_c)}[y_c]\right)}_{\text{EU}_{\text{Var}}:\ \text{epistemic}}. \quad (2)$$

Finally, the third UQ definition comes from subjective logic (SL) [Jøsang, 2018, Sensoy et al., 2018]. While the previous two metrics allow any valid meta-distribution q, this last metric relies on a Dirichlet form for q. The SL interpretation of a Dirichlet parameter vector $\alpha = [\alpha_1, \ldots, \alpha_C]$ views each entry as a decomposable sum, $\alpha_c = e_\theta(x)_c + u_c$, where $e_\theta(x)_c \geq 0$ is the data-informed signal for class c and $u_c > 0$ represents the uninformed prior mass on class

c. SL's uncertainty metric is then the ratio of uninformed to total mass: $\text{EU}_{\text{SL}} = \frac{\sum_c u_c}{\sum_c \alpha_c}$.

EU when q is Dirichlet. For a specific Dirichlet q with parameter vector α whose sum is $\alpha_0 = \sum_c \alpha_c$, the three EU metrics have tractable closed forms (App. B.1):

$$\text{EU}_{\text{MI}} = \sum_{c=1}^{C} \frac{\alpha_c}{\alpha_0} \left[\log \frac{\alpha_0}{\alpha_c} + \psi(\alpha_c{+}1) - \psi(\alpha_0{+}1) \right] \quad (3)$$

$$\text{EU}_{\text{Var}} = \sum_{c=1}^{C} \frac{\alpha_c(\alpha_0 - \alpha_c)}{\alpha_0^2(\alpha_0 + 1)} \quad (4)$$

$$\text{EU}_{\text{SL}} = \frac{\sum_c u_c}{\alpha_0}. \quad \text{EU}_{\text{SL}} = \frac{C}{\alpha_0} \text{when } u_c = 1 \; \forall c. \quad (5)$$

where ψ is the digamma function. All EU terms above would vanish as $\alpha_0 \to \infty$. However, minimizing EDL's loss directs the learned Dirichlet toward a finite fixed constant the depends on a regularization parameter and is independent of N (see Theorem 5.1 and Example 5.2 of Shen et al. [2024]). As we discuss in App. A.2, this failure mode is apparent when models are fully converged in terms of training loss; early stopping may not reveal the full effects.

Laplace approximation. From a Bayesian perspective, model uncertainty is captured by a posterior distribution over weights. We focus on the *last layer* of a DNN for tractability. Let $\beta \in \mathbb{R}^{D \times C}$ denote the weights of the linear classification head that maps the D-dimensional output of DNN $f_\theta(x)$ to logits for the C classes. We use a Laplace approximation to locally approximate the posterior of β via a Gaussian centered at the maximum a posteriori (MAP) estimate: $\beta \sim \mathcal{N}(\hat{\beta}, \hat{\Sigma})$. To find the MAP estimate $\hat{\beta}$, we first use ADAM to minimize a loss that sums cross-entropy of each y with L_2 regularization (implying a Gaussian prior on β). Next, the covariance matrix $\hat{\Sigma}$ for that $\hat{\beta}$ can computed efficiently as described in Liu et al. [2020].

EU for Laplace approx. The earlier EU metrics above all rely on a meta-distribution $q(\pi|x, \theta)$. For any Bayesian method that can yield posterior samples of θ, we can define q implicitly via these samples. For Laplace methods, we draw M samples $\beta^{(m)} \sim \mathcal{N}(\hat{\beta}, \hat{\Sigma})$ and compute the corresponding predictive distribution $\pi^{(m)}(x) = \text{SOFTMAX}\left(f_\theta(x)^\top \beta^{(m)}\right)$. These samples from the induced meta-distribution $q(\pi|x, \theta)$ can be used to calculate Monte Carlo estimates of the expectations in first two EU metrics (App. B.2):

$$\text{EU}_{\text{MI}} \approx -\sum_{c=1}^{C} \bar{\pi}_c(x) \log \bar{\pi}_c(x) + \frac{1}{M} \sum_{m=1}^{M} \sum_{c=1}^{C} \pi_c^{(m)}(x) \log \pi_c^{(m)}(x), \quad (6)$$

$$\text{EU}_{\text{Var}} \approx \frac{1}{M-1} \sum_{c=1}^{C} \sum_{m=1}^{M} \left(\pi_c^{(m)}(x) - \bar{\pi}_c(x) \right)^2. \quad (7)$$

Here, we define the predictive posterior mean $\bar{\pi} = \frac{1}{M} \sum_{m=1}^{M} \pi^{(m)}$ and use the well-known bias-corrected estimator for variance.

The SL metric requires a Dirichlet q, so we crudely construct a concentration parameter via $\alpha = \frac{1}{M} \sum_m \exp(f_\theta(x)^T \beta^{(m)}) + u$, then use Eq. (5).

RFFs for UQ. Reliable UQ requires knowing when new input is "far" from any training data. To achieve this goal, Liu et al. [2020] suggest a feature transformation known as a random Fourier feature (RFF) layer applied as the penultimate step of a DNN. Given the D-dim. output of a DNN $h_i = f_\theta(x_i)$, we map to an R-dim. RFF embedding via

$$\phi_{\text{RFF}}(h_i) = \sigma_k \sqrt{\tfrac{2}{R}} \cos\left(\tfrac{1}{\ell_k} A^\top h_i + b \right), \quad (8)$$

where RFF weights $A \in \mathbb{R}^{D \times R}$, $b \in \mathbb{R}^R$ are sampled *once* via $A_{d,r} \sim \mathcal{N}(0, 1)$, $b_r \sim \text{Unif}(0, 2\pi)$ then held frozen. This RFF embedding allows a linear weight-driven classifier to efficiently mimic the behavior of a distance-aware radial basis function (RBF) kernel classifier. Key hyperparameters here include the outputscale $\sigma_k > 0$ and the lengthscale $\ell_k > 0$. This construction comes from Rahimi and Recht [2007], with generalization to arbitrary lengthscales $\ell_k > 0$ due to Harvey et al. [2026]. Later experiments examine whether this optional penultimate layer improves OOD detection with or without a Laplace approximation.

4 EXPERIMENTS

Datasets, Models, and Features. Inputs are encoded with a frozen ImageNet-pretrained ResNet-50, yielding 2048-dim. embeddings. We evaluate all five models discussed previously (CE baseline, EDL, RFF, Laplace, and RFF + Laplace), which all share the same frozen embeddings of input images and differ only in the classification head and uncertainty mechanism. We train and evaluate epistemic uncertainty on CIFAR-10, and evaluate OOD detection on CIFAR-100, SVHN, FashionMNIST, TinyImageNet, and Random Noise (32×32 images with uniform pixel values).

Epistemic Scaling. Epistemic uncertainty for ID inputs should decrease monotonically towards zero as $N \to \infty$. We train each model on nested CIFAR-10 subsets $N \in \{100, 10^3, 10^4, 5{\times}10^4\}$, selecting model specific hyperparameters via a grid search on validation data (App. C.1). For each EU metric, we evaluate its mean over all test set images. We repeat over 10 random initializations and report the average metric value, for each metric (mutual information, law-of-total-variance, subjective logic) at each train set size N. Since the issues with EDL arise at the loss minimum [Shen et al., 2024], all models are trained to training loss convergence as discussed in Appendix C.1.4.

OOD Detection. OOD detection requires good generalization. Thus, we fit models to a rather large train set

Table 1: OOD Detection AUROC (↑) by uncertainty metric across possible OOD data sources. ID dataset is always CIFAR-10, with training set size N=10,000. Values shown as mean ± standard deviation across 10 random seeds. **Bold** indicates the best overall model/metric pair for an OOD data source and underline indicates the second best.

	CIFAR-100			SVHN			TinyImageNet			FashionMNIST			Random Noise		
Model	MI	Var	SL	MI	Var	SL	MI	Var	SL	MI	Var	SL	MI	Var	SL
CE Baseline	0.840 ±0.004	0.842 ±0.004	0.839 ±0.004	0.867 ±0.025	0.870 ±0.025	0.864 ±0.025	0.830 ±0.009	0.830 ±0.009	0.830 ±0.009	0.904 ±0.016	0.904 ±0.016	0.904 ±0.016	0.851 ±0.031	0.855 ±0.030	0.847 ±0.032
EDL	0.772 ±0.036	0.773 ±0.036	0.774 ±0.035	0.823 ±0.046	0.824 ±0.046	0.824 ±0.046	0.725 ±0.038	0.726 ±0.038	0.727 ±0.038	0.791 ±0.050	0.792 ±0.050	0.793 ±0.051	0.854 ±0.076	0.854 ±0.078	0.854 ±0.076
RFF	0.860 ±0.002	0.863 ±0.002	0.860 ±0.002	**0.946 ±0.005**	0.944 ±0.006	**0.946 ±0.005**	0.880 ±0.003	0.877 ±0.003	0.880 ±0.003	0.969 ±0.006	0.968 ±0.005	0.969 ±0.006	0.995 ±0.003	0.995 ±0.002	0.995 ±0.003
Laplace	**0.886 ±0.004**	0.875 ±0.003	0.839 ±0.005	0.937 ±0.008	0.878 ±0.011	0.874 ±0.019	**0.925 ±0.007**	0.922 ±0.012	0.826 ±0.008	0.979 ±0.004	0.972 ±0.005	0.905 ±0.016	**1.000 ±0.000**	**1.000 ±0.000**	0.851 ±0.030
RFF+Laplace	0.864 ±0.002	0.815 ±0.002	0.860 ±0.002	0.851 ±0.008	0.718 ±0.006	**0.946 ±0.005**	0.924 ±0.002	0.888 ±0.003	0.880 ±0.003	**0.981 ±0.002**	0.936 ±0.004	0.969 ±0.006	0.999 ±0.000	0.976 ±0.008	0.995 ±0.003

(N = 10,000) and apply early stopping on maximum validation balanced accuracy (App. C.1.4). For each model and EU metric, we compute test set AUROC for binary ID-vs-OOD detection (10,000 ID test samples vs. 10,000 OOD samples). Scores come from the closed-form Dirichlet construction of $q(\pi|x,\theta)$ for deterministic models (CE, EDL, RFF) and from Monte Carlo estimates (10,000 samples) for the Bayesian models (Laplace, RFF + Laplace). We report AUROC averaged over ten seeds and include a correlation analysis across seed, model, and N values.

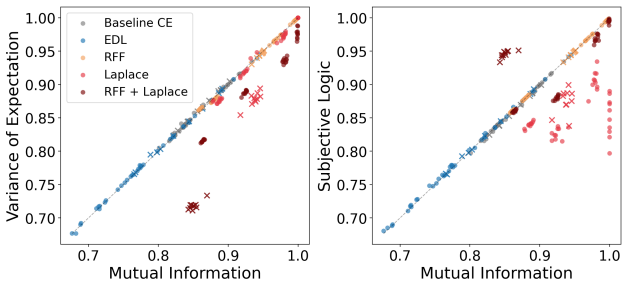

Figure 2: Correlation of OOD detection AUROC between MI (x-axis) and alternative EU metrics. Performance is evaluated across all OOD datasets. SVHN is plotted separately with x markers, due to its anomalous behavior.

5 DISCUSSION

Which methods produce EU estimates that fall as N increases? A valid epistemic uncertainty metric should decrease monotonically toward zero as training dataset size N increases. Empirically, EDL fails this under all three metrics (often actually increasing with N in Fig. 1). This corroborates the theory from Shen et al.. The two Bayesian models (Laplace and RFF + Laplace) recover the expected decrease under the more rigorous metrics: Mutual Information (MI) and Variance of Expectation (Var).

We were surprised by the CE baseline appearing to decrease and plateau near zero in Fig. 1, but we have doubts this reflects solid EU estimation. No metric monotonically decreases, and when evaluated at a non-converged early-stopped checkpoint its estimates are sporadic under all metrics, unlike other methods where trends observed at convergence still hold (Fig. D.4). Only RFF and RFF + Laplace decrease monotonically across all three metrics. We leave why a point-estimate RFF head succeeds (possibly via its RBF-kernel distance-awareness) to future work.

For the SL metric, both Laplace and the CE baseline show EU increase between N=10,000 and N=50,000 (Fig. D.3). We attribute this to SL unnaturally forcing models into its Dirichlet construction.

Which model/metric pairs detect OOD best? Table 1 reports AUROC for OOD detection when N=10,000. See App. D.3 for how performance varies with data size N. Across four of five datasets (CIFAR-100, TinyImageNet, FashionMNIST, Random Noise) the pattern in Tab. 1 is consistent: MI is the top metric, and Laplace and RFF + Laplace outperform all other models, usually constituting the best and second-best pair. In general, Bayesian methods paired with MI outperform their counterparts on OOD detection. SVHN is the exception. It favors the SL metric and is the only dataset where a non-Bayesian model reaches the top three.

EDL performs poorly throughout. Prior benchmarks report equivalent/slightly better OOD AUROC for EDL in comparison to a CE baseline [Mucsányi et al., 2024, Shen et al., 2024], but they refine an entire DNN classifier end-to-end with the EDL loss. In contrast, we only fine-tune the classifier head, which may explain why EDL performs worse. The simple Laplace additions still see large gains under this same last-layer-only protocol.

