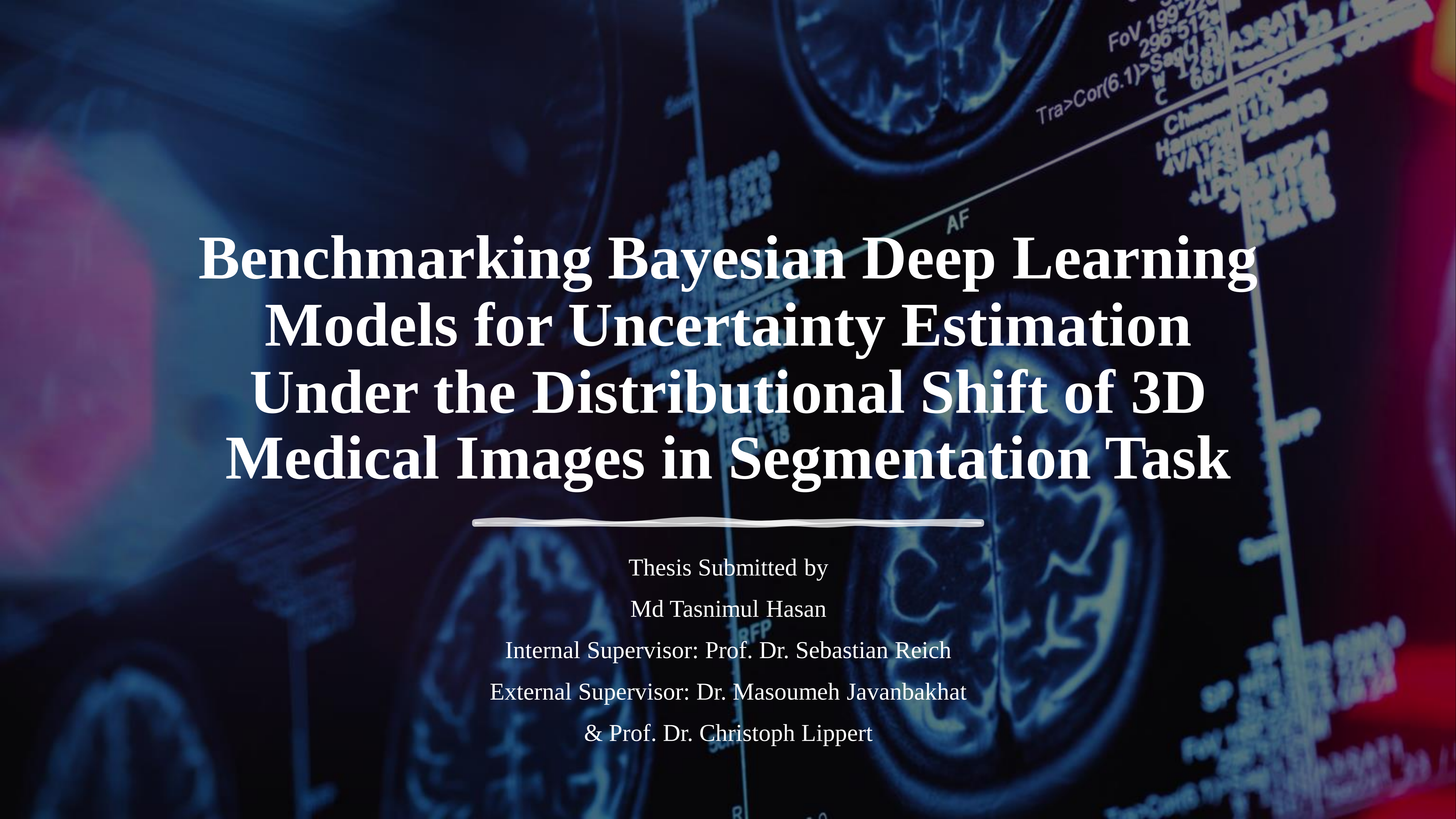


Benchmarking Bayesian Deep Learning Models for Uncertainty Estimation Under the Distributional Shift of 3D Medical Images in Segmentation Task

Thesis Submitted by

Md Tasnimul Hasan

Internal Supervisor: Prof. Dr. Sebastian Reich

External Supervisor: Dr. Masoumeh Javanbakhat

& Prof. Dr. Christoph Lippert

Contents

- Motivations
- Bayesian Neural Networks (BNNs)
- Adapting Bayesian Methods for Deep Learning
- Uncertainty Quantification and Evaluation Metrics
- Segmentation Architecture and Performance Metrics
- Research questions that we want to answer
- Experimental Setting
- Datasets
- Results
- Discussion

Motivations

Typical DL method runs under the assumption of using training and test datasets having the same distribution. Which has a great downside if the deployed model encounters with samples coming from different setting.

