Density-Informed Pseudo-Counts for Calibrated Evidential Deep Learning

Pietro Carlotti Nevena Gligić Arya Farahi



The University of Texas at Austin
Department of Statistics and Data Sciences
College of Natural Sciences

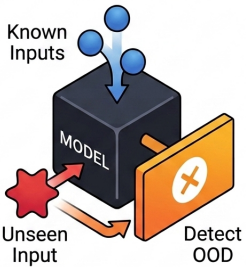
CosmicAI
Funded by the U.S. National Science Foundation and the Simons Foundation

Uncertainty Quantification

Predictions are important, but calibrated **uncertainty** is essential for:



(a) Safety-critical decision-making



(b) Detection of unseen inputs

Yet modern classifiers (e.g. neural networks) achieve good predictive accuracy while remaining **overconfident**, especially under **limited data** availability or **covariate shift**.

Evidential Deep Learning

One way the literature has addressed this is via **Evidential Deep Learning** (EDL). Its key idea is to model class probabilities via a **Dirichlet distribution** parameterized by a neural network,

$$p \sim \mathsf{Dir}(\alpha + \mathsf{NN}_X^{\phi}).$$

Training is performed by minimizing a loss balancing data fit and regularization,

$$\mathcal{L}_{\mathsf{EDL}}^{\lambda}(\phi) = \mathcal{L}_{\mathsf{data}}(\phi) + \lambda\, \mathcal{L}_{\mathsf{reg}}(\phi).$$

This framework is appealing because a single forward pass returns both:

- a **prediction** over the class probabilities.
- an **uncertainty** quantification over the model's confidence.

A Bayesian Interpretation of EDL

To understand EDL, we cast it as a Bayesian model with a **tempered** pseudo-likelihood:

$$Y_i \mid p_i \overset{\text{ind.}}{\propto} \prod_{k=1}^{K} p_i(k)^{\nu \mathbb{1}(Y_i=k)}, \quad i = 1, \dots, n,$$
$$p_i \overset{\text{i.i.d.}}{\sim} \text{Dir}(\alpha), \quad i = 1, \dots, n.$$

Conjugacy yields a closed-form **posterior** given the observed labels:

$$\pi^{\nu}_{Y_i} = \text{Dir}(\alpha + \nu e_{Y_i}), \qquad i = 1, \dots, n,$$

where $e_{Y_i} \in \{0, 1\}^K$ is the one-hot encoding of Y_i.

A Variational Approach

Since the model ignores inputs, we induce **covariate dependence** by approximating the posterior via mean-field amortized **variational inference** (VI) parameterized with a neural network:

$$q^{\phi}_{X_{1:n}} = \prod_{i=1}^{n} q^{\phi}_{X_i}, \quad \text{where} \quad q^{\phi}_{X_i} = \mathsf{Dir}(\alpha + \mathsf{NN}^{\phi}_{X_i}), \quad i = 1, \dots, n.$$

Then, the network parameters are learned by minimizing the KL divergence to the posterior:

$$\hat{\phi}_n = \underset{\phi \in \Phi}{\mathsf{argmin}} \ \mathsf{KL}(q^{\phi}_{X_{1:n}} \,\|\, \pi^{\nu}_{Y_{1:n}}).$$

Equivalence to EDL

Remarkably, minimizing this amortized VI objective is exactly equivalent to EDL training:

$$\hat{\phi}_n = \underset{\phi \in \Phi}{\operatorname{argmin}} \; \mathcal{L}_{\mathsf{EDL}}^{\frac{1}{\nu}}(\phi; X_{1:n}, Y_{1:n}).$$

The EDL loss depends on the choice of ν, with key implications for uncertainty quantification.

EDL's Hidden Hyperparameter Dependence

In the EDL literature, **epistemic** uncertainty of a Dirichlet distribution is measured by **vacuity**:

$$\text{Vacuity}(\alpha) = \frac{K}{\alpha_0}, \qquad \text{where} \qquad \alpha_0 = \sum_{k=1}^{K} \alpha(k).$$

Notably, even if we had access to the **oracle** distribution, uncertainty would remain constant:

$$\text{Vacuity}\left(\alpha + \text{NN}_X^{\phi^\star}\right) = \frac{K}{\alpha_0 + \nu}.$$

EDL's uncertainty quantification is therefore **spurious** because it is hyperparameter-driven.

Decoupling Prediction from Uncertainty Quantification

We propose a new specification that replaces ν with a local, density-informed pseudo-count.