Does the choice of metric matter? We compare the test AUROC for OOD detection for possible pairs of metrics when N=10000 in Fig. 2. The CE, EDL, and RFF methods lie essentially along the diagonal, showing no real preference between metrics, consistent with Tab. 1. Bayesian methods (Laplace and RFF+Laplace) instead sometimes sit off the diagonal and in favor of MI for all datasets (SVHN excepted). This preference is clearest at N=10,000 and above; at smaller N (Fig. D.13), additional datasets show Bayesian methods occasionally preferring the SL metric.

References

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of Epistemic Uncertainty Quantification through Loss Minimisation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior Network: Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Bertrand Charpentier, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. Natural Posterior Network: Deep Bayesian Predictive Uncertainty for Exponential Family Distributions. In *International Conference on Learning Representations (ICLR)*, 2022.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace Redux - Effortless Bayesian Deep Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning (ICML)*, 2016.

Ethan Harvey, Mikhail Petrov, and Michael C. Hughes. Learning Hyperparameters via a Data-Emphasized Variational Objective. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026.

Audun Jøsang. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, 2018.

Anna-Kathrin Kopetzki, Bertrand Charpentier, Daniel Zügner, Sandhya Giri, and Stephan Günnemann. Evaluating Robustness of Predictive Uncertainty Estimation: Are Dirichlet-based Models Reliable? In *International Conference on Machine Learning (ICML)*, 2021.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Leopoldo Julian Lechuga Lopez, Shaza Elsharief, Dhiyaa Al Jorf, Firas Darwish, Congbo Ma, and Farah E. Shamout. Uncertainty Quantification for Machine Learning in Healthcare: A Survey. In *Conference on Health, Inference, and Learning (CHIL)*, 2025.

David J. C. MacKay. Bayesian Interpolation. *Neural Computation*, 4(3):415–447, 1992.

Andrey Malinin and Mark Gales. Predictive Uncertainty Estimation via Prior Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Andrey Malinin and Mark Gales. Reverse KL-Divergence Training of Prior Networks: Improved Uncertainty and Adversarial Robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking Uncertainty Disentanglement: Specialized Uncertainties for Specialized Tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer Science & Business Media, 1996.

Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

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

Maohao Shen, Jongha Jon Ryu, Soumya Ghosh, Yuheng Bu, Prasanna Sattigeri, Subhro Das, and Gregory Wornell. Are Uncertainty Quantification Capabilities of Evidential Deep Learning a Mirage? *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Taeseong Yoon and Heeyoung Kim. Uncertainty Estimation by Flexible Evidential Deep Learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Supplementary Material

A DETAILED BACKGROUND

A.1 SUBJECTIVE LOGIC UNCERTAINTY (DEMPSTER-SHAFER)

This section of the Appendix outlines the subjective logic Dirichlet to Multinomial Opinion mapping as established by Jøsang. Subjective Logic represents belief over the C classes as a multinomial opinion $\omega = (b, u_{\text{mass}}, a)$, where b is a belief mass distribution over the C classes, $u_{\text{mass}} \in [0, 1]$ is a scalar *uncertainty mass* marking the lack of evidence, and a is a prior base-rate distribution assumed from the problem. These satisfy the additivity constraint

$$u_{\text{mass}} + \sum_{c=1}^{C} b_c = 1. \tag{9}$$

Belief and uncertainty mass are computed from the per-class data-informed signal $e_\theta(x)_c \geq 0$ via

$$b_c = \frac{e_\theta(x)_c}{\alpha_0}, \qquad u_{\text{mass}} = \frac{\sum_c u_c}{\alpha_0}, \qquad \alpha_0 = \sum_{c=1}^{C} u_c + \sum_{c=1}^{C} e_\theta(x)_c, \tag{10}$$

where $u_c > 0$ is the uninformed prior mass on class c (conventionally $u_c = 1$, so $\sum_c u_c = C$). This multinomial opinion maps to a Dirichlet meta-distribution $q = \text{Dir}(\alpha)$ through

$$\alpha_c = e_\theta(x)_c + u_c, \tag{11}$$

recovering the main-text decomposition with strength $\alpha_0 = \sum_c \alpha_c$. Taking a uniform base rate with $u_c = 1$ for all c gives the standard EDL parameterization $\alpha_c = e_\theta(x)_c + 1$. The uncertainty mass is then a closed-form function of the Dirichlet strength,

$$u_{\text{mass}} = \frac{\sum_c u_c}{\sum_c u_c + \sum_{c=1}^{C} e_\theta(x)_c} = \frac{\sum_c u_c}{\alpha_0}, \tag{12}$$

which is the subjective logic epistemic metric EU_{SL} used throughout this work (equal to C/α_0 when $u_c = 1$). It depends on x only through the total evidence: $u_{\text{mass}} \to 0$ as $\alpha_0 \to \infty$ (a dogmatic, fully-determined opinion) and $u_{\text{mass}} = 1$ when $e_\theta(x)_c = 0$ for all c (a vacuous opinion). As discussed in Section A.2, this vanishing is guaranteed only if the network drives $\alpha_0 \to \infty$ with N, which the EDL objective does not enforce.

A.2 SHEN ET AL. EPISTEMIC CLAIMS

As described in the main paper epistemic uncertainty on ID inputs should decrease monotonically to zero as $N \to \infty$. With meta-distribution $q(\pi|x, \theta) = \text{Dir}(\alpha(x, \theta))$ and Dirichlet strength $\alpha_0(x) = \sum_c \alpha_c(x)$, all three EU metrics ($\text{EU}_{\text{MI}}$, EU_{Var}, EU_{SL}) vanish only as $\alpha_0(x) \to \infty$, i.e. as the Dirichlet collapses to a point mass. For EU to behave correctly, then, training would have to drive $\alpha_0(x)$ without bound as N increases.

The Shen et al. [2024] critique shows this does not happen. Their Theorem 5.1 characterizes the global optimum of the reverse-KL objective (our SSE + KL Reg above): in the population limit, the learned meta-distribution $q(\pi|x,\theta)$ is forced to match a fixed target $q^\star(\pi|x)$ that does not depend on N. For the Dirichlet prior and categorical likelihood (their Theorem 5.2), this target is itself a Dirichlet of finite strength. Training therefore pulls $\alpha(x,\theta)$ toward this fixed, finite target not dependent on N, not toward the $\alpha_0 \to \infty$ limit that would drive EU to zero, or even to a finite target that increases with N, so all three EDL EU metrics settle to the same non-zero floor rather than vanishing. The fixed target is dependent upon regularization strength of the reverse-KL objective, not training data.

Crucially, this characterizes the *optimum* of the loss, which is why the failure mode is only exposed at convergence. The fixed target $q^\star$ describes where training is headed, not where an arbitrary checkpoint sits. Before convergence, $\alpha_0(x)$ has not yet been pulled to its target: the meta-distribution is still in transit, and its EU estimates can take a wide range of values—sometimes coincidentally small, sometimes erratic—that can superficially resemble well-behaved uncertainty. The non-zero floor only emerges once $q(\pi|x,\theta)$ has actually settled near $q^\star$. Evaluating an early-stopped EDL model therefore risks mistaking an unconverged transient for genuine epistemic uncertainty, which is why we train all models to convergence before assessing EU.

B DERIVATIONS OF EPISTEMIC METRICS

B.1 DIRICHLET BASED UNCERTAINTY METRIC DERIVATIONS

Proof of Closed Form: Variance of Expectation (Dirichlet)

Setup. EDL parameterizes a Dirichlet meta-distribution $q(\pi|x,\theta) = \mathrm{Dir}\big(\pi|\alpha(x,\theta)\big)$, where $\alpha(x,\theta) = e_\theta(x) + u$. Write $\alpha_c := \alpha(x,\theta)_c$ and $\alpha_0 := \sum_{c=1}^{C} \alpha_c$. The marginal likelihood is $y_c|\pi_c \sim \mathrm{Bern}(\pi_c)$, so $\mathbb{E}_{p(y_c|\pi_c)}[y_c] = \pi_c$ and $\mathrm{Var}_{p(y_c|\pi_c)}(y_c) = \pi_c(1-\pi_c)$. We use two standard Dirichlet moments: $\mathbb{E}_{q(\pi|x,\theta)}[\pi_c] = \alpha_c/\alpha_0$ and $\mathrm{Var}_{q(\pi|x,\theta)}(\pi_c) = \alpha_c(\alpha_0-\alpha_c)/(\alpha_0^2(\alpha_0+1))$.

Epistemic. Starting from the variance definition of EU:

$$
\begin{aligned}
\mathrm{EU}(x) &= \sum_{c=1}^{C} \mathrm{Var}_{q(\pi|x,\theta)}\big(\mathbb{E}_{p(y_c|\pi_c)}[y_c]\big) && \text{(definition)} \\
&= \sum_{c=1}^{C} \mathrm{Var}_{q(\pi|x,\theta)}(\pi_c) && \text{(Bernoulli mean is } \pi_c\text{)} \\
&= \sum_{c=1}^{C} \frac{\alpha_c(\alpha_0-\alpha_c)}{\alpha_0^2(\alpha_0+1)} && \text{(Dirichlet variance moment)}
\end{aligned}
$$

Aleatoric. Starting from the variance definition of AU:

$$
\begin{aligned}
\mathrm{AU}(x) &= \sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}\big[\mathrm{Var}_{p(y_c|\pi_c)}(y_c)\big] && \text{(definition)} \\
&= \sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}\big[\pi_c(1-\pi_c)\big] && \text{(Bernoulli variance is } \pi_c(1-\pi_c)) \\
&= \sum_{c=1}^{C} \big(\mathbb{E}_{q(\pi|x,\theta)}[\pi_c] - \mathbb{E}_{q(\pi|x,\theta)}[\pi_c^2]\big) && \text{(linearity)} \\
&= \sum_{c=1}^{C} \Big(\mathbb{E}_{q(\pi|x,\theta)}[\pi_c] - \mathrm{Var}_{q(\pi|x,\theta)}(\pi_c) - \big(\mathbb{E}_{q(\pi|x,\theta)}[\pi_c]\big)^2\Big) && (\mathbb{E}[X^2] = \mathrm{Var}(X) + (\mathbb{E}[X])^2) \\
&= \sum_{c=1}^{C} \left(\frac{\alpha_c}{\alpha_0} - \frac{\alpha_c^2}{\alpha_0^2} - \frac{\alpha_c(\alpha_0-\alpha_c)}{\alpha_0^2(\alpha_0+1)}\right) && \text{(substitute Dirichlet moments)} \\
&= \sum_{c=1}^{C} \left(\frac{\alpha_c(\alpha_0-\alpha_c)}{\alpha_0^2} - \frac{\alpha_c(\alpha_0-\alpha_c)}{\alpha_0^2(\alpha_0+1)}\right) && \text{(combine first two terms)} \\
&= \sum_{c=1}^{C} \frac{\alpha_c(\alpha_0-\alpha_c)}{\alpha_0^2}\left(1 - \frac{1}{\alpha_0+1}\right) && \text{(factor)} \\
&= \sum_{c=1}^{C} \frac{\alpha_c(\alpha_0-\alpha_c)}{\alpha_0^2} \cdot \frac{\alpha_0}{\alpha_0+1} && \text{(simplify)} \\
&= \sum_{c=1}^{C} \frac{\alpha_c(\alpha_0-\alpha_c)}{\alpha_0(\alpha_0+1)} && \text{(cancel one } \alpha_0)
\end{aligned}
$$

Total. Starting from the variance definition of TU:

$$
\begin{aligned}
\mathrm{TU}(x) &= \sum_{c=1}^{C} \mathrm{Var}_{p(y|x,\theta)}(y_c) && \text{(definition)} \\
&= \mathrm{AU}(x) + \mathrm{EU}(x) && \text{(law of total variance)} \\
&= \sum_{c=1}^{C} \frac{\alpha_c(\alpha_0-\alpha_c)}{\alpha_0(\alpha_0+1)} + \sum_{c=1}^{C} \frac{\alpha_c(\alpha_0-\alpha_c)}{\alpha_0^2(\alpha_0+1)} && \text{(substitute closed forms)} \\
&= \sum_{c=1}^{C} \frac{\alpha_c(\alpha_0-\alpha_c)}{\alpha_0+1}\left(\frac{1}{\alpha_0} + \frac{1}{\alpha_0^2}\right) && \text{(factor)} \\
&= \sum_{c=1}^{C} \frac{\alpha_c(\alpha_0-\alpha_c)}{\alpha_0+1} \cdot \frac{\alpha_0+1}{\alpha_0^2} && \text{(common denominator)} \\
&= \sum_{c=1}^{C} \frac{\alpha_c(\alpha_0-\alpha_c)}{\alpha_0^2} && \text{(cancel } \alpha_0+1)
\end{aligned}
$$

Proof of Closed Form: Entropy / Mutual Information (Dirichlet)

Setup: the network. EDL parameterizes a Dirichlet meta-distribution $q(\pi|x,\theta) = \mathrm{Dir}\big(\pi|\alpha(x,\theta)\big)$, where $\alpha(x,\theta) = e_\theta(x) + u$. Write $\alpha_c := \alpha(x,\theta)_c$ and $\alpha_0 := \sum_{c=1}^{C} \alpha_c$.