Another drawback is that the traditional DL methods cannot produce reliable uncertainty. Thus makes the DL models less expressive, and less reliable to the experts.

Because of these problems computer-assisted image-based disease detection and diagnosis are not common practices among medical professionals and practitioners.

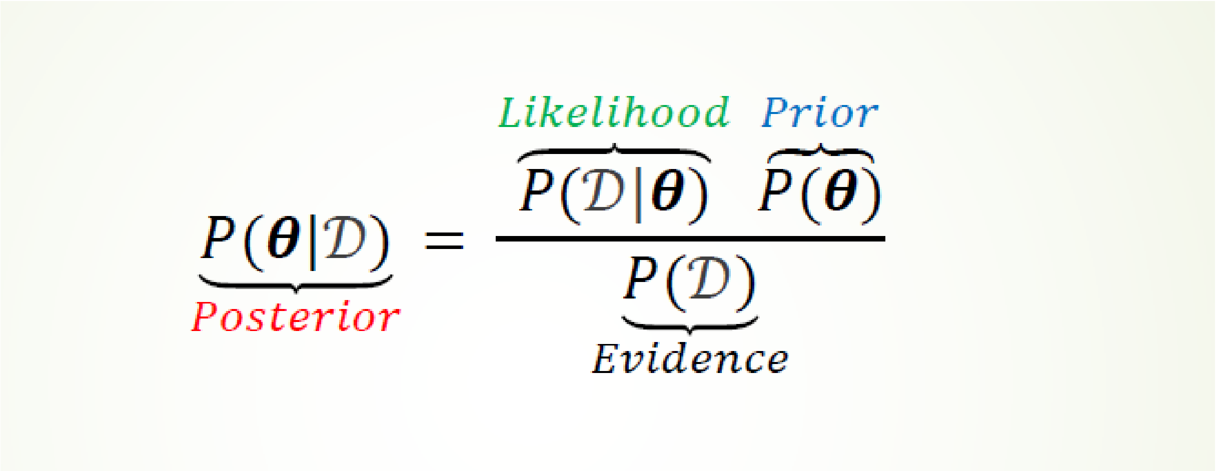
In this master’s thesis, we will provide comprehensive evaluation of uncertainty estimation based on different Bayesian and non-Bayesian DNNs under different dataset shifts for applications specific to the medical image domain.

Bayesian Neural Networks (BNNs)

- **Bayes’ theorem**: Let θ be unknown parameters of a model, and D be the dataset. We can obtain the posterior distribution by the Bayes' theorem.

- If we have the posterior distribution P(θ|D) we can obtain the predictive distribution as follows:

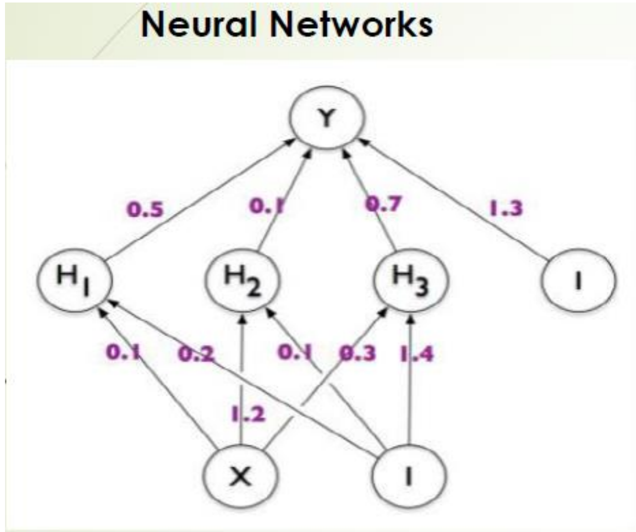
$$\mathrm{P}(\mathrm{y}|\mathrm{D}) = \int P(y|\theta)\mathrm{P}(\theta|\mathrm{D})d\theta$$

