SHBC2025 – D1-148

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | Centre of AI in Medicine

CUPID: A Plug-in Framework for Joint Aleatoric and Epistemic Uncertainty Estimation with a Single Model

Xinran Xu[2], Xiuyi Fan[1,2,3]

1 College of Computing and Data Science, Nanyang Technological University, Singapore

2 Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

3 Centre for Medical Technologies & Innovations, National Health Group, Singapore

INTRODUCTION

Deep models can be overconfident, so reliable prediction requires understanding not only how uncertain a model is, but also why it is uncertain [1]. Uncertainty in deep learning mainly comes from two sources: aleatoric uncertainty, caused by noise or ambiguity in the data, and epistemic uncertainty, caused by limited model knowledge or insufficient training coverage [2]. Distinguishing them is important for trustworthy decision-making [3]. Existing methods often estimate only one uncertainty type, or they require modifying and retraining the base model. This increases computational cost and limits compatibility with pretrained systems.

We propose **CUPID**, a lightweight plug-in module that can be inserted into a pretrained network to jointly estimate aleatoric and epistemic uncertainty without changing or retraining the base model. It also enables layer-wise analysis of how uncertainty evolves inside the network.

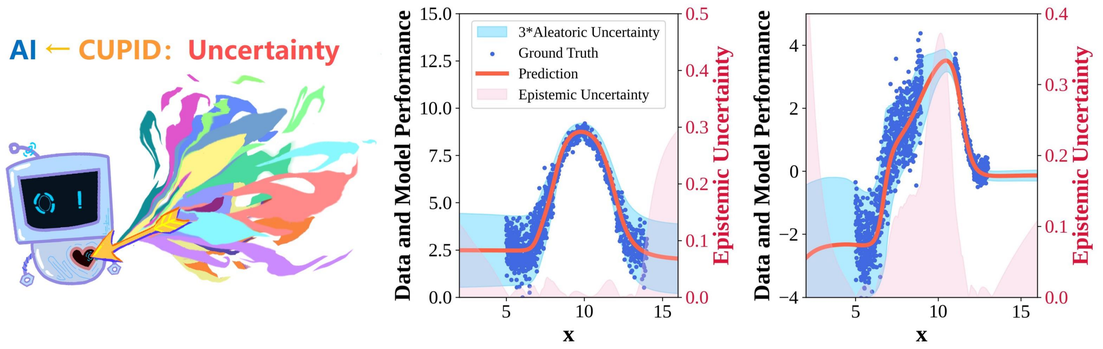
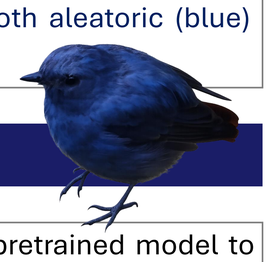


Figure 1: CUPID uncertainty estimation on a 1D regression toy problem. CUPID is inserted into an MLP-based predictive model. CUPID captures both aleatoric (blue) and epistemic (red) uncertainty.



METHODS

CUPID is a plug-in module inserted at any intermediate layer of a pretrained model to jointly estimate aleatoric uncertainty from a variance branch and epistemic uncertainty from prediction changes caused by feature perturbation.

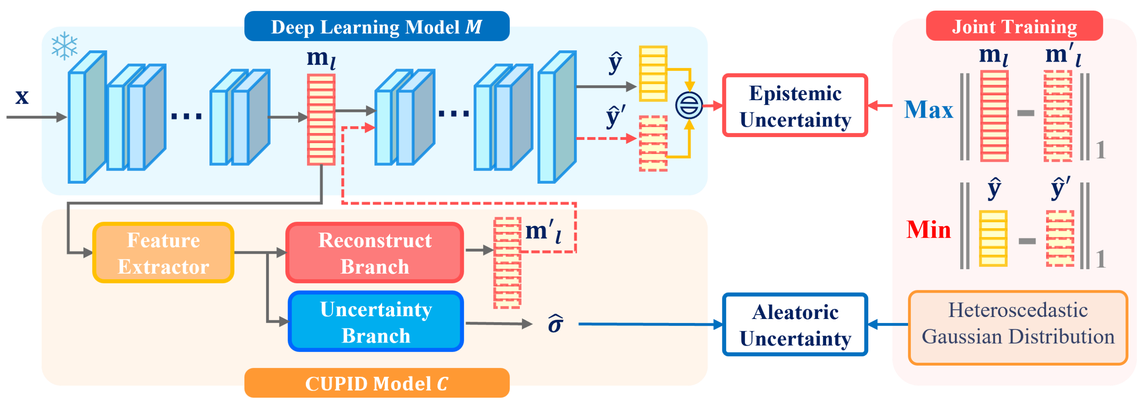


Figure 2: The CUPID pipeline. Aleatoric uncertainty is estimated using a dedicated Uncertainty Branch, while epistemic uncertainty is captured by measuring the variance between the original model output $\hat{y}$ and the perturbed output $\hat{y}'$.