DIP-EDL Parametrization

Consider the following variational family:

$$q_{X_i}^{\psi,\phi} = \mathsf{Dir}(\alpha + n\,\mathsf{DE}_{X_i}^{\psi}\,\mathsf{NN}_{X_i}^{\phi}), \qquad i = 1, \dots, n,$$

where

- n governs the degree of concentration,
- DE^{ψ} learns the true covariate density P_X^*,
- NN^{ϕ} learns the true conditional class distribution $P_{Y|X}^*$.

Salient Properties of DIP-EDL

Our method enjoys some desirable properties:

- **Distributional awareness**: predictions contract to the prior where data is scarce:

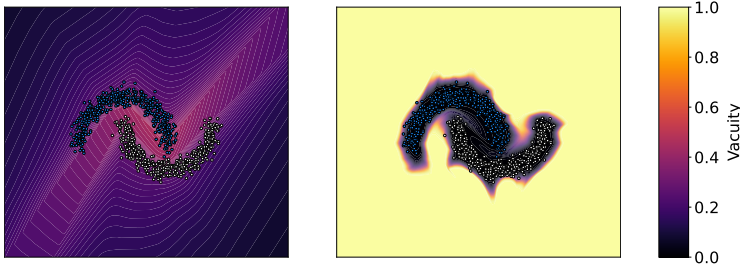
$$\mathsf{DE}_{X_i}^{\psi} \approx 0 \quad \Longrightarrow \quad q_{X_i}^{\psi,\phi} \approx \mathsf{Dir}(\alpha).$$

- **Asymptotic concentration**: if both estimators are consistent, then predictions converge:

$$\mathsf{DE}_{X_i}^{\psi} \xrightarrow{p} P_X^*(X_i), \;\; \mathsf{NN}_{X_i}^{\phi} \xrightarrow{p} P_{Y|X}^*(\cdot \mid X_i) \;\; \Longrightarrow \;\; q_{X_i}^{\psi,\phi} \xrightarrow{p} P_{Y|X}^*(\cdot \mid X_i).$$

Results on Simulated Data

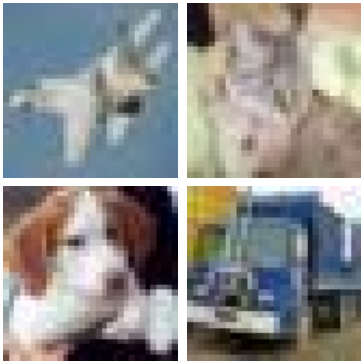
We visualize the vacuity of EDL and DIP-EDL on a synthetic **two-moons** dataset. As expected, EDL's vacuity stays artificially uniform, while DIP-EDL's satisfies distributional awareness.



(a) EDL (b) DIP-EDL

Benchmark Data

We train models on **CIFAR-10** (ID), evaluating on **SVHN** (OOD).



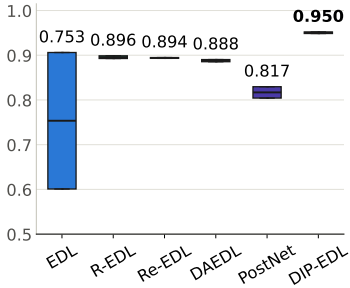
(a) ID: CIFAR-10



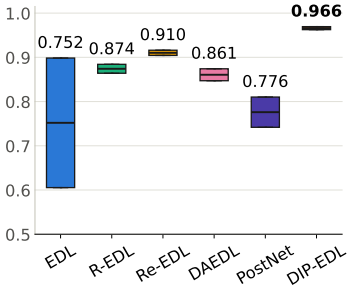
(b) OOD: SVHN

Results on Benchmark Data

We assess prediction with **accuracy** and uncertainty quantification with **AUROC**.



(a) ID Accuracy

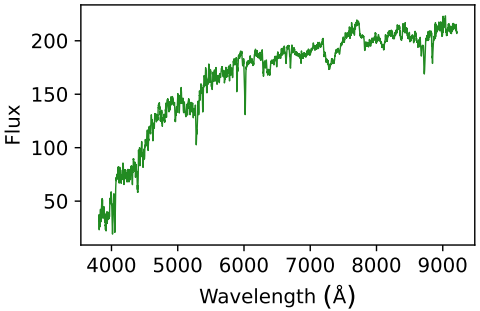


(b) OOD AUROC

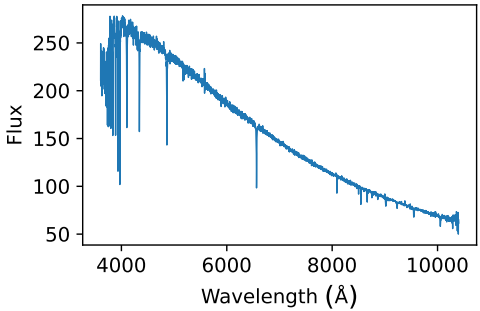
DIP-EDL **outperforms** existing baselines in both ID classification and OOD detection.

