Improve Predictive Model Calibration for Safety-Critical Domains

Le Van Tuan Anh[*]
anh.lvt2416657@sis.hust.edu.vn

Nguyen Tran Minh Tri[*]
tri.ntm2400116@sis.hust.edu.vn

Le Xuan Nhat Khoi[*]
khoi.lxn2416711@sis.hust.edu.vn

Nguyen Duy Tung[*]
tung.nd2416762@sis.hust.edu.vn

1 Introduction

Traditional deterministic neural networks typically provide point estimates without meaningful uncertainty information. As a result, they are unable to distinguish between confident and uncertain predictions, especially under data ambiguity, limited training samples, or domain shift. This limitation motivates the use of Bayesian and approximate Bayesian methods to capture predictive uncertainty and improve model reliability.

In this project, we focus on uncertainty-aware deep learning methods, specifically Monte Carlo Dropout (MC Dropout) and Bayesian Variational Inference (Bayesian VI), as scalable approximations to Bayesian neural networks. These methods allow uncertainty estimation without substantially changing the underlying model architecture or incurring prohibitive computational cost.

The goals of this project are:

- Evaluate the ability of MC Dropout and Bayesian Variational Inference to capture predictive uncertainty and improve model calibration on standard image classification benchmarks, namely CIFAR-10 and CIFAR-100.
- Compare uncertainty-aware models against deterministic baselines using calibration-focused metrics (e.g., Expected Calibration Error) and reliability visualizations, in addition to standard accuracy-based performance.
- Apply the studied uncertainty estimation methods to the task of brain tumor classification from medical images, assessing whether improved calibration and uncertainty awareness translate to more reliable predictions in a safety-critical setting.

Overall, this project aims to demonstrate that incorporating uncertainty estimation into deep learning models can enhance not only predictive reliability but also practical applicability, particularly in domains where decision confidence is as important as raw accuracy.

2 Dataset and Data Source

2.1 CIFAR-10 and CIFAR-100

The CIFAR-10 and CIFAR-100 datasets are widely used benchmarks for image classification and model calibration analysis.

- CIFAR-10 consists of 60,000 color images of size 32×32, evenly distributed across 10 classes, with 50,000 training images and 10,000 test images.

[*] Group 9

- CIFAR-100 has the same image resolution and total number of samples but is divided into 100 fine-grained classes, making it a more challenging classification task with increased class ambiguity.

For both datasets, standard train-test splits provided by the dataset are used.

2.2 Brain Tumor Image Classification Dataset

To evaluate uncertainty estimation in a real-world, safety-critical setting, our project uses the Brain Tumor Classification dataset obtained from [Kaggle].

The dataset consists of MRI brain images categorized into multiple tumor-related classes (glioma, meningioma, pituitary tumor, and no tumor). Compared to CIFAR datasets, these images exhibit:

- Higher semantic complexity
- Greater intra-class variability
- Increased consequences of misclassification

This dataset is particularly suitable for uncertainty-aware modeling, as medical imaging tasks require not only accurate predictions but also reliable confidence estimates to support clinical decision-making.

2.3 Preprocessing and Data Handling

Across all datasets, the following preprocessing steps are applied:

- Resizing and Normalization
- Label encoding
- Image augmentation (e.g., Random flips, Rotation, Random Crops)

3 Approach

Our project investigates uncertainty-aware deep learning methods for image classification by comparing deterministic neural networks with two approximate Bayesian approaches: Monte Carlo Dropout (MC Dropout) and Bayesian Variational Inference (Bayesian VI). All methods are evaluated under a consistent experimental setup to ensure fair comparison.

3.1 Baseline Deterministic Model

As a reference point, a standard deterministic convolutional neural network (CNN) is used as the baseline model. The network is trained using maximum likelihood estimation with a cross-entropy loss function and produces point estimates for class probabilities via the softmax output.

While this model can achieve high classification accuracy, its predicted confidence scores are often poorly calibrated, motivating the need for uncertainty-aware alternatives.

3.2 Monte Carlo Dropout

Monte Carlo Dropout is an efficient approximation to Bayesian inference that interprets dropout as a form of variational inference. Unlike standard dropout, which is disabled during inference, MC Dropout keeps dropout active at test time, enabling stochastic forward passes through the network.