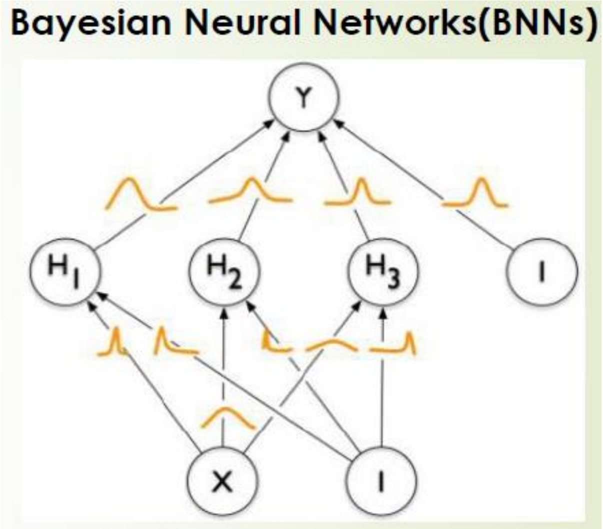


$$\underbrace{P(\boldsymbol{\theta}|\mathcal{D})}_{Posterior} = \frac{\overbrace{P(\mathcal{D}|\boldsymbol{\theta})}^{Likelihood}\;\overbrace{P(\boldsymbol{\theta})}^{Prior}}{\underbrace{P(\mathcal{D})}_{Evidence}}$$

Bayesian Neural Networks (BNNs)

- In point estimation, the model’s parameters have single deterministic value.
- The model’s parameters in the Bayesian inference are random variables, and we can learn the distribution.
- Estimating posterior distribution using Bayes theorem is intractable.
- Two widely used approaches to approximate the posterior distribution:
    - Markov chain Monte Carlo (MCMC)
    - variational inference (VI)

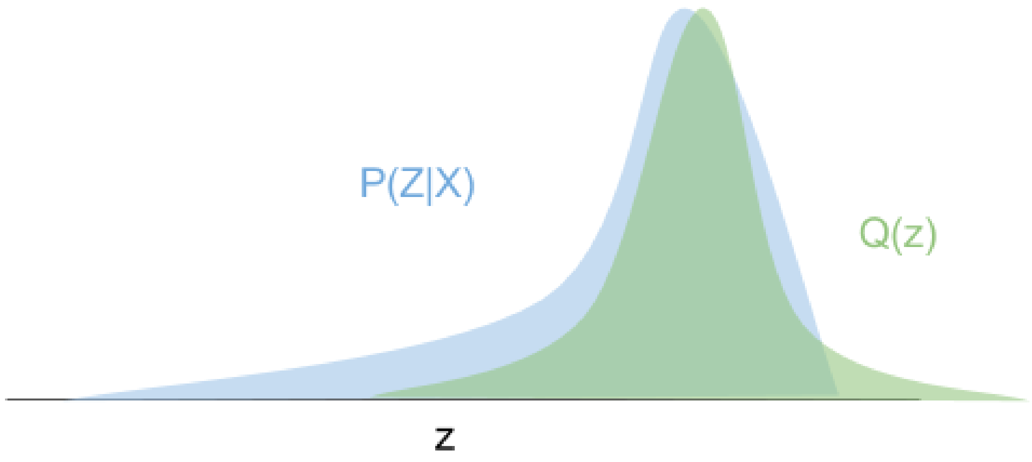




Variational Inference

- ❑ The **Variational Inference** method used to find a probability distribution Q also called **variational distribution**, has some parameters which will be learned in such a way that Q becomes as close as possible to the **true posterior** $\mathrm{P}(\theta|D)$.

- ❑ We have to find the approximation distribution such that it minimizes the KL divergence .