Scientific Data

We train models on **LAMOST Galaxy/Quasar** (ID), evaluating on **Star** (OOD).



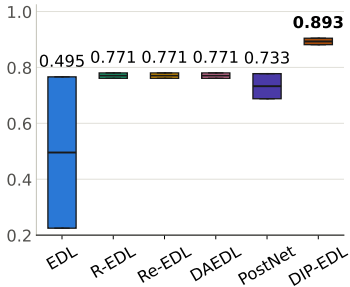
(a) ID: Galaxy



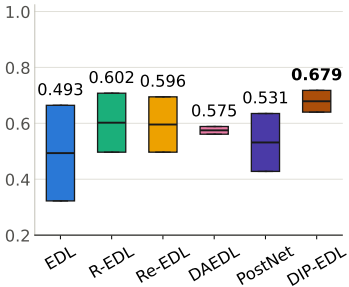
(b) OOD: Star

Results on Scientific Data

We assess prediction with **accuracy** and uncertainty quantification with **AUROC**.



(a) ID Accuracy



(b) OOD AUROC

DIP-EDL **outperforms** existing baselines in both ID classification and OOD detection.

Summary

To review what we discussed today:

1. **EDL** is equivalent to **amortized VI** on a Bayesian model with a tempered pseudo-likelihood.
2. EDL's uncertainty quantification is **spurious** because it is hyperparameter-driven.
3. We propose **DIP-EDL**, a new method that decouples prediction from uncertainty.
4. DIP-EDL exhibits **distributional awareness** and **asymptotic concentration**.
5. DIP-EDL **outperforms** existing baselines on simulated, benchmark, and scientific data.

Thank you for your attention!



Scan for the paper

Contact: pietro.carlotti@utexas.edu

Evaluation Metrics

Our experiments rely on two metrics: **ID classification** and **OOD detection**.

- **Accuracy** is the fraction of correctly classified points:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(\hat{Y}_i = Y_i\right).$$

- **AUROC** is the fraction of ID/OOD pairs correctly ordered by vacuity:

$$\text{AUROC} = \frac{1}{|\text{ID}|\,|\text{OOD}|}\sum_{i\in\text{ID}}\sum_{j\in\text{OOD}} \mathbb{1}\left(\text{Vacuity}\left(\alpha_{X_j}\right) > \text{Vacuity}\left(\alpha_{X_i}\right)\right).$$

Hyperparameter-Driven Vacuity

EDL minimizes $\mathrm{KL}\left(q^{\phi}_{X_{1:n}} \,\|\, \pi^{\nu}_{Y_{1:n}}\right)$, which is zero only if **concentration parameters** match:

$$\mathrm{Dir}\left(\alpha + \mathrm{NN}^{\widehat{\phi}_n}_{X_i}\right) = \mathrm{Dir}(\alpha + \nu\, e_{Y_i}) \quad \Longleftrightarrow \quad \mathrm{NN}^{\widehat{\phi}_n}_{X_i} = \nu\, e_{Y_i}.$$

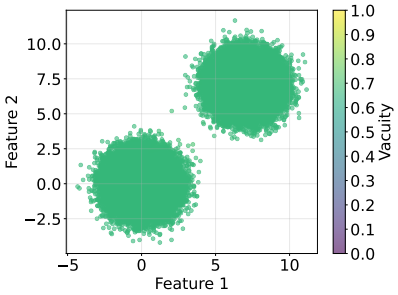
This fixes every training point's total concentration, which directly determines its vacuity:

$$u\left(\alpha + \mathrm{NN}^{\widehat{\phi}_n}_{X_i}\right) = \frac{K}{\alpha_0 + \sum_{k=1}^{K} \mathrm{NN}^{\widehat{\phi}_n}_{X_i}(k)} = \frac{K}{\alpha_0 + \nu}, \qquad i = 1, \ldots, n.$$

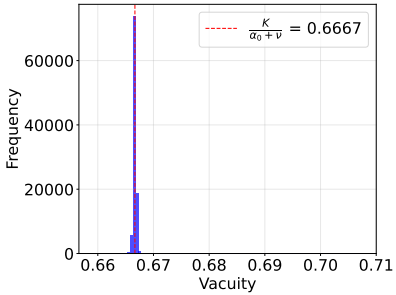
Vacuity depends only on the hyperparameter ν, **not** on the structure of the data.