- **Plug-in formulation.** Let $M(x) = F_l(B_l(x))$, where B_l extracts the intermediate feature and F_l maps it to the final prediction. CUPID model takes $m_l = B_l(x)$ and produces $(m_l', \hat{\sigma})$.
- **Aleatoric uncertainty branch.** The uncertainty branch models input-dependent noise with a heteroscedastic distribution. The predicted variance is used as the aleatoric uncertainty score. For training, CUPID predicts the log-variance for numerical stability.
- **Epistemic uncertainty branch.** The reconstruction branch learns a feature perturbation that changes the internal representation while preserving the final prediction. Epistemic uncertainty is measured by the discrepancy between the original output and the perturbed output.
- **Joint objective.** Both branches are optimized jointly in a single model. This lets CUPID estimate both uncertainty types without modifying or retraining the base predictor.

$$(m'_{l,n}, \hat{\sigma}_n) = C(m_{l,n}; \omega)$$

$$U_{\text{alea}}(x_n) = \hat{\sigma}_n^2$$

$$U_{\text{epis}}(x) = \|\hat{y}_n - \hat{y}'_n\|_1$$

$$L_{\text{CUPID}} = L_{\text{epis}} + \lambda_2 L_{\text{alea}}$$

REFERENCES

[1] K. Zou, Z. Chen, X. Yuan, X. Shen, M. Wang, and H. Fu, "A review of uncertainty estimation and its application in medical imaging," Meta-Radiology, p. 100003, 2023.

[2] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al., "A survey of uncertainty in deep neural networks," Artificial Intelligence Review, vol. 56, suppl. 1, pp. 1513–1589, 2023.

[3] M. Chan, M. Molina, and C. Metzler, "Estimating epistemic and aleatoric uncertainty with a single model," in Advances in Neural Information Processing Systems, vol. 37, pp. 109845–109870, 2024.

RESULTS

We evaluate CUPID on three complementary settings: medical image misclassification detection, out-of-distribution (OOD) detection, and image super-resolution as a regression task. These experiments are designed to test whether CUPID can separate aleatoric and epistemic uncertainty across both classification and regression problems. We further include an ablation study on insertion location to analyze how uncertainty emerges across network depth. All experiments are repeated three times, and results are reported as mean ± standard deviation. The best model for each metric is in bold, and the second best is underlined.

Table 1: Performance of misclassification detection (misclassified samples as positive). CUPID Aleatoric achieved the best performance on GLV2, while CUPID Epistemic performed best on HAM10000, suggesting different dominant sources of uncertainty across datasets.

Method	GLV2 AUC (↑)	GLV2 AURC (↓)	GLV2 Spearman (↑)	HAM10000 AUC (↑)	HAM10000 AURC (↓)	HAM10000 Spearman (↑)
CUPID Alea.	**0.870 ± 0.002**	**0.018 ± 0.001**	<u>0.941 ± 0.004</u>	0.769 ± 0.023	0.067 ± 0.007	0.722 ± 0.014
CUPID Epis.	0.769 ± 0.015	0.034 ± 0.002	0.701 ± 0.051	**0.855 ± 0.006**	**0.047 ± 0.001**	<u>0.907 ± 0.001</u>
MC Dropout	0.768 ± 0.006	0.027 ± 0.001	0.888 ± 0.005	0.829 ± 0.001	0.076 ± 0.001	0.861 ± 0.002
Rate-in	0.815 ± 0.006	<u>0.024 ± 0.001</u>	0.816 ± 0.004	<u>0.846 ± 0.001</u>	<u>0.048 ± 0.000</u>	**0.915 ± 0.000**
IGRUE	0.642 ± 0.007	0.058 ± 0.002	0.199 ± 0.004	0.548 ± 0.004	0.157 ± 0.002	0.027 ± 0.018
PostNet Alea.	0.671 ± 0.006	0.182 ± 0.004	0.641 ± 0.011	0.793 ± 0.007	0.142 ± 0.003	0.764 ± 0.006
PostNet Epis.	0.559 ± 0.031	0.238 ± 0.019	0.284 ± 0.054	0.751 ± 0.017	0.158 ± 0.010	0.698 ± 0.033
BNN	<u>0.829 ± 0.018</u>	0.025 ± 0.003	**0.954 ± 0.007**	0.793 ± 0.006	0.096 ± 0.004	0.821 ± 0.009
DEC	0.503 ± 0.012	0.192 ± 0.006	0.803 ± 0.139	0.837 ± 0.017	0.082 ± 0.004	0.874 ± 0.007