Markov chain Monte Carlo

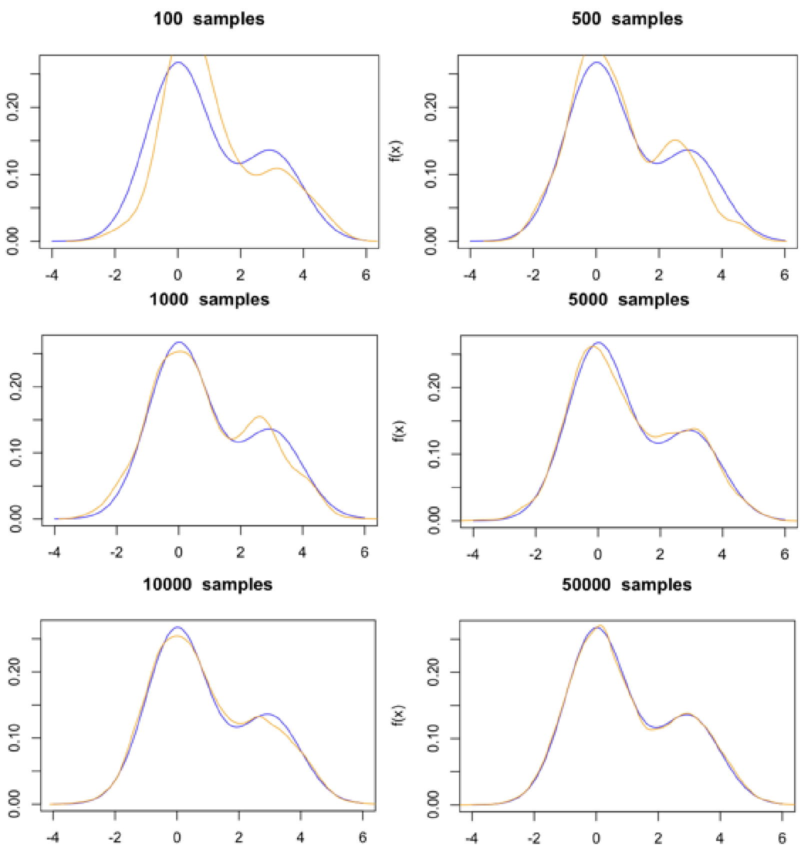
- The **Markov chain Monte Carlo** method is based on sampling from the Markov chain.

- The Markov chain can be defined as the sequence of random samples that depend only on the previous sample.

- It can produce exact samples from the posterior distribution.

- Then the Monte Carlo integration will be applied to those samples with respect to the target distribution.

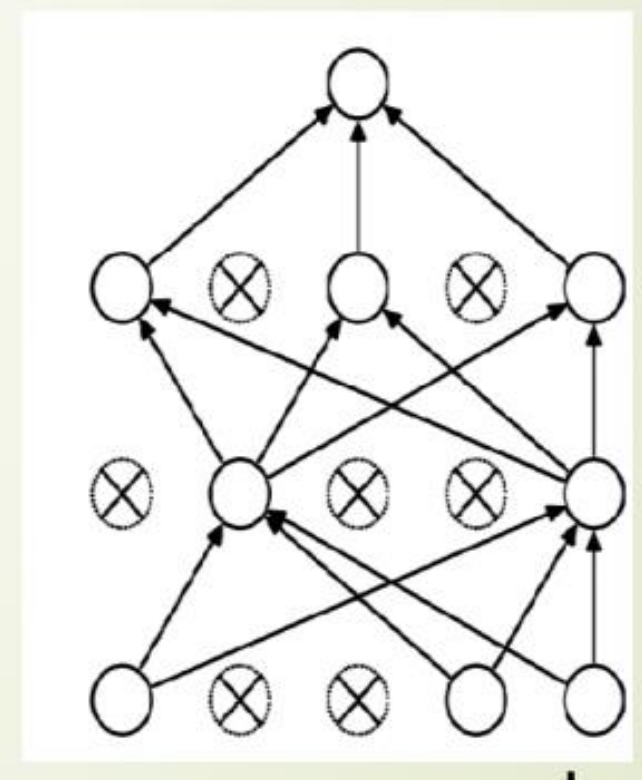
- The metropolis-Hastings algorithm and the Hamiltonian Monte Carlo are the most common MCMC method.



Adapting Bayesian Methods for Deep Learning

Bayes via Dropout:

