OUT-OF-DISTRIBUTION DETECTION FROM SMALL TRAINING SETS USING BAYESIAN NEURAL NETWORK CLASSIFIERS

Kevin Raina, Tanya Schmah
Department of Mathematics and Statistics - University of Ottawa, Canada
krain033@uottawa.ca, tschmah@uottawa.ca

BMVC 2025

Introduction

Out-of-Distribution (OOD) detection is critical to AI reliability and safety. The goal is to determine whether a test sample belongs to the In-Distribution (ID) or is OOD. This problem is especially important in high-stakes domains such as medical imaging, finance, and cybersecurity, where models can make confident yet incorrect predictions on unfamiliar data.

Why Bayesian?
Bayesian Neural Networks (BNNs) explicitly model *uncertainty* by placing a probability distribution over weights:

$$p(\omega|\mathcal{D}) = \frac{p(\mathcal{D}|\omega)\, p(\omega)}{p(\mathcal{D})} \quad \text{where} \quad p(\mathcal{D}) = \int p(\mathcal{D}|\omega)p(\omega)d\omega$$

This allows the model to quantify how confident it should be about its predictions.

Posterior Predictive Distribution:

$$p(y|x,\mathcal{D}) = \int p(y|x,\omega)\, p(\omega|\mathcal{D})\, d\omega$$

Bayesian Neural Networks incorporate both:

- **Aleatoric uncertainty:** inherent noise in the data (e.g. ambiguous inputs);
- **Epistemic uncertainty:** uncertainty about the model parameters ω.

Why it matters for small data:

When data are scarce, point-estimate networks (*MLE*) can be overconfident.
Bayesian models share belief across plausible parameter configurations, reducing overconfidence.
Likelihood: $\hat{y} = f(x;\hat{\omega}) \Rightarrow$ Bayesian: $p(y|x,\mathcal{D}) = \mathbb{E}_{p(\omega|\mathcal{D})}[p(y|x,\omega)]$

In essence: BNNs let us reason about epistemic uncertainty. This is crucial for OOD detection, where we must know not only *what* the model predicts, but also *how sure* it is.

Methods

Our study compares a family of deterministic and Bayesian post-hoc OOD scores applied to neural network classifiers. These scores are computed using the logits or softmax probabilities and then thresholded to determine whether a sample is ID or OOD.

Table 1: Comparison of deterministic and Bayesian OOD scores.

Deterministic Scores	**Bayesian Scores**
Softmax Entropy (SE) $H_{\text{SE}}(x) = -\sum_{c=1}^{K} p(y_c\|x) \log p(y_c\|x)$ Entropy of predictive probabilities; high values indicate uncertainty.	**Predictive Entropy (PE)** $H_{\text{PE}}(x) = -\sum_{c=1}^{K} p(y_c\|x,\mathcal{D}) \log p(y_c\|x,\mathcal{D})$ Computed from the posterior predictive distribution $p(y\|x,\mathcal{D}) = \mathbb{E}_{q(\omega)}[p(y\|x,\omega)]$.
Max Logit (ML) $S_{\text{ML}}(x) = \max_c z_c(x)$ Confidence derived directly from logits. Low max logit $\rightarrow$ OOD.	**Expected Logit (EL ML)** $S_{\text{EL-ML}}(x) = \max_c \mathbb{E}_{q(\omega)}[z_c(x;\omega)]$ Expectation of logits over posterior samples.
kNN in Logit Space $S_{\text{kNN}}(x) = \|z(x) - z(x_k)\|_2$ Distance in logit space from nearest ID neighbours.	**Expected Logit kNN (EL kNN)** $S_{\text{EL-kNN}}(x) = \|\hat{z}(x) - \hat{z}(x_k)\|_2$, $\hat{z}(x) = \mathbb{E}_{q(\omega)}[z(x;\omega)]$ Same, but with expected logits.
kNN+ (class-conditioned) $S_{\text{kNN+}}(x) = d_{\text{kNN}}(x) + \min_c d_c(x) - \frac{1}{C-1}\sum_{c'\neq c} d_{c'}(x)$ Balances aleatoric and epistemic uncertainty.	**Expected Logit kNN+ (EL kNN+)** Same as left but using $\hat{z}(x)$ (posterior mean logits).
No direct deterministic analog. Though, kNN is a surrogate measure of epistemic uncertainty.	**Mutual Information (MI)** $I(\omega;y\|x,\mathcal{D}) = H[p(y\|x,\mathcal{D})] - \mathbb{E}_{q(\omega)}[H[p(y\|x,\omega)]]$ Captures pure epistemic uncertainty. No deterministic counterpart.

Results

Our experiments show that Bayesian methods consistently outperform deterministic ones in the small-data regime. For MNIST, the class-conditional Bayesian kNN+ score (EL kNN+) achieves the best overall performance, with the highest AUC-ROC and lowest FPR95 across OOD datasets. For CIFAR-10, the Bayesian Maximum Logit (EL ML) yields the strongest results.