Given an input image x, multiple forward passes are performed:

$$\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\}$$

where each prediction results from a different dropout mask. The final predictive distribution is obtained by averaging these stochastic predictions:

$$p(y \mid x) \approx \frac{1}{T} \sum_{t=1}^{T} p(y \mid x, \theta_t)$$

3.3 Bayesian Variational Inference

Bayesian Variational Inference explicitly models uncertainty in network parameters by learning a probabilistic distribution over weights, rather than fixed values. In this framework, the posterior is approximated by a variational distribution $q(\theta)$, typically chosen to be a tractable family such as a Gaussian

Training is performed by minimizing the variational objective, known as the Evidence Lower Bound (ELBO):

$$\mathcal{L}_{ELBO} = \mathbb{E}[\log p(y \mid x, \theta)] - \mathbb{KL}[q(\theta) \parallel p(\theta)]$$

The first term encourages accurate predictions, while the KL divergence term regularizes the model by keeping the learned posterior close to the prior distribution. At inference time, multiple samples from the learned weight distribution are used to generate predictive distributions.

Compared to MC Dropout, Bayesian VI provides a more principled Bayesian formulation but typically incurs higher computational cost.

3.4 Key Design Choice

To ensure fair and meaningful comparison:

- The same base CNN architecture is used across all methods
- Identical training data splits and preprocessing steps are applied
- The number of stochastic forward passes is fixed across uncertainty-aware methods
- Deterministic and Bayesian models are evaluated using the same metrics

4 Experiments and Evaluation

4.1 Training, Validation and Test Splits

For CIFAR-10 and CIFAR-100, the standard dataset splits provided by the dataset are used, consisting of predefined training and test sets. A portion (20%) of the training data is further held out as a validation set for hyperparameter tuning and early stopping.

Given that the Brain Tumor Dataset is pre-partitioned into train and test sets, we will derive a validation set from the training data, similar to the methodology used for the CIFAR datasets.

All models are trained using the same data splits to ensure comparability.

4.2 Training Procedure

All models are trained using mini-batch stochastic gradient descent with identical optimization settings across methods. Deterministic and Bayesian models share the same base architecture and training pipeline, differing only in how uncertainty is modeled.

For uncertainty-aware models:

- MC Dropout uses dropout during both training and inference
- Bayesian VI models are trained by optimizing the ELBO objective

4.3 Evaluation Metrics

Model performance is evaluated using both accuracy-based metrics and calibration-focused metrics, reflecting the dual goal of predictive correctness and reliability.

Classification Performance:

- Accuracy (ACC)

Calibration Metrics:

- Expected Calibration Error (ECE)
- Maximum Calibration Error (MCE)

These metrics quantify how well predicted probabilities align with empirical correctness.

4.4 Calibration Visualization

To complement numerical metrics, reliability diagrams are used to visualize model calibration. These plots compare predicted confidence against observed accuracy across probability bins, enabling intuitive interpretation of overconfidence and underconfidence.

4.5 Baselines and Comparisons

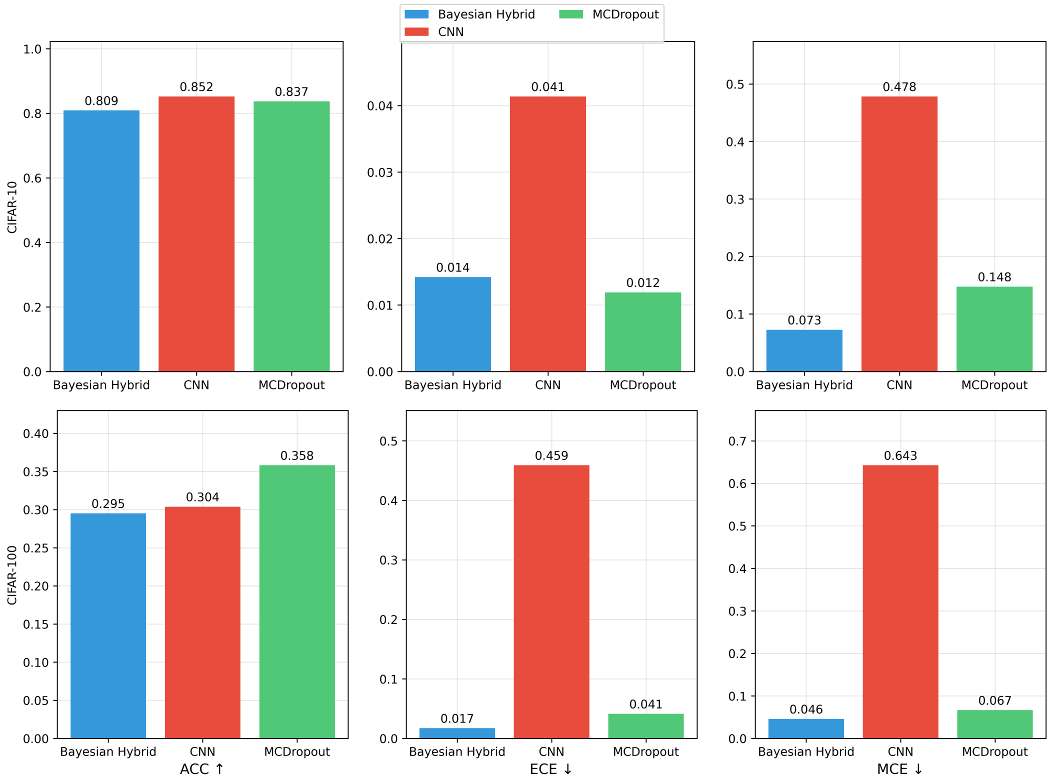
The deterministic CNN serves as the primary baseline for evaluating the effectiveness of uncertainty-aware methods.