Setup: the label model. We assume $y|\pi \sim \mathrm{Cat}(\pi)$, which is just shorthand for the statement

$$p(y = c|\pi) = \pi_c. \tag{13}$$

The predictive $p(y|x,\theta)$ is then obtained by marginalizing π out under the meta-distribution:

$$p(y=c|x,\theta) \;=\; \int p(y=c|\pi)\, q(\pi|x,\theta)\, d\pi \;=\; \int \pi_c\, q(\pi|x,\theta)\, d\pi \;=\; \mathbb{E}_{q(\pi|x,\theta)}[\pi_c]. \tag{14}$$

That is, the class-c probability the model assigns to input x is the q-expectation of π_c.

Setup: entropies via Shannon. Both entropies that appear in the MI decomposition come from the Shannon formula $H = -\sum_c p_c \log p_c$ applied to a categorical distribution. Plugging the probability vector from (13) gives the conditional entropy, and plugging the probability vector from (14) gives the predictive entropy:

$$\mathrm{H}_{p(y|\pi)}(y) = -\sum_{c=1}^{C} \pi_c \log \pi_c, \tag{15}$$

$$\mathrm{H}_{p(y|x,\theta)}(y) = -\sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}[\pi_c] \;\log \mathbb{E}_{q(\pi|x,\theta)}[\pi_c]. \tag{16}$$

These are the only two entropies needed below.

Setup: Dirichlet moments. We use two moments of $\pi \sim \mathrm{Dir}(\alpha)$:

$$\mathbb{E}_{q(\pi|x,\theta)}[\pi_c] = \frac{\alpha_c}{\alpha_0}, \qquad \mathbb{E}_{q(\pi|x,\theta)}[\pi_c \log \pi_c] = \frac{\alpha_c}{\alpha_0}\big[\psi(\alpha_c+1) - \psi(\alpha_0+1)\big], \tag{17}$$

where ψ is the digamma function.

Total. Starting from the entropy definition of TU:

$$\begin{aligned}
\mathrm{TU}(x) &= \mathrm{H}_{p(y|x,\theta)}(y) && \text{(definition)} \\
&= -\sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}[\pi_c] \;\log \mathbb{E}_{q(\pi|x,\theta)}[\pi_c] && \text{(predictive entropy)} \\
&= -\sum_{c=1}^{C} \frac{\alpha_c}{\alpha_0} \log \frac{\alpha_c}{\alpha_0} && \text{(Dirichlet mean, eq. (17))}
\end{aligned}$$

Aleatoric. Starting from the entropy definition of AU:

$$\begin{aligned}
\mathrm{AU}(x) &= \mathbb{E}_{q(\pi|x,\theta)}\big[\mathrm{H}_{p(y|\pi)}(y)\big] && \text{(definition)} \\
&= \mathbb{E}_{q(\pi|x,\theta)}\left[-\sum_{c=1}^{C} \pi_c \log \pi_c\right] && \text{(conditional entropy, eq. (15))} \\
&= -\sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}[\pi_c \log \pi_c] && \text{(linearity)} \\
&= -\sum_{c=1}^{C} \frac{\alpha_c}{\alpha_0}\big[\psi(\alpha_c+1) - \psi(\alpha_0+1)\big] && \text{(Dirichlet log-moment, eq. (17))} \\
&= \sum_{c=1}^{C} \frac{\alpha_c}{\alpha_0}\big[\psi(\alpha_0+1) - \psi(\alpha_c+1)\big] && \text{(flip sign)}
\end{aligned}$$

Epistemic. Starting from the mutual-information definition of EU:

$$\begin{aligned}
\mathrm{EU}(x) &= I(y;\pi|x,\theta) && \text{(definition)}\\
&= \mathrm{TU}(x) - \mathrm{AU}(x) && \text{(MI decomposition)}\\
&= -\sum_{c=1}^{C} \frac{\alpha_c}{\alpha_0} \log \frac{\alpha_c}{\alpha_0} \; - \; \sum_{c=1}^{C} \frac{\alpha_c}{\alpha_0} \big[\psi(\alpha_0+1) - \psi(\alpha_c+1)\big] && \text{(substitute closed forms)}\\
&= \sum_{c=1}^{C} \frac{\alpha_c}{\alpha_0} \Big[-\log \frac{\alpha_c}{\alpha_0} - \psi(\alpha_0+1) + \psi(\alpha_c+1)\Big] && \text{(combine sums)}\\
&= \sum_{c=1}^{C} \frac{\alpha_c}{\alpha_0} \left[\log \frac{\alpha_0}{\alpha_c} + \psi(\alpha_c+1) - \psi(\alpha_0+1)\right] && \text{(flip log sign)}
\end{aligned}$$

B.2 BAYESIAN BASED UNCERTAINTY METRIC DERIVATIONS

Setup: the network and posterior. A backbone network f_θ produces an embedding that is fed through a (possibly RFF) feature map $\Phi(x) \in \mathbb{R}^R$ and a linear last-layer head with weights $\beta \in \mathbb{R}^{R\times C}$. After MAP training, a last-layer Laplace approximation places a Gaussian posterior over the head,

$$\beta \;\sim\; \mathcal{N}(\hat{\beta}, \hat{\Sigma}). \tag{18}$$

Each draw β maps to a point on the simplex via the per-class softmax

$$\pi_c(\beta, x) \;:=\; \big[\mathrm{softmax}(\Phi(x)^\top \beta)\big]_c, \tag{19}$$

so sampling β from the posterior (18) and applying (19) yields samples of π from the meta-distribution $q(\pi|x,\theta)$. This q has no analytic form, but because (19) is a deterministic function of β, any expectation or variance under $q(\pi|x,\theta)$ is just the corresponding expectation or variance under the weight posterior (18). We therefore estimate all $\mathbb{E}_{q(\pi|x,\theta)}$ and $\mathrm{Var}_{q(\pi|x,\theta)}$ terms by Monte Carlo over posterior draws of β below.

Setup: the label model. We assume $y|\pi \sim \mathrm{Cat}(\pi)$, i.e. $p(y=c|\pi) = \pi_c$. Marginally then $y_c|\pi_c \sim \mathrm{Bern}(\pi_c)$, giving

$$\mathbb{E}_{p(y_c|\pi_c)}[y_c] = \pi_c, \qquad \mathrm{Var}_{p(y_c|\pi_c)}(y_c) = \pi_c(1-\pi_c). \tag{20}$$

The predictive marginalizes π out under the meta-distribution:

$$p(y=c|x,\theta) \;=\; \int p(y=c|\pi)\, q(\pi|x,\theta)\, d\pi \;=\; \mathbb{E}_{q(\pi|x,\theta)}[\pi_c]. \tag{21}$$

Setup: Monte Carlo approximation. Draw $\beta^{(m)} \sim \mathcal{N}(\hat{\beta}, \hat{\Sigma})$ for $m = 1, \ldots, M$ and define

$$\pi_c^{(m)}(x) \;:=\; \pi_c(\beta^{(m)}, x), \qquad \bar{\pi}_c(x) \;:=\; \frac{1}{M}\sum_{m=1}^{M} \pi_c^{(m)}(x). \tag{22}$$

For any function g, $\mathbb{E}_{q(\pi|x,\theta)}[g(\pi)] \approx \frac{1}{M}\sum_m g(\pi^{(m)}(x))$; in particular $\mathbb{E}_{q(\pi|x,\theta)}[\pi_c] \approx \bar{\pi}_c(x)$.

Setup: entropies via Shannon. Plugging the categorical probability vectors from (20) (conditional on π) and (21) (marginal) into $H = -\sum_c p_c \log p_c$ gives the two entropies needed below:

$$\mathrm{H}_{p(y|\pi)}(y) = -\sum_{c=1}^{C} \pi_c \log \pi_c, \tag{23}$$

$$\mathrm{H}_{p(y|x,\theta)}(y) = -\sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}[\pi_c] \; \log \mathbb{E}_{q(\pi|x,\theta)}[\pi_c]. \tag{24}$$

Variance-based closed forms

Epistemic. Starting from the variance definition of EU:

$$\begin{aligned}
\mathrm{EU}(x) &= \sum_{c=1}^{C} \mathrm{Var}_{q(\pi|x,\theta)}\big(\mathbb{E}_{p(y_c|\pi_c)}[y_c]\big) && \text{(definition)} \\
&= \sum_{c=1}^{C} \mathrm{Var}_{q(\pi|x,\theta)}(\pi_c) && \text{(Bernoulli mean, eq. (20))} \\
&= \sum_{c=1}^{C} \Big(\mathbb{E}_{q(\pi|x,\theta)}[\pi_c^2] - \big(\mathbb{E}_{q(\pi|x,\theta)}[\pi_c]\big)^2\Big) && (\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2) \\
&\approx \sum_{c=1}^{C} \frac{1}{M-1} \sum_{m=1}^{M} \big(\pi_c^{(m)}(x) - \bar{\pi}_c(x)\big)^2 && \text{(MC approximation, eq. (22))}
\end{aligned}$$

The third line is the exact population variance via the identity $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, which corresponds to the biased ($\frac{1}{M}$) Monte Carlo estimator. In the final line we instead report the bias-corrected ($\frac{1}{M-1}$) estimator, matching the form used in the main text (Sec. 3). The two differ only by the $\frac{M}{M-1}$ factor and coincide as $M \to \infty$; we use the bias-corrected version throughout for consistency.

Aleatoric. Starting from the variance definition of AU:

$$\begin{aligned}
\mathrm{AU}(x) &= \sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}\big[\mathrm{Var}_{p(y_c|\pi_c)}(y_c)\big] && \text{(definition)} \\
&= \sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}\big[\pi_c(1-\pi_c)\big] && \text{(Bernoulli variance, eq. (20))} \\
&\approx \sum_{c=1}^{C} \frac{1}{M} \sum_{m=1}^{M} \pi_c^{(m)}(x)\big(1 - \pi_c^{(m)}(x)\big) && \text{(MC approximation, eq. (22))}
\end{aligned}$$

Total. Starting from the variance definition of TU:

$$\begin{aligned}
\mathrm{TU}(x) &= \sum_{c=1}^{C} \mathrm{Var}_{p(y|x,\theta)}(y_c) && \text{(definition)} \\
&= \mathrm{AU}(x) + \mathrm{EU}(x) && \text{(law of total variance)} \\
&= \sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}[\pi_c(1-\pi_c)] + \sum_{c=1}^{C} \mathrm{Var}_{q(\pi|x,\theta)}(\pi_c) && \text{(substitute closed forms)} \\
&= \sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}[\pi_c] - \sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}[\pi_c^2] + \sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}[\pi_c^2] - \sum_{c=1}^{C} \big(\mathbb{E}_{q(\pi|x,\theta)}[\pi_c]\big)^2 && \text{(expand)} \\
&= \sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}[\pi_c] - \sum_{c=1}^{C} \big(\mathbb{E}_{q(\pi|x,\theta)}[\pi_c]\big)^2 && \text{(cancel)} \\
&= 1 - \sum_{c=1}^{C} \big(\mathbb{E}_{q(\pi|x,\theta)}[\pi_c]\big)^2 && (\textstyle\sum_c \pi_c = 1, \text{ so } \sum_c \mathbb{E}_{q(\pi|x,\theta)}[\pi_c] = 1) \\
&\approx 1 - \sum_{c=1}^{C} \bar{\pi}_c(x)^2 && \text{(MC approximation)}
\end{aligned}$$

Entropy / Mutual Information closed forms

Total. Starting from the entropy definition of TU:

$$\begin{aligned}
\mathrm{TU}(x) &= \mathrm{H}_{p(y|x,\theta)}(y) && \text{(definition)} \\
&= -\sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}[\pi_c]\ \log \mathbb{E}_{q(\pi|x,\theta)}[\pi_c] && \text{(predictive entropy, eq. (24))} \\
&\approx -\sum_{c=1}^{C} \bar{\pi}_c(x)\ \log \bar{\pi}_c(x) && \text{(MC approximation, eq. (22))}
\end{aligned}$$

Aleatoric. Starting from the entropy definition of AU:

$$\begin{aligned}
\mathrm{AU}(x) &= \mathbb{E}_{q(\pi|x,\theta)}\big[\mathrm{H}_{p(y|\pi)}(y)\big] && \text{(definition)} \\
&= \mathbb{E}_{q(\pi|x,\theta)}\left[-\sum_{c=1}^{C} \pi_c \log \pi_c\right] && \text{(conditional entropy, eq. (23))} \\
&= -\sum_{c=1}^{C} \mathbb{E}_{q(\pi|x,\theta)}\big[\pi_c \log \pi_c\big] && \text{(linearity)} \\
&\approx -\frac{1}{M}\sum_{m=1}^{M}\sum_{c=1}^{C} \pi_c^{(m)}(x) \log \pi_c^{(m)}(x) && \text{(MC approximation, eq. (22))}
\end{aligned}$$

Epistemic. Starting from the mutual-information definition of EU:

$$\begin{aligned}
\mathrm{EU}(x) &= I(y;\pi|x,\theta) && \text{(definition)} \\
&= \mathrm{TU}(x) - \mathrm{AU}(x) && \text{(MI decomposition)} \\
&\approx -\sum_{c=1}^{C} \bar{\pi}_c(x)\ \log \bar{\pi}_c(x)\ -\ \left(-\frac{1}{M}\sum_{m=1}^{M}\sum_{c=1}^{C} \pi_c^{(m)}(x) \log \pi_c^{(m)}(x)\right) && \text{(substitute MC estimates)} \\
&= -\sum_{c=1}^{C} \bar{\pi}_c(x) \log \bar{\pi}_c(x) + \frac{1}{M}\sum_{m=1}^{M}\sum_{c=1}^{C} \pi_c^{(m)}(x) \log \pi_c^{(m)}(x) && \text{(simplify)}
\end{aligned}$$

C DETAILED EXPERIMENTS SECTION

C.1 EPISTEMIC V DATASET SIZE

C.1.1 General Setup

We evaluate epistemic uncertainty quantification across five models trained on a common feature representation of CIFAR-10. This experimental setup was chosen in order to isolate the contribution of each model and its uncertainty mechanism from the representation. All models operate on a frozen feature embedding rather than on raw pixels of the CIFAR-10 images. In practice, we extract a fixed 2048-dimensional embedding from a pretrained ResNet-50 backbone (Section C.1.3) and treat this embedding as the input space for every downstream model. These representations are held constant across all five methods, thus any difference in accuracy of epistemic uncertainty trends are attributable to the model architecture, training objective, and uncertainty methodology rather than to differences in the learned features. It also reduces each model to a comparatively small classifier over a 2048-dimensional input, which makes it tractable to train every configuration to convergence over a large hyperparameter grid and across ten random seeds for more robust results.