Notably, logit-space methods (ML, kNN, kNN+) tend to outperform probability-based ones (SE, PE, MI). This highlights that information encoded in raw logits carries a stronger signal for distinguishing OOD samples.

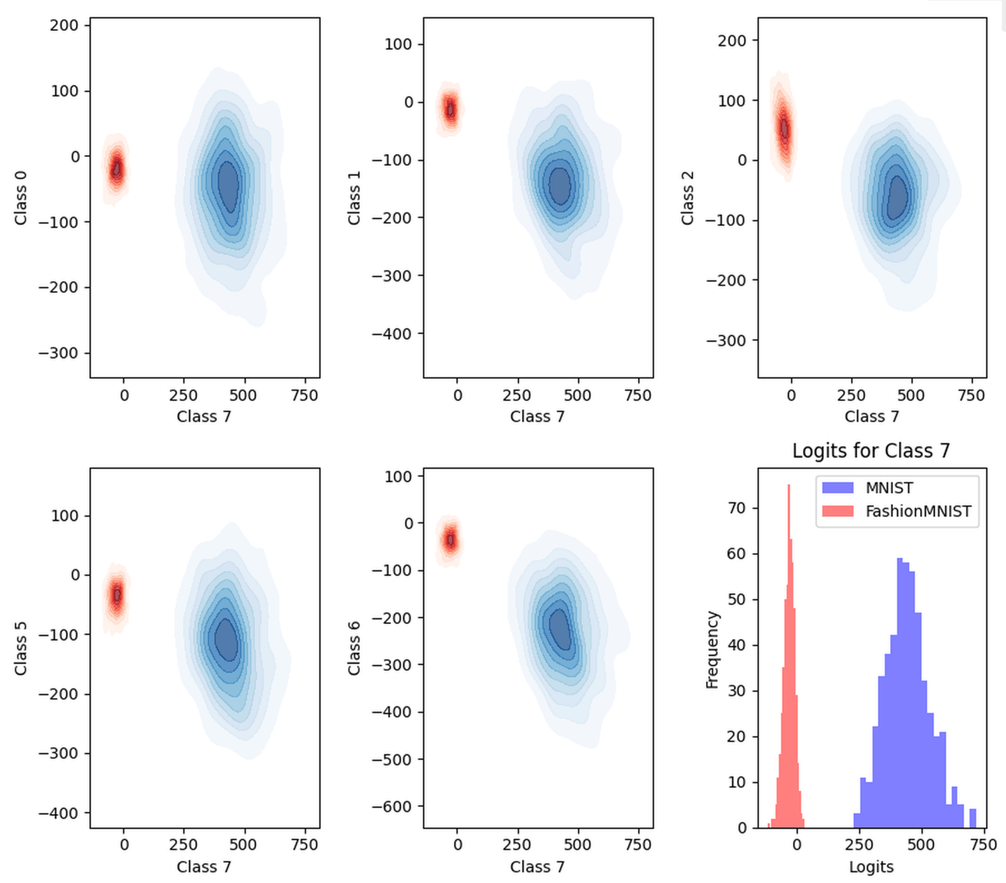

Figure 1: Joint distribution of logit components from the posterior distribution for blue (MNIST sample) and red (FashionMNIST) sample. The MNIST sample shows strong logit signals for class 7 and very weak signals for the remaining clsses, while the FashionMNIST sample doesn't follow this pattern.

Table 2: Summary of OOD detection performance. Best methods bolded.

Method	MNIST (AUC-ROC)	CIFAR-10 (AUC-ROC)
Softmax Entropy	73.3	90.5
MLE/FinT ML	50.4	90.7
MLE kNN	85.4	81.8
MLE/FinT kNN+	87.0	90.3
Predictive Entropy	91.8	91.2
Mutual Information	90.4	68.7
EL ML	93.5	**92.6**
EL kNN	86.5	81.3
EL kNN+	**93.8**	90.9

Discussion

Our findings highlight three key insights:

- **Bayesian Advantage**: BNNs outperform deterministic networks in OOD detection, especially when training data is limited. Explicit uncertainty modelling is the main driver of this improvement.
- **Logit-Space Power**: Scores computed in logit space (particularly EL kNN+) provide more robust signals for OOD detection than probability-based metrics.
- **Role of Priors**: Incorporating informative priors via empirical Bayes (MOPED) further enhances performance, demonstrating the importance of prior knowledge in low-data regimes.

These insights suggest that future work should explore richer Bayesian inference techniques, such as the Laplace approximation, and adapt novel OOD scores that exploit internal model representations.

Acknowledgements

This work was supported by the University of Ottawa. We thank colleagues and reviewers for valuable discussions. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).