Vacuity Is Constant

We train EDL on a mixture of two well-separated Gaussians: $\frac{1}{2}\mathcal{N}\left([0,0]^\top, I_2\right)+\frac{1}{2}\mathcal{N}\left([7,7]^\top, I_2\right)$:



(a) Vacuity across feature space ($\nu = 1$)



(b) Vacuity histogram ($\nu = 1$)

Vacuity is uniform across clusters and concentrates sharply around the theoretical value $\frac{K}{\alpha_0+\nu}$.

Vacuity Matches the Theoretical Prediction

We repeated this experiment with different values of ν:

ν	$\frac{K}{\alpha_0+\nu}$	Vacuity Average	Vacuity Standard Deviation
1	0.667	0.667	0.0004
5	0.286	0.288	0.0021
10	0.167	0.167	0.0007
50	0.038	0.038	0.0002
100	0.020	0.020	0.0002
500	0.004	0.004	0.0005
1000	0.002	0.002	0.0001

The computed vacuity closely matches the theoretical anchor $\frac{K}{\alpha_0+\nu}$ across all values of ν.

Practical Advantages of DIP-EDL

Beyond its theoretical guarantees, our proposed construction offers two practical benefits:

- **Modularity**: DIP-EDL allows to train DE^{ψ} and NN^{ϕ} independently:

$$\hat{\psi}_n = \underset{\psi \in \Psi}{\operatorname{argmin}}\ \mathcal{L}_{\mathsf{DE}}(\psi; X_{1:n}), \qquad \hat{\phi}_n = \underset{\phi \in \Phi}{\operatorname{argmin}}\ \mathcal{L}_{\mathsf{NN}}(\phi; X_{1:n}, Y_{1:n}).$$

- **Architecture-agnosticism**: DIP-EDL accommodates any valid density estimator/classifier.

Overview of Full Experimental Results

Beyond what is shown in the previous slides, the paper additionally:

- Reports experiments on **MNIST**, another benchmark dataset.
- Compares against multi-pass baselines: **MC-Dropout** and **Deep Ensembles**.
- Reports **AUPR**, omitted here for space.
- Reports the **Brier Score** (BS), the squared error of predicted probabilities to true labels:

$$\mathrm{BS} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\hat{p}_i(k) - \mathbb{1}(Y_i = k) \right)^2 .$$

MNIST: Full Results

We train DIP-EDL and baselines on MNIST (ID) vs. K-MNIST/Omniglot (near/far-OOD):

Model	ID Accuracy	Near-OOD AUROC	Far-OOD AUROC
EDL	0.9166	0.9116	0.9204
R-EDL	*0.9940*	0.9729	0.9879
Re-EDL	*0.9940*	0.9910	**1.0000**
DAEDL	0.9937	*0.9987*	0.9997
PostNet	0.9925	0.9602	0.9936
MC Dropout	*0.9940*	0.9800	0.9880
Deep Ensemble	*0.9940*	0.9830	0.9970
DIP-EDL	**0.9953**	**0.9997**	*0.9998*

Bold indicates the best result, *italics* the second best.

CIFAR-10: Full Results

We train DIP-EDL and baselines on CIFAR-10 (ID) vs. CIFAR-100/SVHN (near/far-OOD):

Model	ID Accuracy	Near-OOD AUROC	Far-OOD AUROC
EDL	0.7535	0.7588	0.7519
R-EDL	0.8957	0.8484	0.8741
Re-EDL	0.8939	0.8603	0.9102
DAEDL	0.8879	0.8323	0.8606
PostNet	0.8169	0.7691	0.7760
MC Dropout	*0.9534*	0.8833	0.8802
Deep Ensemble	**0.9614**	**0.9072**	*0.9628*
DIP-EDL	0.9503	*0.9002*	**0.9660**

Bold indicates the best result, *italics* the second best.

LAMOST: Full Results

We train DIP-EDL and baselines on LAMOST Galaxy/Quasar data (ID) vs. Star data (OOD):

Model	ID Accuracy	OOD AUROC
EDL	0.4955	0.4934
R-EDL	*0.7706*	0.6022
Re-EDL	*0.7706*	0.5957
DAEDL	*0.7706*	0.5748
PostNet	0.7328	0.5315
MC Dropout	0.5851	0.4974
Deep Ensemble	**0.8929**	**0.7425**
DIP-EDL	**0.8929**	*0.6789*

Bold indicates the best result, *italics* the second best.