All five models share the same data pipeline, the same train/validation/test splits, the same optimizer family (Adam), and the same stopping protocol. They differ only in (i) the structure of the classifier head, (ii) the training objective, and (iii) the use of either Dirichlet or Laplace based uncertainty formation. The five models we look at in this work are the cross-entropy (MAP) baseline (CE Baseline), Evidential Deep Learning (EDL), the Random Fourier Feature (RFF) model, the Laplace model, and the combined RFF+Laplace model, each described in Section C.2. Here we document the shared infrastructure and then the cross-entropy baseline in full detail.

C.1.2 Datasets

We use CIFAR-10 as our in-distribution dataset to test epistemic trends.

CIFAR-10 (in-distribution). We construct training subsets of size n by sampling from the CIFAR-10 training split of size $50{,}000$ under a fixed random seed, holding out a validation split and using the standard CIFAR-10 test split for in-distribution evaluation. We include a train-val split in whatever n value so as to operate under the assumption that we are in a real-world learning scenario and have n total training samples. We utilize an 80/20 split for all dataset sizes as a reasonable implementation (i.e. for $n = 1000$ we would have 800 train and 200 val). The train-validation split is the same for all models.

C.1.3 Feature Embeddings

We extract embeddings using a ResNet-50 pretrained on ImageNet (`ResNet50_Weights.IMAGENET1K_V1`). We replace the final fully connected classification layer with an identity mapping (`model.fc` $=$ Identity), so that a forward pass returns a 2048-dimensional feature vector rather than ImageNet logits. The backbone is used purely as a fixed feature extractor, it is run in inference mode, its weights are never updated, and no gradients flow into it. We pass all the train and test CIFAR-10 data through the backbone once, and the resulting embeddings and labels are stacked and cached to disk. We then sample from these embeddings such that they are nested subsets, or in other words all $n = 100$ are contained within our $n = 50,000$ dataset.

Because the embeddings are cached, every downstream model and every hyperparameter configuration reads the *same* fixed 2048-dimensional inputs, and no model incurs the cost of re-running the backbone.

C.1.4 Shared Training and Stopping Protocol

Every model in the epistemic experiments must be trained to convergence. Thus because we cannot rely on validation we must come up with an alternative stopping mechanism during the training of each model. Shen et al.'s critique of EDL is only relevant in the converged training loss method and thus we want to ensure that all models are compared in this same vein. (The early-stopped, best-validation counterparts are reported separately for robustness of experimentation in Figure D.4.)

Optimizer. All models are optimized with ADAM. Weight decay in the optimizer is set to zero; any L_2-style regularization is supplied explicitly through the prior term in the loss objective rather than through the optimizer (see Section C.1.5). The learning rate is a swept hyperparameter for all models discussed later.

Stopping rule. We have two stopping rules, one that relies upon training loss and one that relies upon validation balanced accuracy:

- *Convergence mode* (used for the experiments reported here): training continues until the training loss plateaus, as measured by a rolling-mean convergence test over a window of W epochs with a relative-decrease threshold. We declare convergence when the relative decrease of the windowed-mean training loss falls below 10^{-6} over a window of $W = 50$ epochs.
- *Early Stopping mode* (val-balanced accuracy early stopping): training stops when the validation loss fails to improve. We always track the best-validation balanced accuracy at each epoch and keep track of the best performing model and return that as seen in Appendix D.

In convergence mode the maximum epoch budget is set very high ($10{,}000$) so that the plateau criterion, not the epoch cap, is what terminates training in practice.

C.1.5 Loss Objectives

Every model in this work is trained with one of two objectives: a maximum-a-posteriori (MAP) objective or the Evidential Deep Learning (EDL) objective. We define both here and then, in each model's description (Section C.2), simply state which objective it uses and with what settings.

MAP Objective The MAP objective is the categorical cross-entropy likelihood plus an isotropic zero-mean Gaussian prior over the weights if we are thinking in a Bayesian framing. This boils down to cross entropy plus L_2-Regularization. For a batch of size B with logits Wx_i and labels y_i, and a Gaussian prior of variance τ on the weights W,

$$\mathcal{L}_{\mathrm{MAP}}(W) = \underbrace{\frac{1}{B}\sum_{i=1}^{B} \mathrm{CE}\big(\mathrm{softmax}(Wx_i),\, y_i\big)}_{\text{mean cross-entropy}} \;-\; \frac{1}{n}\log p(W), \tag{25}$$

where n is the training-set size and the log-prior of the isotropic Gaussian is

$$\log p(W) = -\tfrac{1}{2}\Big[d\log(2\pi) + d\log\tau + \tfrac{1}{\tau}\|W\|^2\Big], \tag{26}$$

with d the number of weight parameters.

This is just cross-entropy + L_2 regularization. Dropping the W-independent constants in (26), the only term that depends on W is the quadratic $\frac{1}{2\tau}\|W\|^2$. The loss objective is thus:

$$\mathcal{L}_{\mathrm{MAP}}(W) = \frac{1}{B}\sum_{i=1}^{B} \mathrm{CE}\big(\mathrm{softmax}(Wx_i),\, y_i\big) \;+\; \underbrace{\frac{1}{2n\tau}}_{\lambda_{\text{eff}}}\|W\|^2 \;+\; \text{const}, \tag{27}$$

i.e. mean cross-entropy plus an L_2 penalty on the weights, with effective regularization strength

$$\lambda_{\text{eff}} = \frac{1}{2n\tau}. \tag{28}$$

The prior variance τ is therefore inversly related to the L_2 weight: large τ means a weak penalty, small τ means a strong penalty. The $1/n$ factor places the penalty on the same scale as the mean likelihood so we don't overweight the prior, and that λ_{eff} is correctly normalized for the dataset size n. We parameterize regularization through τ rather than through a raw weight-decay coefficient precisely so that this normalization is explicit, and we set the optimizer's own weight decay to zero so that all L_2 regularization comes from this prior term.

EDL Objective The EDL objective follows the sum-of-squares (Bayes-risk) directly from Sensoy et al. [2018]. While this may seem an odd choice for a classification task this was chosen by the original authors of EDL for its "empirical findings" and thus we decided to follow their implementation exactly. As stated in the main paper the whole crux of EDL methods is that rather than producing class probabilities directly, we design the network to output per-class evidence $e_c = g(\phi_c) \geq 0$ via a non-negative activation g, which parametrizes a Dirichlet meta-distribution through its concentrations

$$\alpha_c = e_c + 1, \qquad \alpha_0 = \sum_{c=1}^{C} \alpha_c, \tag{29}$$

so that $q(\pi|x) = \mathrm{Dir}(\alpha)$. The training loss is the expected sum of squares between the one-hot label y and the predicted probabilities π under this Dirichlet. We can write this as follows:

$$\mathcal{L}_{\mathrm{SSE}}(x) = \mathbb{E}_q\big[\|y - \pi\|_2^2\big] = \sum_{c=1}^{C}\Big((y_c - \mathbb{E}_q[\pi_c])^2 + \mathrm{Var}_q(\pi_c)\Big), \tag{30}$$

where the Dirichlet moments are

$$\mathbb{E}_q[\pi_c] = \frac{\alpha_c}{\alpha_0}, \qquad \mathbb{E}_q[\pi_c^2] = \frac{\alpha_c(\alpha_c+1)}{\alpha_0(\alpha_0+1)}. \tag{31}$$

Following, Sensoy et al. we add a Kullback–Leibler regularizer to our loss function that penalizes evidence assigned to the *incorrect* classes by pulling their concentrations back toward the uniform Dirichlet $\mathrm{Dir}(\mathbf{1})$ creating a single spiked distribution. Defining the misleading-evidence-removed concentrations $\tilde{\alpha} = y + (1-y)\odot\alpha$, the full objective is

$$\mathcal{L}_{\mathrm{EDL}}(x) = \mathcal{L}_{\mathrm{SSE}}(x) \;+\; \lambda_t\,\mathrm{KL}\big[\,\mathrm{Dir}(\tilde{\alpha})\,\big\|\,\mathrm{Dir}(\mathbf{1})\,\big], \tag{32}$$

where $\mathbf{1}$ is the all-ones concentration vector (the uniform Dirichlet over the simplex). Following Sensoy et al. [2018], the regularization coefficient is annealed over training,

$$\lambda_t = \lambda \cdot \min\Big(1,\ \tfrac{t}{T_{\text{anneal}}}\Big), \tag{33}$$

with t the current epoch and T_{anneal} the annealing temperature. The annealing prevents the KL term from dominating early in training, before the network has accumulated meaningful evidence: it lets the model first fit the data and only gradually enforces the regularization. This is consistent with the training methodology as laid out by Shen et al. [2024]. The evidence activation g is treated as a hyperparameter (ReLU, softplus, or a clamped exponential); ReLU is the default used in the original paper during training.

C.2 MODELS:

C.2.1 Cross-Entropy (MAP) Baseline

The cross-entropy baseline is a single linear classifier mapping the 2048-dimensional ResNet-50 embedding directly to $C = 10$ class logits, logits $= Wx$ with $W \in \mathbb{R}^{10\times 2048}$, with class probabilities obtained by a softmax over the logits. Because the model is linear in a fixed feature space, this baseline is effectively multinomial logistic regression on top of the frozen embedding. It is trained with the MAP objective of Section C.1.5, where the prior variance τ is a swept hyperparameter within each run and the optimizer's own weight decay is set to zero, so that all regularization flows through the prior term. This model serves as the reference point against which the uncertainty-aware models are compared: it produces a single point estimate of the class probabilities and carries no explicit mechanism for epistemic uncertainty beyond softmax confidence.

The baseline is also the foundation for the Laplace model (Section C.2.4), which reuses these trained checkpoints and the same prior variance τ to fit a last-layer posterior post-hoc. We sweep τ over three values, together with four learning rates, four dataset sizes, and ten random seeds; the full grid is given in Table C.1. As with all models, every configuration is trained to convergence and both the best-validation and final-converged snapshots are retained, with each run written to its own result directory keyed by its four hyperparameters so that runs are individually addressable and reproducible.

Hyperparameter	Values
lrule Training-set size n	$100,\ 1000,\ 10000,\ 50000$
Random seed	$1001, 2001, \ldots, 10001$ (ten seeds)
Learning rate	$0.1,\ 0.01,\ 0.001,\ 0.0001$
Prior variance τ	$0.1,\ 1.0,\ 10.0$

Table C.1: Hyperparameter grid for the cross-entropy (MAP) baseline. Every combination is trained independently to convergence, for a total of 480 runs.

Each run follows the shared training and stopping protocol of Section C.1.4: per epoch we take one optimization pass over the training loader under the MAP objective, evaluate the unregularized cross-entropy on the validation loader for tracking (so the validation curve reflects predictive fit rather than the penalized objective), and record the learning rate, epoch, training loss, validation loss, and validation balanced accuracy in a per-run history table. The best-validation balanced accuracy snapshot is updated whenever the validation balanced accuracy reaches a new maximum, and the convergence criterion is tested on the training-loss history; training halts when the training loss plateaus under the $W = 50$, threshold $= 10^{-6}$ rolling test, or at the 10,000-epoch cap. When training stops we snapshot the final converged parameters, evaluate both the best-validation and final-converged parameter sets on all four loaders, and persist each as a separate set of result files. Runs are dispatched through a SLURM batch script with one job per configuration, each requesting a single GPU, 32 GB of memory, and a six-hour wall-clock limit under a CUDA-enabled environment.