Table 2: Performance of OOD detection (OOD samples as positive). PAPILA and ACRIMA share the same research problem (glaucoma detection) with the ID dataset while CIFAR10 is a general classification dataset. CUPID Epistemic performs best on PAPILA and ACRIMA, suggesting that it is more sensitive to subtle distribution shifts within related clinical domains. In contrast, CUPID Aleatoric performs best on CIFAR-10, showing stronger response to extreme domain mismatch.

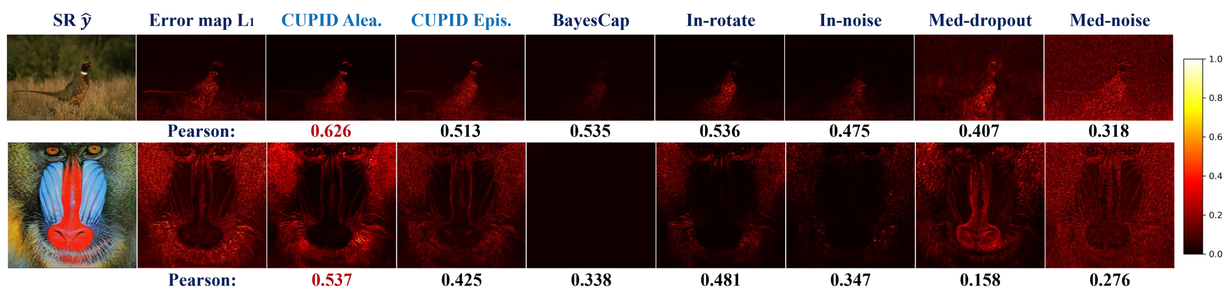
Method	PAPILA AUC(↑)	PAPILA AUPR(↑)	ACRIMA AUC(↑)	ACRIMA AUPR(↑)	CIFAR10 AUC(↑)	CIFAR10 AUPR(↑)
CUPID Alea.	0.379 ± 0.027	0.333 ± 0.007	0.717 ± 0.029	0.661 ± 0.027	**0.983 ± 0.005**	**0.998 ± 0.001**
CUPID Epis.	**0.877 ± 0.032**	**0.854 ± 0.027**	**0.978 ± 0.010**	**0.984 ± 0.007**	0.898 ± 0.054	<u>0.991 ± 0.005</u>
MC Dropout	<u>0.733 ± 0.002</u>	0.586 ± 0.007	0.869 ± 0.003	0.816 ± 0.009	0.887 ± 0.004	0.986 ± 0.001
Rate-in	0.328 ± 0.005	0.329 ± 0.008	0.363 ± 0.003	0.390 ± 0.003	0.620 ± 0.001	0.927 ± 0.002
IGRUE	0.636 ± 0.114	0.486 ± 0.097	<u>0.941 ± 0.008</u>	<u>0.944 ± 0.008</u>	<u>0.978 ± 0.005</u>	**0.998 ± 0.001**
PostNet Alea.	0.638 ± 0.060	0.487 ± 0.067	0.549 ± 0.040	0.487 ± 0.040	0.657 ± 0.032	0.952 ± 0.005
PostNet Epis.	0.577 ± 0.097	0.425 ± 0.088	0.685 ± 0.154	0.654 ± 0.151	0.773 ± 0.082	0.976 ± 0.011
BNN	0.707 ± 0.040	<u>0.612 ± 0.050</u>	0.708 ± 0.073	0.699 ± 0.042	0.643 ± 0.108	0.959 ± 0.013
DEC	0.515 ± 0.024	0.457 ± 0.024	0.680 ± 0.003	0.685 ± 0.012	0.660 ± 0.015	0.963 ± 0.003

Table 3: Performance on natural image datasets (Set5, Set14, BSDS100) and medical imaging dataset IXI (MRI scans). CUPID Aleatoric gives the best overall quantitative performance on the natural-image benchmarks, suggesting that reconstruction uncertainty is mainly driven by data-related ambiguity in this setting. On the IXI dataset, however, CUPID Epistemic becomes more informative, which is consistent with the larger domain shift from the training distribution.