Ablation Study: MNIST

We ablate DIP-EDL's three components to isolate their contributions: n, DE^{ψ}, and NN^{ϕ}.

n	DE^{ψ}	NN^{ϕ}	ID Accuracy	ID BS	Near-OOD AUROC	Far-OOD AUROC
✓	✓	✓	0.9955	0.0079	0.9998	0.9998
×	✓	✓	0.9952	0.7202	0.9996	0.9996
✓	×	✓	0.9958	0.0069	0.5138	0.5299
✓	✓	×	0.0980	0.9000	0.9998	0.9998

This reveals three distinct roles:

- Removing n compromises **ID BS**.
- Removing DE^{ψ} compromises **OOD AUROC**.
- Removing NN^{ϕ} compromises **ID Accuracy** and **ID BS**.

Ablation Study: CIFAR-10

We ablate DIP-EDL's three components to isolate their contributions: n, DE^{ψ}, and NN^{ϕ}.

n	DE^{ψ}	NN^{ϕ}	ID Accuracy	ID BS	Near-OOD AUROC	Far-OOD AUROC
✓	✓	✓	0.9496	0.0978	0.9000	0.9686
×	✓	✓	0.9496	0.7653	0.9000	0.9686
✓	×	✓	0.9496	0.0831	0.5036	0.5014
✓	✓	×	0.1000	0.9000	0.9000	0.9686

This reveals three distinct roles:

- Removing n compromises **ID BS**.
- Removing DE^{ψ} compromises **OOD AUROC**.
- Removing NN^{ϕ} compromises **ID Accuracy** and **ID BS**.

Robustness to Density Misestimation

We test DIP-EDL's robustness to two forms of error in its density estimator DE^{ψ}:

- **Likelihood scaling**: we scale the concentration parameters by a factor $\gamma > 0$:

$$\alpha = 1 + \gamma\, n\, \mathrm{DE}_X^{\psi}\, \mathrm{NN}_X^{\phi},$$

- **Density corruption**: we corrupt the density's log-probability with additive Gaussian noise:

$$\log \mathrm{DE}_X^{\psi} + \mathcal{N}(0, \sigma^2),$$

In both cases, DIP-EDL remains **robust** to imprecision in the density estimate.

Robustness to Likelihood Scaling

We compare MNIST (ID) vs. K-MNIST (OOD) and CIFAR-10 (ID) vs. SVHN (OOD):

γ	MNIST				CIFAR-10			
	ID Accuracy	ID BS	OOD AUROC	OOD BS	ID Accuracy	ID BS	OOD AUROC	OOD BS
0.1	0.9951	0.0092	0.9997	0.0005	0.9480	0.1148	0.9630	0.0604
0.5	0.9951	0.0090	0.9997	0.0012	0.9480	0.1029	0.9630	0.1009
1.0	0.9952	0.0088	0.9998	0.0017	0.9480	0.0996	0.9630	0.1217
2.0	0.9951	0.0089	0.9997	0.0021	0.9480	0.0969	0.9630	0.1439
5.0	0.9954	0.0087	0.9999	0.0030	0.9480	0.0940	0.9630	0.1747
10.0	0.9953	0.0088	0.9998	0.0034	0.9480	0.0920	0.9630	0.1988

Uniformly scaling the concentration parameter preserves rankings but not magnitudes:

- ID Accuracy, OOD AUROC, and OOD AUPR are **unaffected**.
- ID BS **improves** while OOD BS **worsens**.

Robustness to Density Corruption

We compare MNIST (ID) vs. K-MNIST (OOD) and CIFAR-10 (ID) vs. SVHN (OOD):

σ	MNIST				CIFAR-10			
	ID Accuracy	ID BS	OOD AUROC	OOD BS	ID Accuracy	ID BS	OOD AUROC	OOD BS
0.0	0.9951	0.0089	0.9998	0.0017	0.9480	0.0996	0.9630	0.1217
0.5	0.9952	0.0089	0.9998	0.0017	0.9480	0.0996	0.9623	0.1220
1.0	0.9952	0.0089	0.9998	0.0017	0.9480	0.1000	0.9601	0.1231
2.0	0.9952	0.0093	0.9997	0.0019	0.9480	0.1023	0.9514	0.1274
5.0	0.9952	0.0433	0.9984	0.0035	0.9480	0.1451	0.8995	0.1505
∞	0.9955	0.0078	0.5016	0.6957	0.9480	0.0851	0.5016	0.6514

Random noise preserves predictions and only mildly affects uncertainty rankings:

- ID Accuracy is **unaffected**.
- OOD AUROC and OOD BS both **degrade**, collapsing to near-random detection.