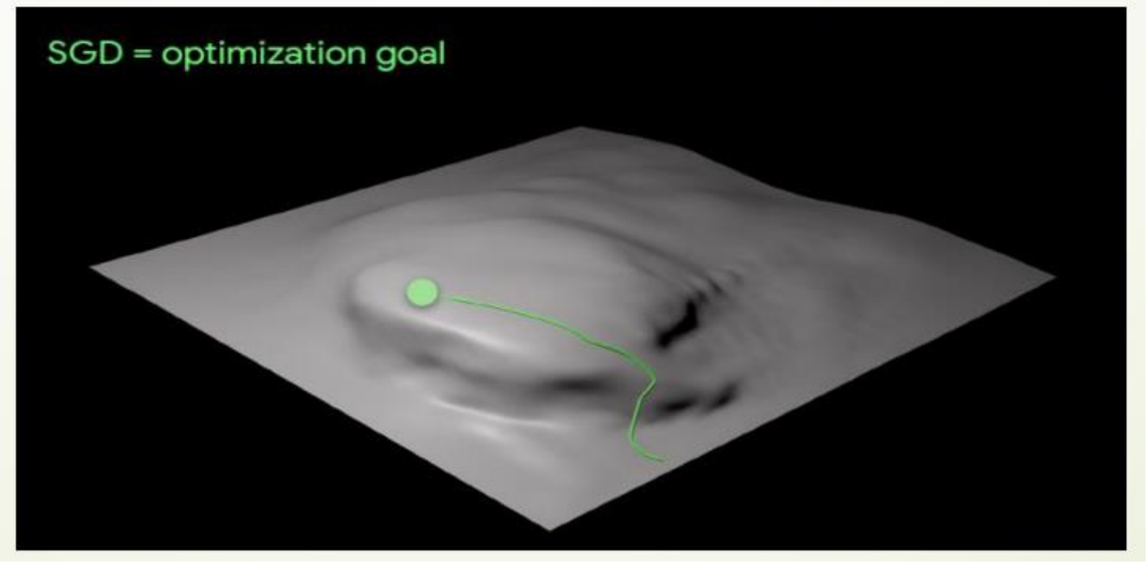
- Gal and Ghahremani showed that, training a deterministic NN with dropout and a different type of regularization such as ℓ2 regularizer (which acts as the prior) approximately corresponds to a VI method in BNNs.
- To sample from the approximate predictive distribution, the dropout is also applied at the test time, which results in a distribution for the output prediction.
- This method is popularly known as the Monte Carlo dropout.



Adapting Bayesian Methods for Deep Learning

Bayes via Stochastic Gradient Decent:

- The main idea behind stochastic gradient optimization is to find an optimal point estimation solution for the parameters θ based on the observed dataset D.



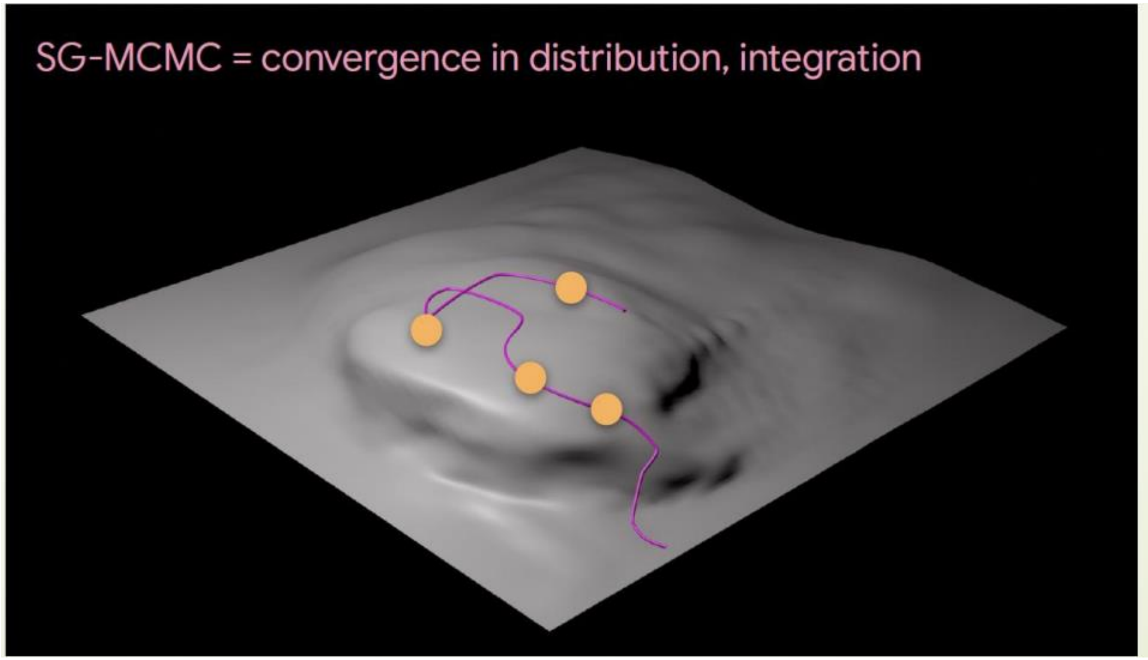
Adapting Bayesian Methods for Deep Learning

Bayes via Stochastic Gradient Decent:

- **SGLD** uses the same gradient steps as stochastic gradient optimization but with an additional Gaussian noise term.