C.2.2 Evidential Deep Learning Model

The EDL model uses the same architecture as the cross-entropy baseline, a single linear head mapping the 2048-dimensional ResNet-50 embedding to $C = 10$ output, but reinterprets those outputs and trains them under a different objective. Rather than producing logits that are softmax-normalized into a single probability vector, the head produces per-class *evidence*

$e_c = g(\phi_c) \geq 0$ via a non-negative activation g, which parameterizes a Dirichlet meta-distribution through the concentrations $\alpha_c = e_c + 1$ (Eq. 29). The model is trained with the EDL objective of Section C.1.5—the expected sum-of-squares data term plus the annealed KL regularizer toward the uniform Dirichlet. Unlike the other four models, EDL is therefore the only method whose epistemic uncertainty is routed through Dirichlet strength: its uncertainty estimates are read off directly from the predicted concentrations $\alpha(x)$ at a single forward pass, with no weight-space posterior and no Monte Carlo sampling.

The regularization weight λ (Eq. 32) is the principal EDL-specific hyperparameter and is swept over a comparatively wide range, since it controls the strength with which misleading evidence is suppressed and, as discussed in Section A.2, sets the finite Dirichlet strength that the loss drives the network toward. We use the ReLU evidence activation of the original formulation and anneal λ linearly over the first T_{anneal} epochs. The hyperparameter grid is given in Table C.2; as with all models, every configuration is trained to convergence across ten random seeds and both the best-validation and final-converged snapshots are retained.

Hyperparameter	Values
lrule Training-set size n	$100,\ 1000,\ 10000,\ 50000$
Random seed	$1001, 2001, \ldots, 10001$ (ten seeds)
Learning rate	$0.1,\ 0.01,\ 0.001,\ 0.0001$
Regularization weight λ	$0.0001,\ 0.001,\ 0.01,\ 0.1,\ 0.5,\ 1.0$

Table C.2: Hyperparameter grid for the EDL model. Every combination is trained independently to convergence.

C.2.3 Random Fourier Feature Model

The RFF model maps the 2048-dimensional embedding through a random Fourier feature expansion $\Phi(x) \in \mathbb{R}^R$ that approximates an RBF kernel, followed by a bias-free linear head from the R-dimensional feature space to the $C = 10$ logits. The feature map draws its frequencies ω and phases b once at initialization and holds them fixed, so that

$$\Phi(x) = \sigma_{\text{out}} \sqrt{\tfrac{2}{R}} \cos\Big(\tfrac{1}{\ell}\,\omega\, x + b\Big), \tag{34}$$

with lengthscale ℓ and outputscale σ_{out} as kernel hyperparameters; only the linear head is trained, under the MAP objective of Section C.1.5.

Importantly, the RFF model and the RFF–Laplace model share the *same trained checkpoint*: they differ only at evaluation time. For the RFF model we take the trained `RFFLaplace` network and evaluate it as a *plain softmax classifier*—we apply $\mathrm{softmax}(\Phi(x)^\top \beta)$ at the point estimate $\hat{\beta}$ and do *not* fit or apply the Laplace covariance, and we draw no Monte Carlo samples. Epistemic uncertainty for this model is therefore computed from a single deterministic forward pass, using the same Dirichlet-style metrics applied to the point logits. The RFF model thus isolates the contribution of the kernel feature map alone, separate from the weight-space posterior that the RFF–Laplace model adds on top. Its hyperparameter grid is given in Table C.3.

Hyperparameter	Values
lrule Training-set size n	$100,\ 1000,\ 10000,\ 50000$
Random seed	$1001, 2001, \ldots, 10001$ (ten seeds)
Feature rank R	2048
Lengthscale ℓ	20.0
Outputscale σ_{out}	1.0
Learning rate	$0.1,\ 0.01,\ 0.001,\ 0.0001$
Prior variance τ	1.0

Table C.3: Hyperparameter grid shared by the RFF and RFF–Laplace models; the two differ only at evaluation time (plain softmax vs. Laplace posterior).

C.2.4 Laplace Model

The Laplace model adds a last-layer Laplace approximation *on top of the already trained cross-entropy baseline*; it introduces no new training. We take the converged linear head $\hat{\beta}$ from a CE-baseline run (Section C.2, the cross-entropy baseline) and fit a Gaussian posterior $\beta \sim \mathcal{N}(\hat{\beta}, \hat{\Sigma})$ over its weights, where the posterior precision is the Gaussian-prior precision $\frac{1}{\tau} I$ plus the data Fisher information accumulated over the training set,

$$\hat{\Sigma}^{-1} = \tfrac{1}{\tau} I + \sum_i \pi_i (1 - \pi_i)\, \phi(x_i) \phi(x_i)^\top, \tag{35}$$

computed per class from the training embeddings $\phi(x_i)$ and the baseline's predicted probabilities π_i. The prior variance τ reused here is the same Gaussian-prior variance that defined the baseline's L_2 regularization (Eq. 28), keeping the prior consistent between training and the posterior fit.

At prediction time we draw Monte Carlo samples of the logits from the induced predictive Gaussian, $\beta^{(m)} \sim \mathcal{N}(\hat{\beta}, \hat{\Sigma})$, push each through the softmax, and measure epistemic uncertainty as the spread of the resulting probability vectors across draws—the Bayesian mechanism described in Section B.2, where vanishing epistemic uncertainty corresponds to $\hat{\Sigma} \to 0$ rather than to diverging logits. Because the Laplace model is fit post-hoc to existing checkpoints, it inherits the baseline's hyperparameter grid exactly (Table C.1); the only model-specific evaluation setting is the number of posterior samples, which we set to $10{,}000$ to keep the Monte Carlo estimate of the epistemic metrics low-variance.

C.2.5 Random Fourier Features + Laplace Model

The RFF–Laplace model is the full Bayesian model: the random Fourier feature map of Section C.2.3 *combined with* a last-layer Laplace posterior over the linear head. It uses the identical trained checkpoint as the RFF model (Table C.3); the distinction is purely at evaluation. Whereas the RFF model discards the covariance and evaluates a plain softmax, the RFF–Laplace model fits the Gaussian posterior $\beta \sim \mathcal{N}(\hat{\beta}, \hat{\Sigma})$ over the head weights—with precision $\frac{1}{\tau} I$ plus the Fisher information accumulated over the RFF-transformed training features—and then draws Monte Carlo samples to form the predictive meta-distribution

$$\pi^{(m)}(x) = \mathrm{softmax}\big(\Phi(x)^\top \beta^{(m)}\big), \qquad \beta^{(m)} \sim \mathcal{N}(\hat{\beta}, \hat{\Sigma}). \tag{36}$$

Epistemic uncertainty is then the spread of $\pi^{(m)}$ across draws, computed with the Monte Carlo estimators of Section B.2, again using $10{,}000$ posterior samples.

This pairing is the natural counterpart to the two single-mechanism models: relative to the RFF model it adds the weight-space posterior, and relative to the Laplace model it adds the kernel feature map. As discussed in Section A.2, this is the model whose epistemic uncertainty converges toward zero with increasing N—driven by $\hat{\Sigma} \to 0$—in contrast to the finite floor that EDL plateaus at, which is the central comparison of this work.

C.3 OOD AUROC DETECTION

Having characterized how each method's in-distribution epistemic uncertainty behaves as a function of dataset size (Section C.1), we now ask a complementary question: do these epistemic uncertainty estimates actually *separate* in-distribution inputs from out-of-distribution inputs? A well-behaved epistemic uncertainty measure should assign systematically higher uncertainty to OOD inputs than to ID inputs, so that thresholding on the uncertainty score recovers an OOD detector. We quantify this separation with the area under the receiver operating characteristic curve (AUROC), treating each epistemic metric in turn as the detection score.

C.3.1 Shared Setup with the Epistemic Experiments

The OOD detection experiments reuse the exact infrastructure of the epistemic study, so that the two sets of results are directly comparable. In particular, they share:

- **The same frozen feature representation.** Every input—ID or OOD—is first passed through the same fixed 2048-dimensional ResNet-50 embedding described in Section C.1.3 (`ResNet50_Weights.IMAGENET1K_V1` with the

final classification layer replaced by an identity map). The backbone is run purely as a feature extractor in inference mode; its weights are never updated and no gradients flow into it. Crucially, the OOD images are embedded through this *same* backbone, so that OOD detection is performed entirely in the common 2048-dimensional feature space rather than on raw pixels.

- **The same five models.** We evaluate the cross-entropy (MAP) baseline, Evidential Deep Learning (EDL), the Random Fourier Feature (RFF) model, the Laplace model, and the combined RFF–Laplace model, each defined in Section C.2. The non-Bayesian models (CE, EDL, RFF) read their epistemic metrics off a single deterministic forward pass, while the Bayesian models (Laplace, RFF–Laplace) draw $10{,}000$ Monte Carlo samples from the last-layer Gaussian posterior to form the predictive meta-distribution. For the Bayesian models we use the Monte Carlo variants of Mutual Information and Variance of Expectation, and retain the Dirichlet-style Subjective Logic metric.
- **The same hyperparameter grids and seeds.** Each model is swept over the grid given in its model description (Tables C.1–C.3), over the same four dataset sizes $n \in \{100, 1000, 10000, 50000\}$ and the same ten random seeds $\{1001, 2001, \ldots, 10001\}$.

Model selection: best validation balanced accuracy. Unlike the train-to-convergence epistemic plots, the OOD experiments report the *early-stopped, best-validation balanced accuracy* snapshot of each run. Concretely, for each $(\text{method}, n, \text{seed})$ cell we select a single hyperparameter configuration by maximizing the validation balanced accuracy (`val_balanced_accuracy`). Ties are broken in favor of stronger regularization: first the smaller outputscale σ_{out}, then the smaller prior variance τ, on the principle that among configurations that fit the validation data equally well we prefer the simpler (more heavily regularized) one. The metrics are then read from the best-validation snapshot of the selected configuration rather than from its converged snapshot. This is the selection rule used consistently across all OOD and correlation results.

C.3.2 In-Distribution and Out-of-Distribution Data

The in-distribution data is the standard CIFAR-10 test split, embedded through the frozen backbone. The ID uncertainty scores are computed on the test-set logits saved at the best-validation snapshot of the selected run.

We use five out-of-distribution datasets, each precomputed once into the same 2048-dimensional embedding space and cached to disk:

- **CIFAR-100** — natural images from disjoint semantic classes, a "near-OOD" set sharing CIFAR-10's low-level statistics;
- **SVHN** — street-view house numbers;
- **TinyImageNet** — a downsampled subset of ImageNet classes;
- **FashionMNIST** — grayscale clothing images;
- **CIFAR10_noise** — a synthetic random noise, where there is no semantic information as a baseline for far-OOD.

Each OOD set is loaded as a feature tensor and labels are not required for the detection task (they default to a placeholder), since the only quantity of interest is each sample's epistemic uncertainty score. Each OOD set contains 10,000 samples, matching the size of the CIFAR-10 test split used as the in-distribution reference.

C.3.3 Metrics and AUROC Computation

For every model we compute the same three epistemic uncertainty metrics used throughout this work: Subjective Logic uncertainty mass, Mutual Information, and Variance of Expectation (with the Monte Carlo estimators of Section B.2 substituted for the Bayesian models). Each metric produces a single per-sample epistemic score.

For a given (model, n, seed, OOD dataset, metric) combination, we form the binary detection problem in which ID test samples are the negative class and OOD samples are the positive class. Treating the epistemic uncertainty score directly as the detector output—higher score $\Rightarrow$ more likely OOD—we compute the AUROC over the pooled ID/OOD scores. Any non-finite scores are sanitized to zero before scoring. The result is one AUROC value per metric, per OOD dataset, for each selected run; an AUROC of 0.5 indicates the metric cannot distinguish ID from OOD, while values approaching 1.0 indicate near-perfect separation. We additionally save the full ROC curve (FPR, TPR) for each metric so that the operating characteristics can be inspected directly.

C.4 CORRELATION ANALYSIS

The three epistemic metrics—Subjective Logic, Mutual Information, and Variance of Expectation—are nominally distinct quantities derived from different decompositions of predictive uncertainty (Section B). A natural question is whether, in practice, they carry the same information for the purpose of OOD detection: if two metrics induce essentially the same ranking of inputs, they will produce the same AUROC and one of them is redundant. The correlation analysis quantifies the degree to which the three metrics agree, both overall and within finer slices of the experimental grid.

C.4.1 Setup and Inputs

The correlation analysis operates on exactly the same selected runs as the OOD detection experiments: the same five models, the same four dataset sizes, the same ten seeds, and the *same* best-validation-balanced-accuracy model selection rule (with the identical outputscale-then-τ tiebreak). Rather than correlating per-sample scores, we correlate at the level of *AUROC values*: each selected run contributes, for each OOD dataset, a triple of AUROCs—one for Subjective Logic, one for Mutual Information, one for Variance of Expectation. A single "dot" in the analysis is therefore one such triple, and we ask how tightly the three coordinates of these dots move together across the experiment. Each method is plotted in its own color so that method-specific structure remains visible in the pooled views.

C.4.2 Data Slices

To check whether metric agreement is uniform or varies with the experimental conditions, we recompute the analysis over several slices of the grid:

- **Per OOD dataset.** Fixing one OOD dataset and pooling over n and seeds yields one dot per (method, n, seed).
- **Per dataset size n.** Fixing one n and pooling over seeds and OOD datasets yields up to $10 \times 5 = 50$ dots per method (ten seeds times five OOD sets).
- **Averaged over OOD.** Averaging each run's AUROC across the five OOD datasets gives one dot per (method, n, seed).
- **Pooled.** Treating every (method, n, seed, OOD) combination as its own dot gives the fully pooled view.