Method	Set5 Pearson (↑)	Set5 AUSE (↓)	Set5 UCE (↓)	Set14 Pearson (↑)	Set14 AUSE (↓)	Set14 UCE (↓)
CUPID Alea.	**0.528 ± 0.006**	**0.010 ± 0.000**	**0.045 ± 0.018**	**0.527 ± 0.002**	**0.012 ± 0.000**	**0.049 ± 0.005**
CUPID Epis.	0.416 ± 0.004	0.018 ± 0.001	0.266 ± 0.007	0.449 ± 0.005	0.019 ± 0.000	0.226 ± 0.003
BayesCap	0.485 ± 0.038	**0.010 ± 0.000**	0.098 ± 0.001	0.422 ± 0.064	**0.012 ± 0.000**	0.100 ± 0.000
in-rotate	<u>0.493 ± 0.000</u>	**0.010 ± 0.000**	0.071 ± 0.000	<u>0.490 ± 0.000</u>	<u>0.013 ± 0.000</u>	<u>0.072 ± 0.000</u>
in-noise	0.370 ± 0.006	0.019 ± 0.000	<u>0.051 ± 0.035</u>	0.354 ± 0.001	0.022 ± 0.000	0.826 ± 0.006
med-dropout	0.219 ± 0.023	0.030 ± 0.001	0.680 ± 0.043	0.271 ± 0.012	0.024 ± 0.000	0.292 ± 0.022
med-noise	0.312 ± 0.003	0.022 ± 0.000	0.826 ± 0.006	0.293 ± 0.002	0.022 ± 0.000	0.826 ± 0.006

Method	BSDS100 Pearson (↑)	BSDS100 AUSE (↓)	BSDS100 UCE (↓)	IXI Pearson (↑)	IXI AUSE (↓)	IXI UCE (↓)
CUPID Alea.	**0.536 ± 0.001**	<u>0.012 ± 0.000</u>	**0.042 ± 0.012**	<u>0.677 ± 0.008</u>	**0.004 ± 0.000**	**0.021 ± 0.004**
CUPID Epis.	0.464 ± 0.007	0.018 ± 0.000	0.185 ± 0.007	**0.734 ± 0.018**	**0.004 ± 0.000**	0.298 ± 0.013
BayesCap	0.427 ± 0.034	**0.011 ± 0.000**	0.100 ± 0.000	0.447 ± 0.034	**0.004 ± 0.000**	0.100 ± 0.000
in-rotate	<u>0.465 ± 0.000</u>	<u>0.012 ± 0.000</u>	<u>0.077 ± 0.000</u>	0.598 ± 0.000	**0.004 ± 0.000**	0.093 ± 0.000
in-noise	0.353 ± 0.001	0.022 ± 0.000	0.826 ± 0.006	0.461 ± 0.001	0.005 ± 0.000	<u>0.091 ± 0.002</u>
med-dropout	0.397 ± 0.002	0.020 ± 0.000	0.136 ± 0.008	0.570 ± 0.001	0.007 ± 0.000	0.337 ± 0.026
med-noise	0.293 ± 0.000	0.024 ± 0.000	0.700 ± 0.002	0.439 ± 0.000	0.006 ± 0.000	0.859 ± 0.002

Figure 3: Comparison of visual results between error and uncertainty maps. CUPID Aleatoric shows the best texture alignment and highest correlation with error maps.



The ablation results provide insight into how uncertainty develops inside the network. As shown in Figure 4, aleatoric uncertainty is estimated more effectively when CUPID is placed closer to the output, whereas epistemic uncertainty benefits more from earlier insertion points. This trend suggests that the two uncertainty types emerge differently across network depth, and that CUPID can be used not only for uncertainty estimation but also for analyzing uncertainty propagation.

Figure 4: Performance of CUPID inserted at varying locations: misclassification detection (Left) and super-resolution (Right).