Comparative analysis focuses on:

- Differences in calibration quality
- Trade-offs between accuracy and uncertainty estimation
- Behavior under increased task difficulty (CIFAR-10 vs CIFAR-100)
- Robustness in a real-world medical classification scenario

5 Results

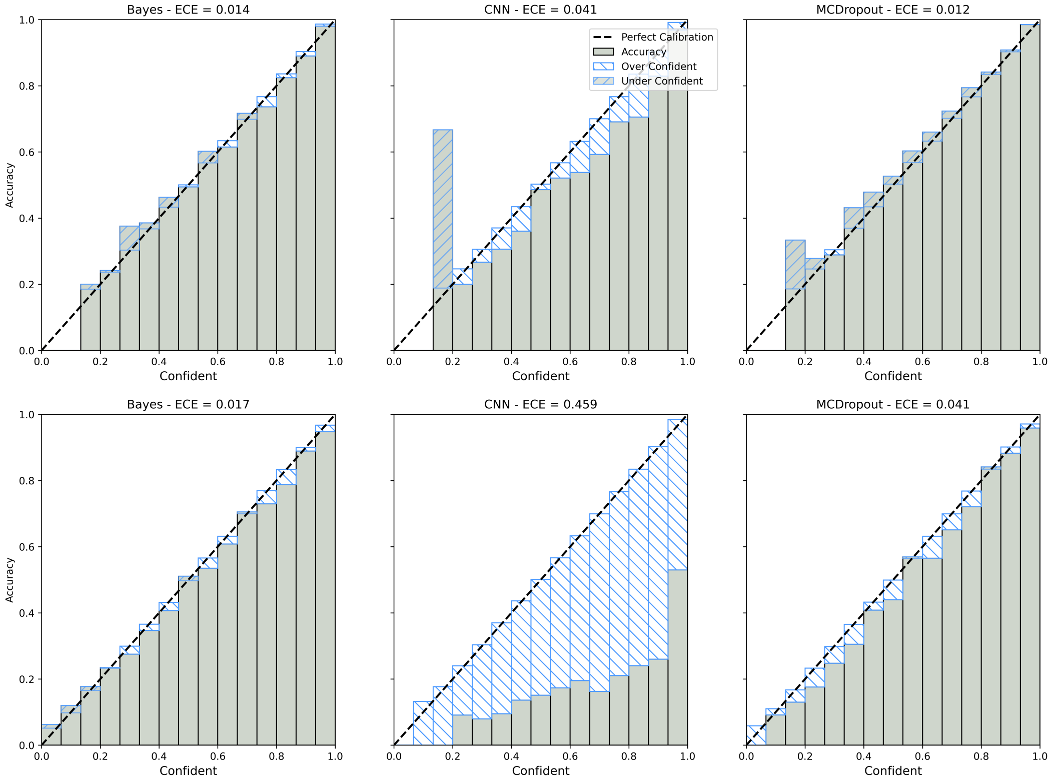
5.1 Results on the CIFAR Datasets



In terms of raw accuracy, the models performed differently depending on the dataset's complexity. On the simpler CIFAR-10 dataset, the baseline CNN achieved the highest accuracy (0.852), marginally outperforming MC Dropout (0.837) and the Bayesian Hybrid (0.809). However, on the more complex

CIFAR-100 task, MC Dropout surpassed the baseline, achieving the highest accuracy of 0.358 compared to the CNN's 0.304. This suggests that while probabilistic layers may introduce a slight regularization penalty on simple tasks, techniques like MC Dropout effectively prevent overfitting on harder problems.