- This injected noise is used to prevent the parameters to collapse to the MAP solution.

- **SGHMC** is based on HMC, and an improved counterpart of SGLD which introduces a momentum variable *m* and allows the parameters to explore the larger state spaces efficiently.


SG-MCMC = convergence in distribution, integration

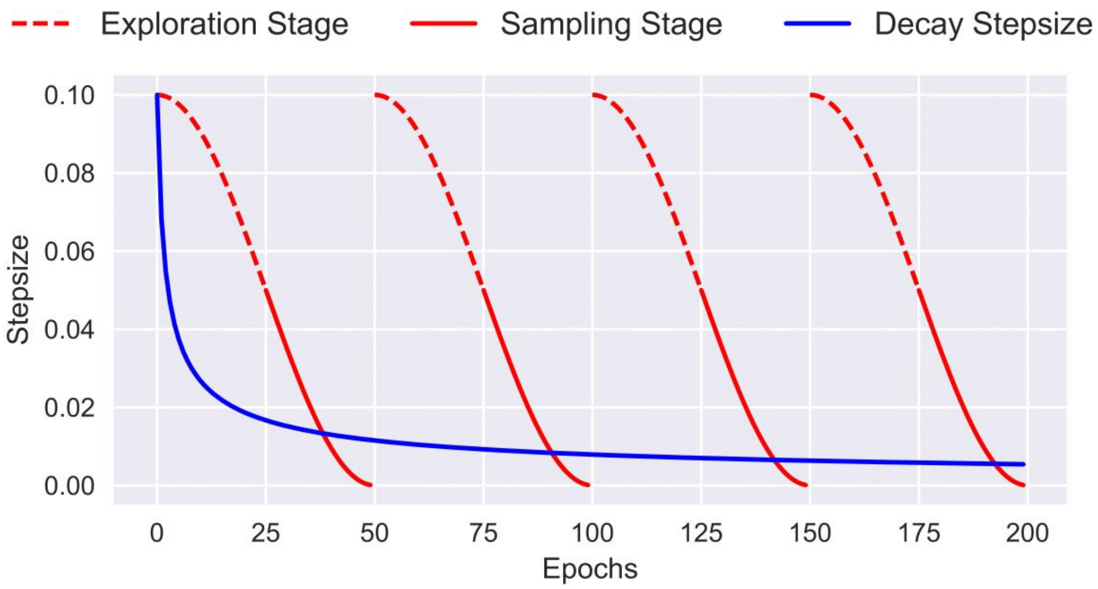
Adapting Bayesian Methods for Deep Learning

Cyclical Stochastic Gradient MCMC (cSG-MCMC):

- This method is the combination of Cyclical learning rate scheduler and the traditional SG-MCMC based method.

- It starts with a large initial learning rate which helps to escape the local mode and *explore* the parameter space efficiently.