C.4.3 Quantities Reported

Visually, each slice is rendered as scatter plots of one metric's AUROC against another's, with a $y = x$ reference line indicating perfect agreement.

D ADDITIONAL RESULTS AND PLOTS

D.1 UNSCALED EPISTEMIC PLOTS

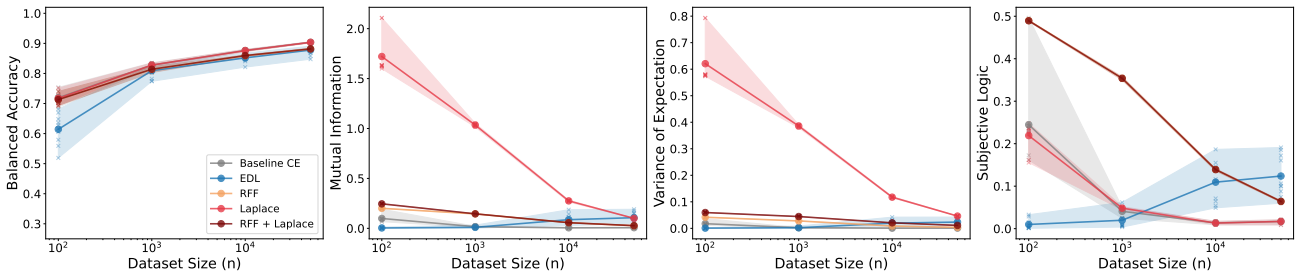

Figure D.1: Epistemic Uncertainty metrics on CIFAR-10 test split as a function of dataset size increase for different models trained on nested subsets. Each model is trained to convergence as described in Appendix A. Laplace is unscaled in this plot unlike in the main paper so you can see the original performance. Of note, both Baseline CE and RFF are overlapped by Laplace and RFF + Laplace in the subjective logic case.

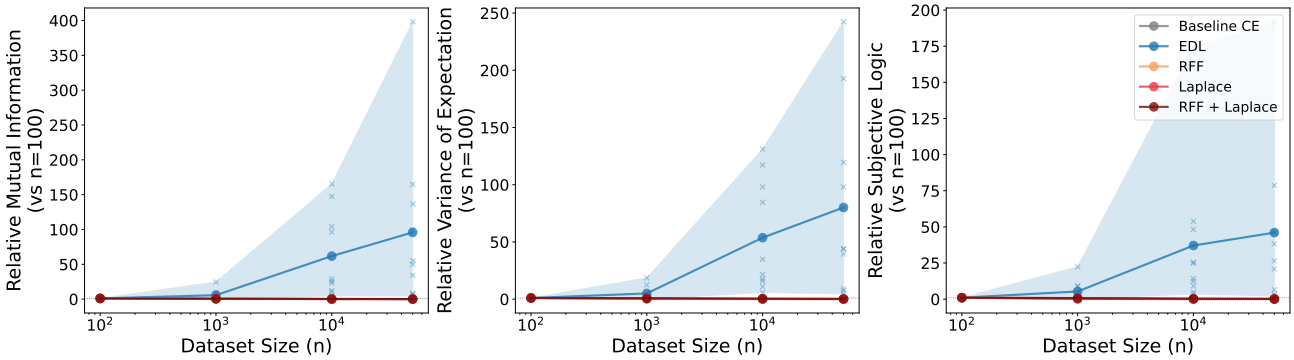

Figure D.2: Relative epistemic uncertainty metrics on the CIFAR-10 test split as a function of dataset size for different models trained on nested subsets. Each metric is normalized per-seed by its value at n=100, so all curves start at 1.0 and lower values indicate a larger proportional decrease in epistemic uncertainty as dataset size increases. Each model is trained to convergence as described in Appendix C.

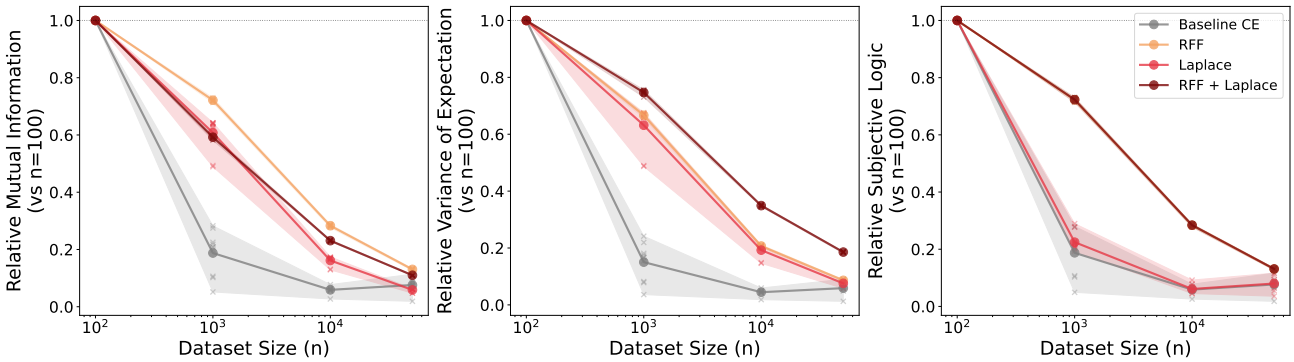

Figure D.3: Relative epistemic uncertainty metrics on the CIFAR-10 test split as a function of dataset size for different models trained on nested subsets. Each metric is normalized per-seed by its value at n=100, so all curves start at 1.0 and lower values indicate a larger proportional decrease in epistemic uncertainty as dataset size increases. Each model is trained to convergence as described in Appendix C. EDL has been removed due to its consistent increase and masking the overall trend.

D.2 EPISTEMIC PLOTS WITH EARLY STOPPING

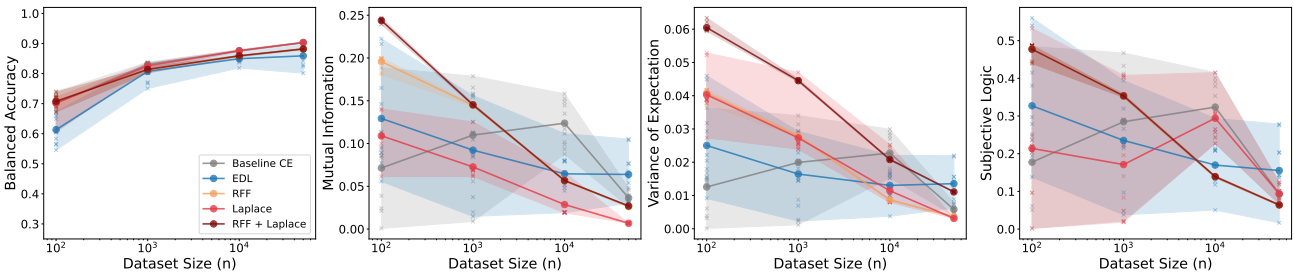

Figure D.4: Epistemic Uncertainty metrics on CIFAR-10 test split as a function of dataset size increase for different models trained on nested subsets. Each model is trained and early stopped with best validation accuracy as a metric metric. Of note, Laplace has been scaled to fit plot for Mutual Information and Variance of expectation by dividing by 15. It is interesting to note that only the CE-baseline and EDL change trends between the converged and non-converged models. EDL when early stopped is not subject to the Shen et. al critique.

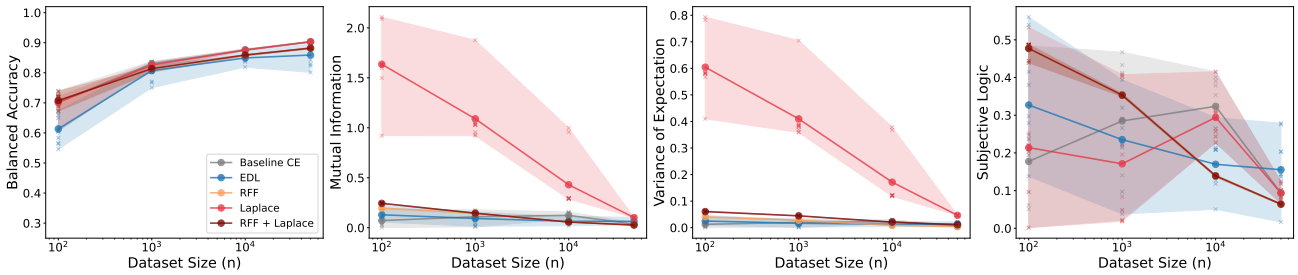

Figure D.5: Epistemic Uncertainty metrics on CIFAR-10 test split as a function of dataset size increase for different models trained on nested subsets. Each model is trained and early stopped with best validation accuracy as a metric metric.

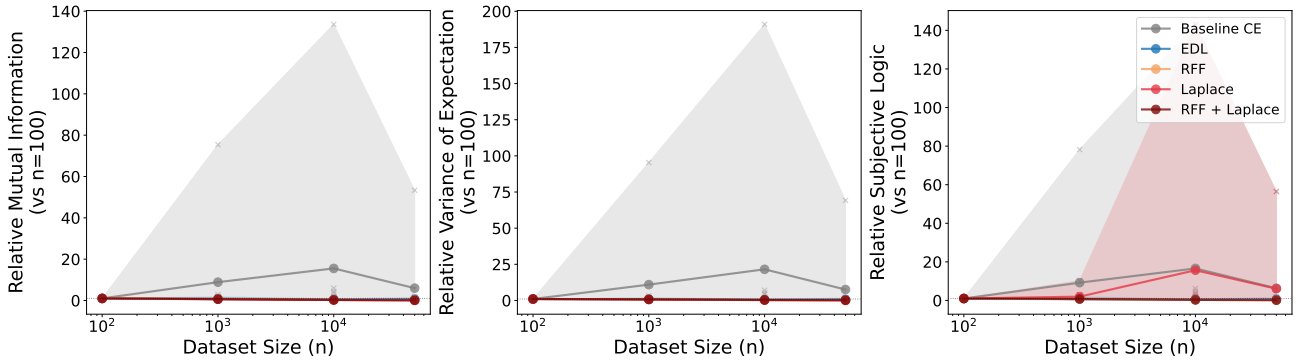

Figure D.6: Relative epistemic uncertainty metrics on the CIFAR-10 test split as a function of dataset size for different models trained on nested subsets. Each metric is normalized per-seed by its value at n=100, so all curves start at 1.0 and lower values indicate a larger proportional decrease in epistemic uncertainty as dataset size increases. Each model is trained and early stopped with best validation accuracy as the metric as described in Appendix C.

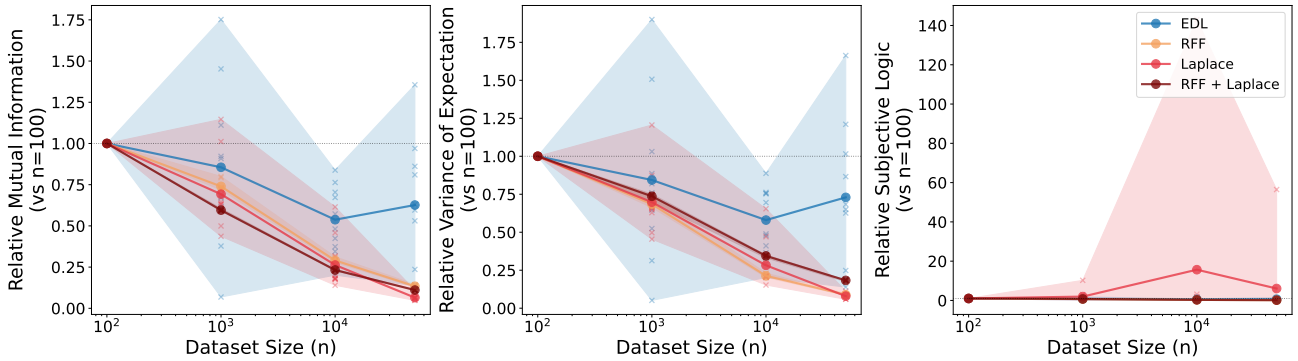

Figure D.7: Relative epistemic uncertainty metrics on the CIFAR-10 test split as a function of dataset size for different models trained on nested subsets. Each metric is normalized per-seed by its value at n=100, so all curves start at 1.0 and lower values indicate a larger proportional decrease in epistemic uncertainty as dataset size increases. Each model is trained and early stopped with best validation accuracy as the metric as described in Appendix C. CE Baseline has been removed due to its increase and masking the overall trend.