Despite the competitive accuracy of the baseline CNN, the calibration metrics reveal a critical flaw: significant overconfidence.



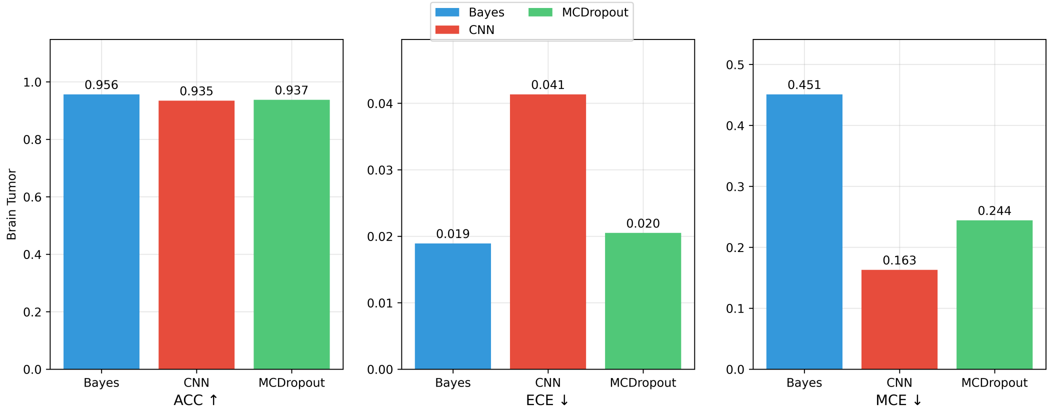
The deterministic CNN consistently produced the highest calibration errors. On CIFAR-100, it exhibited a massive ECE of 0.459 and an MCE of 0.643, indicating extreme misalignment between confidence and correctness. In strong contrast, both probabilistic methods drastically improved reliability. The Bayesian Hybrid proved the most robust, achieving the lowest ECE on CIFAR-100 (0.017) and a low MCE on CIFAR-10 (0.073), effectively reducing the expected error by over 90% compared to the baseline.

This disparity is visually evident in the reliability diagrams. The CNN's confidence bars consistently below the diagonal "Perfect Calibration" line, confirming that the model is often "overconfident". Conversely, both the Bayesian Hybrid and MC Dropout models produce bars that closely adhere to the diagonal.

These results highlight a clear distinction in model behavior: while the deterministic CNN maximizes accuracy on simple data, it fails to capture its own ignorance. The Bayesian Hybrid and MC Dropout methods successfully mitigate this overconfidence, providing accurate uncertainty estimates with only minor (or no) trade-offs in classification accuracy.

5.2 Results on the Brain Tumor Dataset

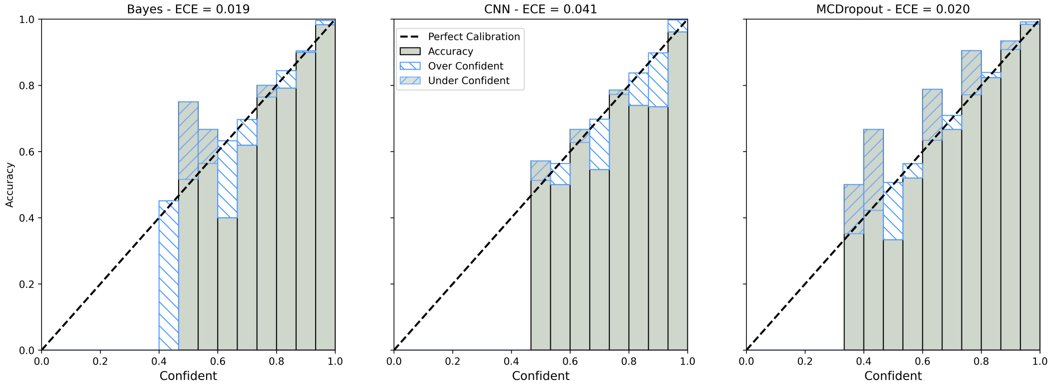
Following the benchmark on CIFAR, we applied the same methodologies to the domain-specific task of Brain Tumor classification. This experiment aimed to verify if the calibration benefits observed in general benchmarks translate to a high-stakes medical imaging context.



Contrary to the CIFAR experiments where the deterministic baseline often led in accuracy, the Bayesian Hybrid demonstrated superior predictive performance on this medical dataset.