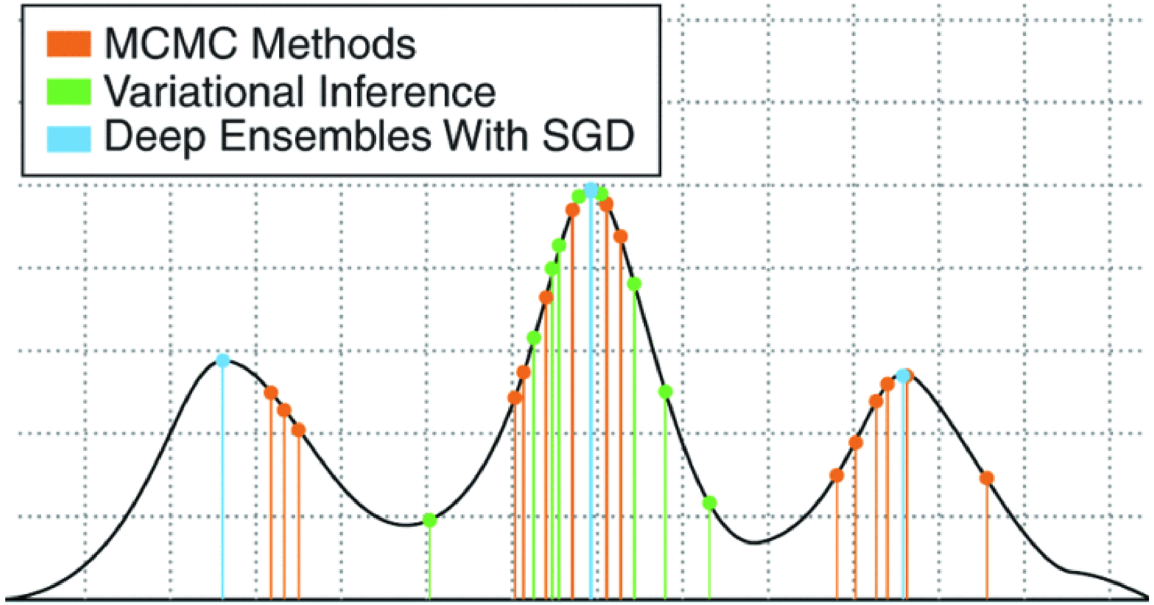
- When the learning rate is close to zero, we can collect *samples*.



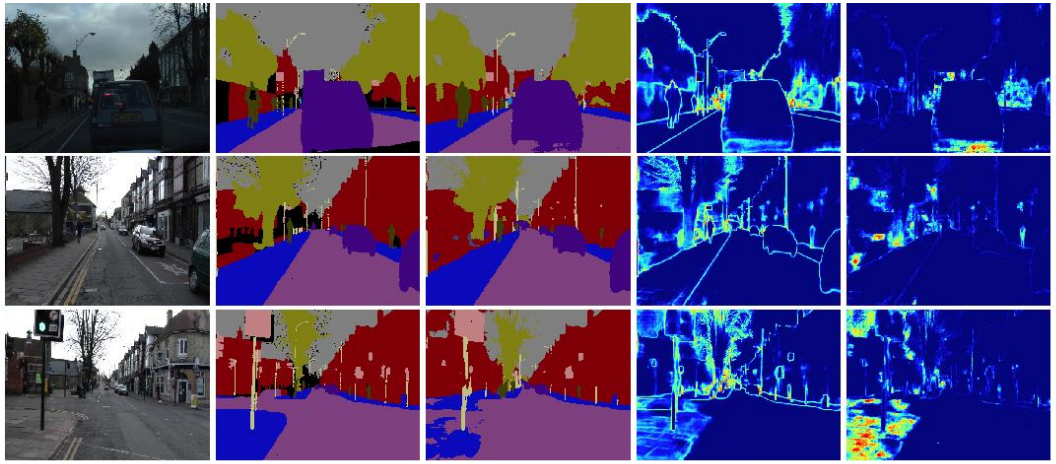
Adapting Bayesian Methods for Deep Learning

Deep Ensembles:

- Ensemble learning or ensemble of *N* models is the procedure of training the *N* number of models independently on the entire dataset using using different random initialization.



Uncertainty Quantification and evaluation metrics



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

- ❑ **Aleatoric Uncertainty:**
  - ➢ It can be described as the randomness in the dataset because of the noisy data, and low-resolution images.
  - ➢ This cannot be explained away by adding more training data.
- ❑ **Epistemic uncertainty:**
  - ➢ Refers to as model uncertainty comes from the lack of knowledge about the true parameters that generated the data.
  - ➢ Can be explained away by adding more training data.

Uncertainty Quantification and evaluation metrics

Uncertainty Quantification:

- **Variance of Predictions** compute the uncertainty of the predictions by simply taking the variance of predicted values.
- The **predictive entropy** is a measure of how much information is in the model predictive density function.
- **Mutual information** is the difference between the entropy of the predictions and the expected value of the entropy of the predictions.