D.3 ADDITIONAL OOD DETECTION PLOTS

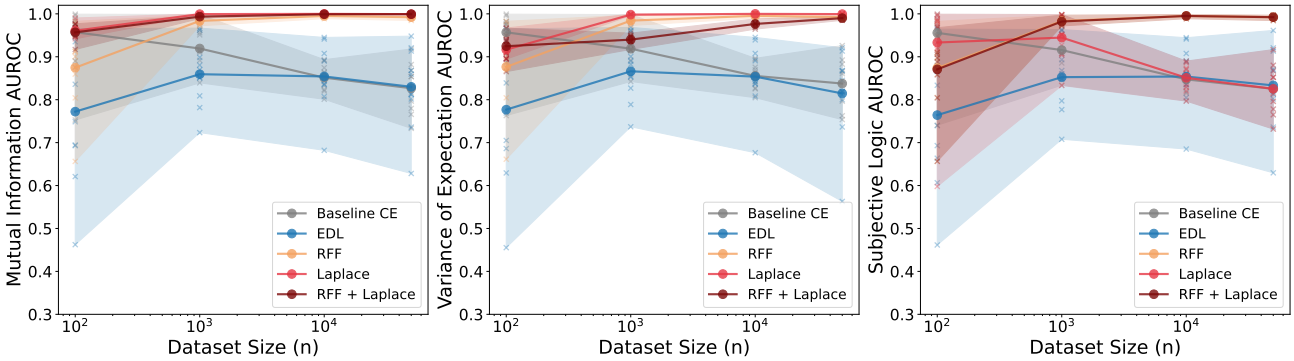

Figure D.8: AUROC for detecting random-noise (OOD) against CIFAR-10 (ID) as a function of dataset size increase, shown across all three epistemic uncertainty metrics for each model. ID is the CIFAR-10 test split and pure noise images embeddings are the OOD positive class, both embedded through the same frozen ResNet-50 backbone. Curves show the mean across ten seeds with the min/max envelope shaded, evaluated at the best-validation-balanced-accuracy snapshot of each run.

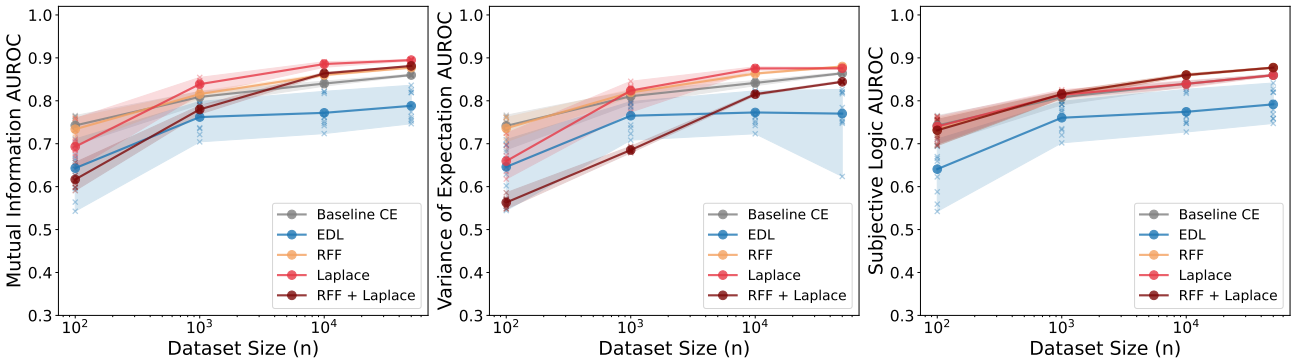

Figure D.9: AUROC for detecting CIFAR-100 (OOD) against CIFAR-10 (ID) as a function of dataset size increase, shown across all three epistemic uncertainty metrics for each model. ID is the CIFAR-10 test split and CIFAR-100 samples are the OOD positive class, both embedded through the same frozen ResNet-50 backbone. Curves show the mean across ten seeds with the min/max envelope shaded, evaluated at the best-validation-balanced-accuracy snapshot of each run.

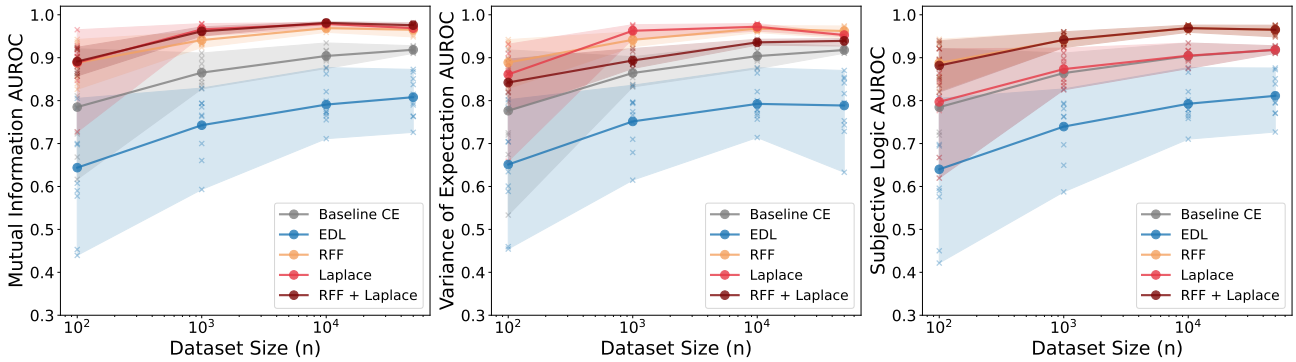

Figure D.10: AUROC for detecting FashionMNIST (OOD) against CIFAR-10 (ID) as a function of dataset size increase, shown across all three epistemic uncertainty metrics for each model. ID is the CIFAR-10 test split and FashionMNIST samples are the OOD positive class, both embedded through the same frozen ResNet-50 backbone. Curves show the mean across ten seeds with the min/max envelope shaded, evaluated at the best-validation-balanced-accuracy snapshot of each run.

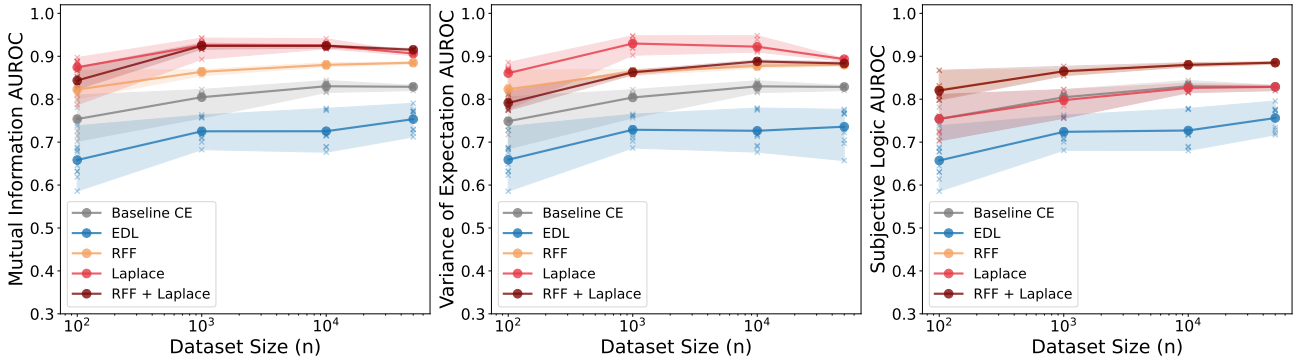

Figure D.11: AUROC for detecting TinyImageNet (OOD) against CIFAR-10 (ID) as a function of dataset size increase, shown across all three epistemic uncertainty metrics for each model. ID is the CIFAR-10 test split and TinyImageNet samples are the OOD positive class, both embedded through the same frozen ResNet-50 backbone. Curves show the mean across ten seeds with the min/max envelope shaded, evaluated at the best-validation-balanced-accuracy snapshot of each run.

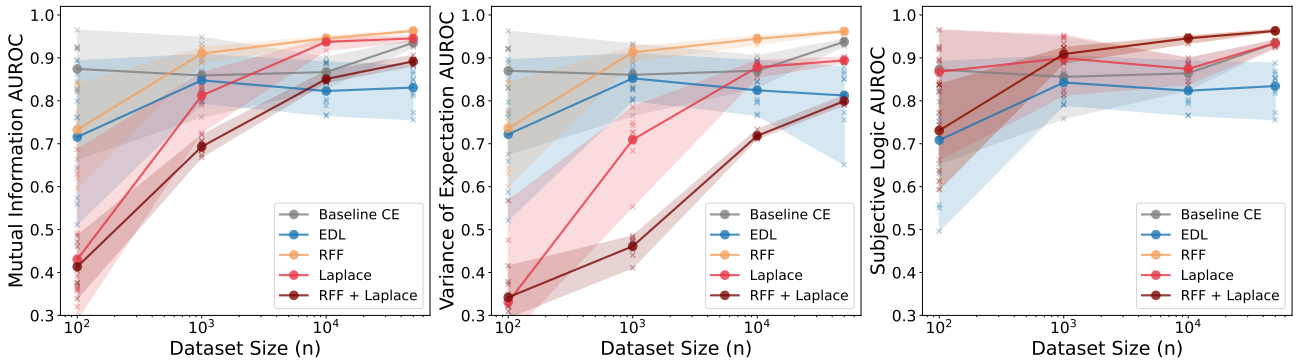

Figure D.12: AUROC for detecting SVHN (OOD) against CIFAR-10 (ID) as a function of dataset size increase, shown across all three epistemic uncertainty metrics for each model. ID is the CIFAR-10 test split and SVHN samples are the OOD positive class, both embedded through the same frozen ResNet-50 backbone. Curves show the mean across ten seeds with the min/max envelope shaded, evaluated at the best-validation-balanced-accuracy snapshot of each run.

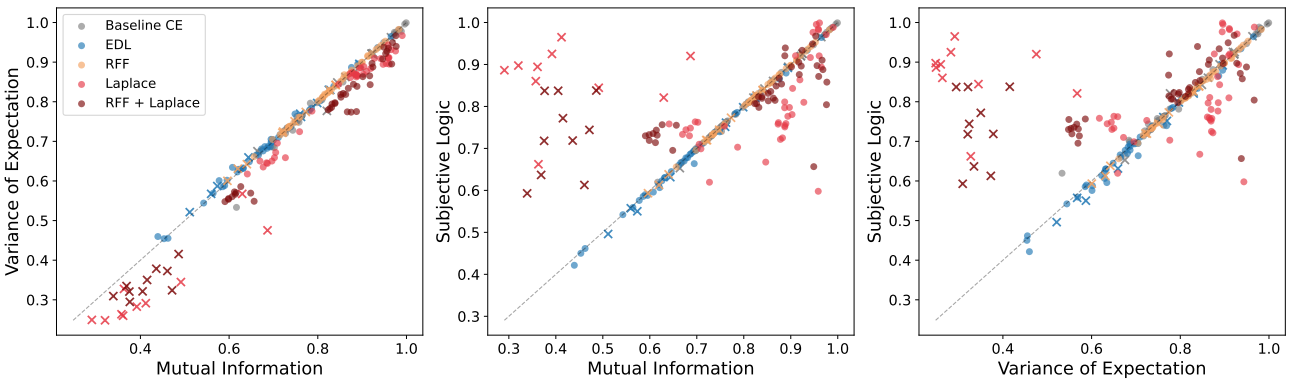

Figure D.13: Correlation analysis of the three epistemic metrics' AUROCs at $n = 100$, pooling over all five OOD datasets and ten seeds. Each dot is one dataset/model pair AUROC (Subjective Logic, Mutual Information, Variance of Expectation), scattered pairwise against a $y = x$ reference line and colored by method.

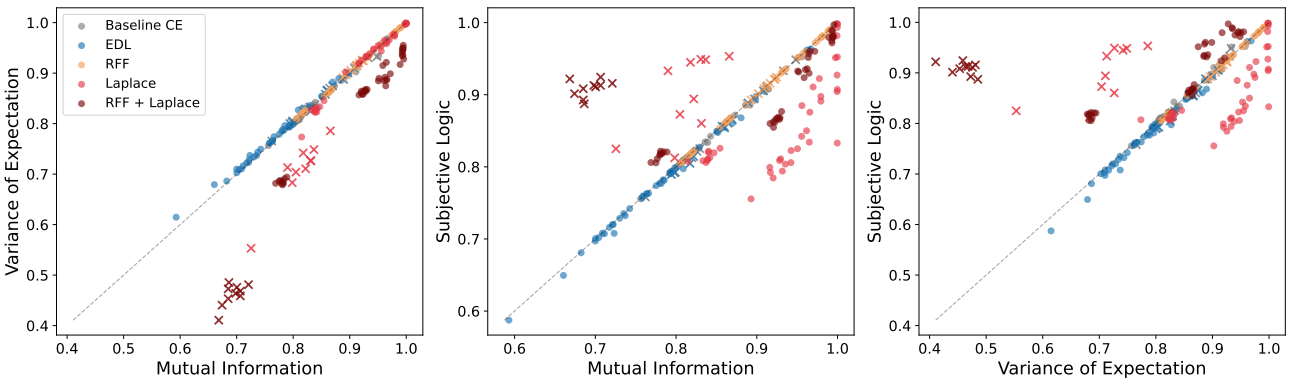

Figure D.14: Correlation analysis of the three epistemic metrics' AUROCs at $n = 1{,}000$, pooling over all five OOD datasets and ten seeds. Each dot is one dataset/model pair AUROC (Subjective Logic, Mutual Information, Variance of Expectation), scattered pairwise against a $y = x$ reference line and colored by method.