- The Bayesian Hybrid achieved the highest accuracy of 0.956, outperforming both the baseline and the dropout ensemble. This suggests that the variational inference layers successfully captured complex features in the MRI scans that the standard models missed.
- The MC Dropout model (0.937) and the baseline CNN (0.935) performed comparably to each other but lagged slightly behind the Bayesian approach.

The calibration metrics present a compelling trade-off between average reliability and worst-case stability



- In terms of general trustworthiness, the probabilistic models clearly outperformed the baseline. The Bayesian Hybrid achieved the lowest expected error of 0.019, followed closely by MCDropout at 0.020. The deterministic CNN was significantly less calibrated, with an ECE of 0.041, indicating that its confidence scores are generally less representative of true correctness.
- A notable anomaly appeared in the maximum error metrics. Despite having the worst average calibration, the CNN achieved the lowest MCE (0.163). In contrast, the Bayesian Hybrid, despite its excellent average performance, suffered the highest MCE of 0.451. This suggests that while the Bayesian model is highly reliable overall, it contains specific "blind spots" (particular confidence bins) where its probability estimates deviate drastically from reality.

6 Conclusion

The experiments conducted in this study highlight the critical role of uncertainty quantification in deployment scenarios, particularly in medical imaging where trust is as important as accuracy.

6.1 The Accuracy-Calibration Trade-off

A common hypothesis in Bayesian Deep Learning is that introducing probabilistic layers often incurs a penalty in raw predictive accuracy in exchange for better calibration. Our results on CIFAR-10 supported this, where the deterministic CNN reigned supreme. However, the Brain Tumor experiment challenged this assumption. The Bayesian Hybrid model achieved the highest accuracy (0.956), outperforming the deterministic baseline (0.935). This suggests that for complex, domain-specific tasks with potentially limited data (compared to CIFAR), the regularization effects of Variational Inference can actually enhance feature learning rather than hinder it.

6.2 Reliability in Medical Contexts (ECE vs. MCE)

- Both MC Dropout and Bayesian VI successfully mitigated the "overconfidence" problem inherent in standard CNNs. The Bayesian Hybrid reduced the Expected Calibration Error to 0.019 (vs. 0.041 for CNN), meaning that on average, a clinician can trust the probability scores provided by the model.
- However, there was a discrepancy in the calibration metrics. While the Bayesian Hybrid had the best average calibration score (ECE of 0.019), it showed a high Maximum Calibration Error (0.451). Looking at the reliability diagrams, this was caused by a specific spike in overconfidence in the 0.4-0.5 probability range.

6.3 Limitations

- **Computational Cost:** Both MC Dropout and Bayesian VI require multiple forward passes during inference to estimate uncertainty. This increases latency, which could be a bottleneck for real-time diagnostic tools, though less critical for offline analysis.
- **Instability:**The high MCE observed in the Bayesian Hybrid indicates that while Variational Inference provides excellent global calibration, it is sensitive to hyperparameter choices (such as the prior distribution) and may behave erratically in specific confidence regions.

6.4 Future Improvements

To address the stability issues observed in the Brain Tumor results, future works could explore:

1. **Temperature Scaling:** Applying post-hoc calibration (Temperature Scaling) to the Bayesian output could smooth out the specific under-confidence spike observed in the reliability diagrams.
2. **Hybrid Ensembles:** Combining the stability of the CNN (low MCE) with the sensitivity of the Bayesian model (low ECE) via a weighted ensemble could offer a "best of both worlds" solution.
3. **Epistemic vs. Aleatoric Uncertainty:** Further decomposing the uncertainty into epistemic (model knowledge) and aleatoric (data noise) components could help explain why the Bayesian model struggled specifically in the 0.4-0.5 confidence bin.

7 Reference

1. **Monte Carlo Dropout:** Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. Proceedings of the 33rd International Conference on Machine Learning (ICML).
2. **Bayesian Variational Inference:** Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight Uncertainty in Neural Networks. Proceedings of the 32nd International Conference on Machine Learning (ICML).
3. **Calibration and ECE:** Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. Proceedings of the 34th International Conference on Machine Learning (ICML).
4. **CIFAR-10 and CIFAR-100:** Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images. Technical Report, University of Toronto.

5. **Brain Tumor MRI Dataset:** Available at: [Kaggle].

6. **Pytorch:** Paszke, A., Gross, S., Massa, F., Lerer, A., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems (NeurIPS).

7. **Blitz-Bayesian-Pytorch:** Available at: [Github]