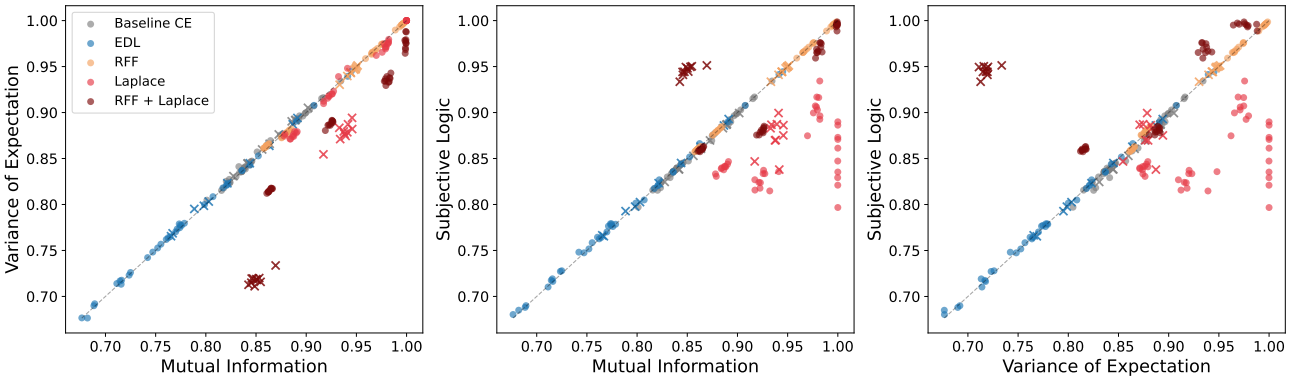

Figure D.15: Correlation analysis of the three epistemic metrics' AUROCs at $n = 10{,}000$, pooling over all five OOD datasets and ten seeds. Each dot is one dataset/model pair AUROC (Subjective Logic, Mutual Information, Variance of Expectation), scattered pairwise against a $y = x$ reference line and colored by method.

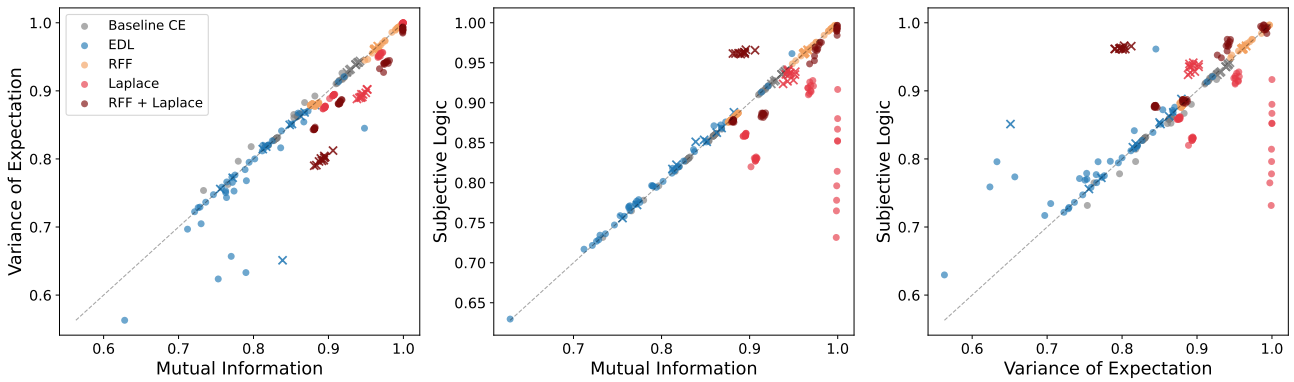

Figure D.16: Correlation analysis of the three epistemic metrics' AUROCs at $n = 50{,}000$, pooling over all five OOD datasets and ten seeds. Each dot is one dataset/model pair AUROC (Subjective Logic, Mutual Information, Variance of Expectation), scattered pairwise against a $y = x$ reference line and colored by method.

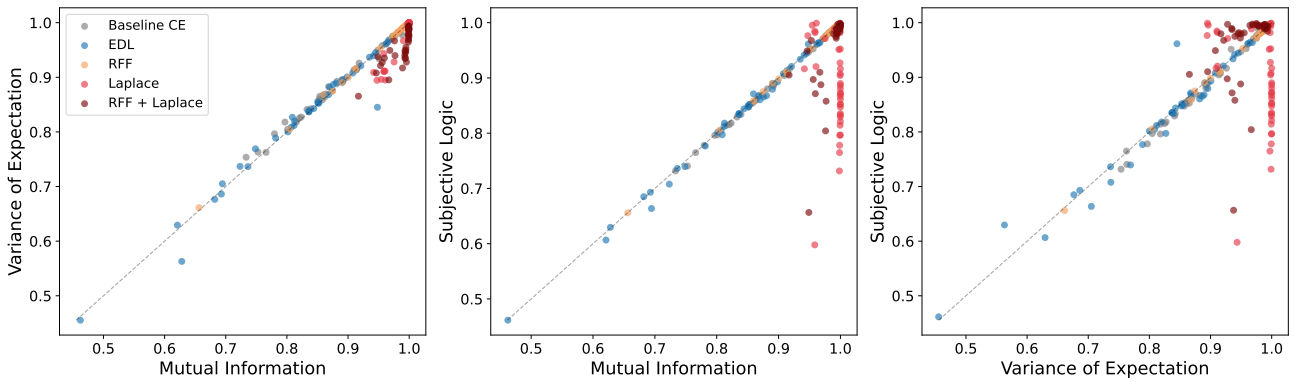

Figure D.17: Correlation analysis of the three epistemic metrics' AUROCs for the random-noise OOD set, pooling over all four dataset sizes and ten seeds. Each dot is one run's triple of AUROCs (Subjective Logic, Mutual Information, Variance of Expectation), scattered pairwise against a $x = y$ reference line and colored by method.

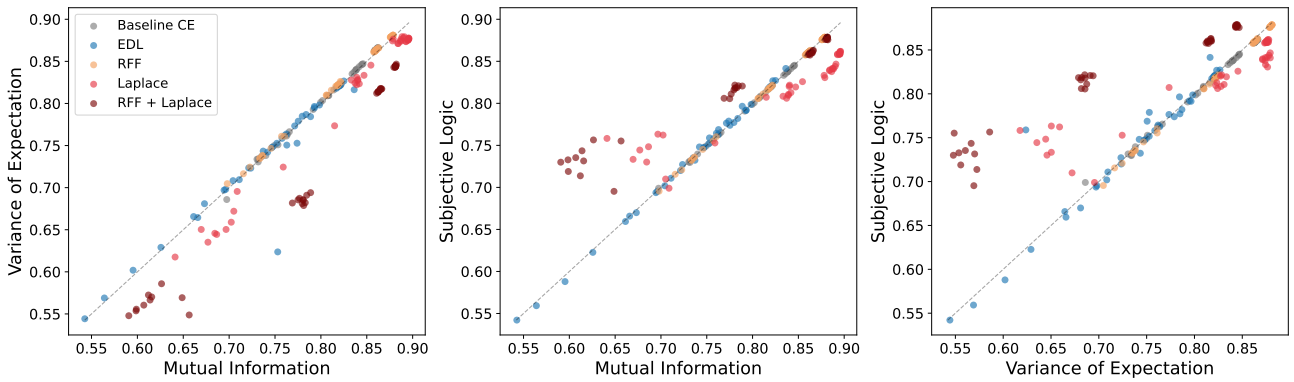

Figure D.18: Correlation analysis of the three epistemic metrics' AUROCs for the CIFAR-100 OOD set, pooling over all four dataset sizes and ten seeds. Each dot is one run's triple of AUROCs (Subjective Logic, Mutual Information, Variance of Expectation), scattered pairwise against a $x = y$ reference line and colored by method.

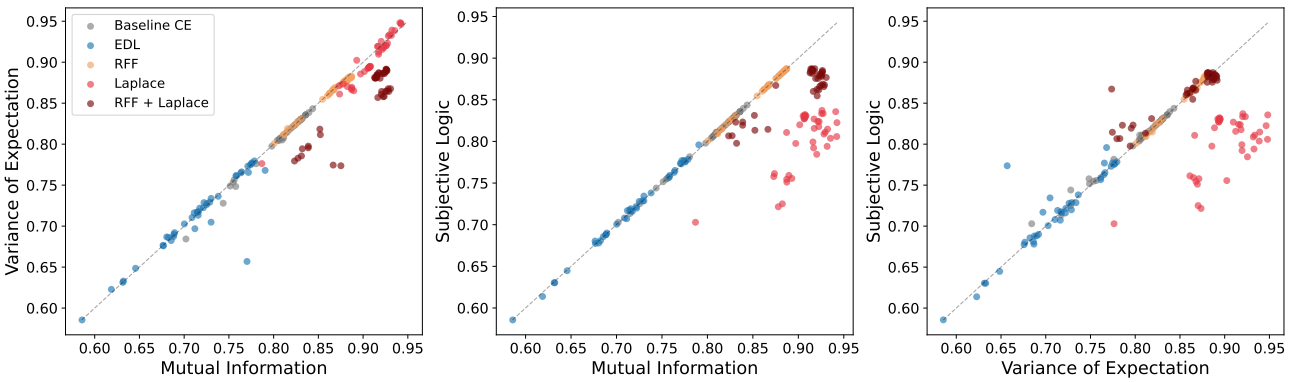

Figure D.19: Correlation analysis of the three epistemic metrics' AUROCs for the TinyImageNet OOD set, pooling over all four dataset sizes and ten seeds. Each dot is one run's triple of AUROCs (Subjective Logic, Mutual Information, Variance of Expectation), scattered pairwise against a $x = y$ reference line and colored by method.

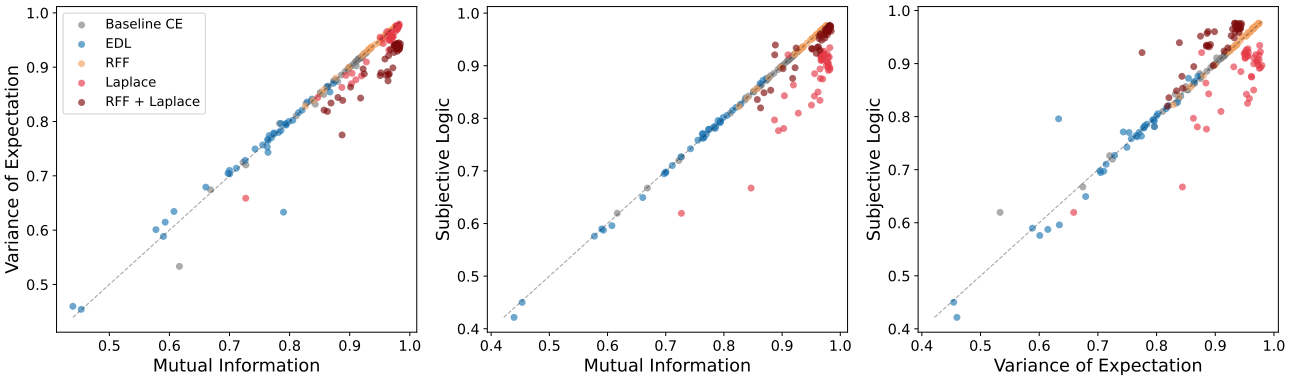

Figure D.20: Correlation analysis of the three epistemic metrics' AUROCs for the FashionMNIST OOD set, pooling over all four dataset sizes and ten seeds. Each dot is one run's triple of AUROCs (Subjective Logic, Mutual Information, Variance of Expectation), scattered pairwise against a $x = y$ reference line and colored by method.

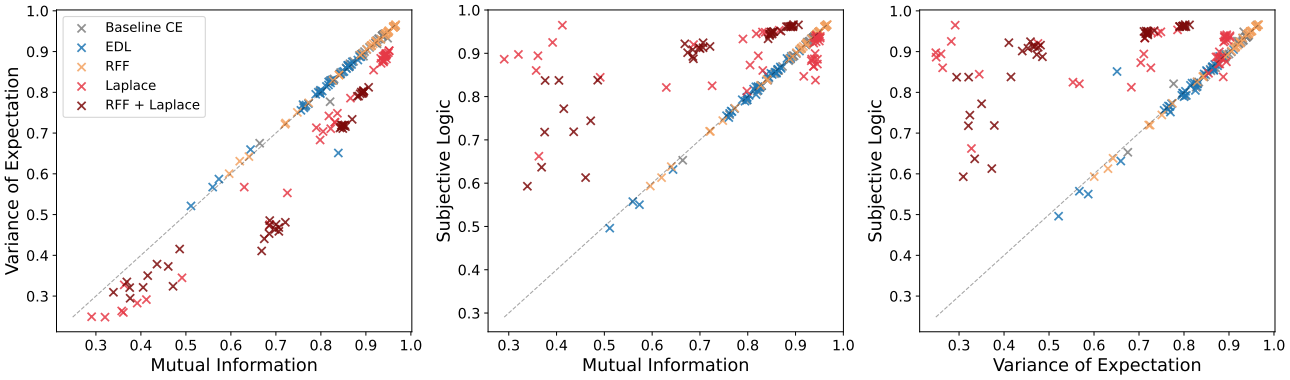

Figure D.21: Correlation analysis of the three epistemic metrics' AUROCs for the SVHN OOD set, pooling over all four dataset sizes and ten seeds. Each dot is one run's triple of AUROCs (Subjective Logic, Mutual Information, Variance of Expectation), scattered pairwise against a $x = y$ reference line and colored by method.