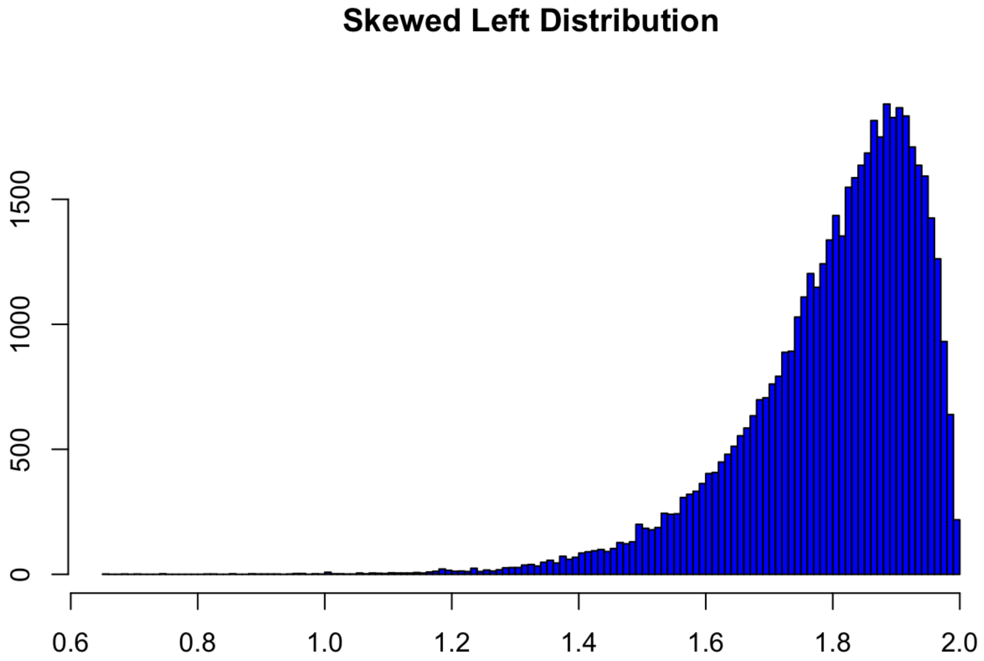
Calibration Metrics:

- **Brier Score** can be defined as the mean of the squared difference of the one-hot encoded ground truth, y and prediction probabilities $\widehat{p}$.
- **Reliability Diagrams** are a visual representation of model calibration that plots samples confidence vs. accuracy.
- **Expected Calibration Error (ECE)** can be defined as the expected difference in the accuracy and confidence of the predictions.
- **Negative log likelihood(NLL)** can be defined as the cross-entropy loss.

Uncertainty Quantification and evaluation metrics

Uncertainty Evaluation Metrics:

- **Probability Density Functions (PDF)** of the entropy to assess the quality of uncertainty.
- For Incorrect predictions and samples that their distribution is different from training data we expect the model assign high uncertainty which means that the mode of the histogram concentrates on high values.

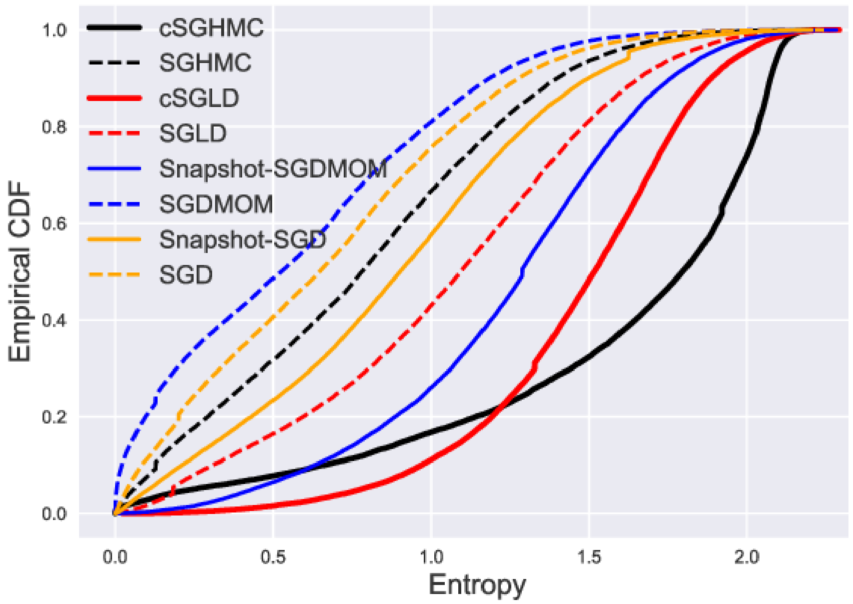


Uncertainty Quantification and evaluation metrics

Uncertainty Evaluation Metrics:

- **Cumulative Distribution Function (CDF)** of the predictive entropy to assess the quality of the uncertainties.

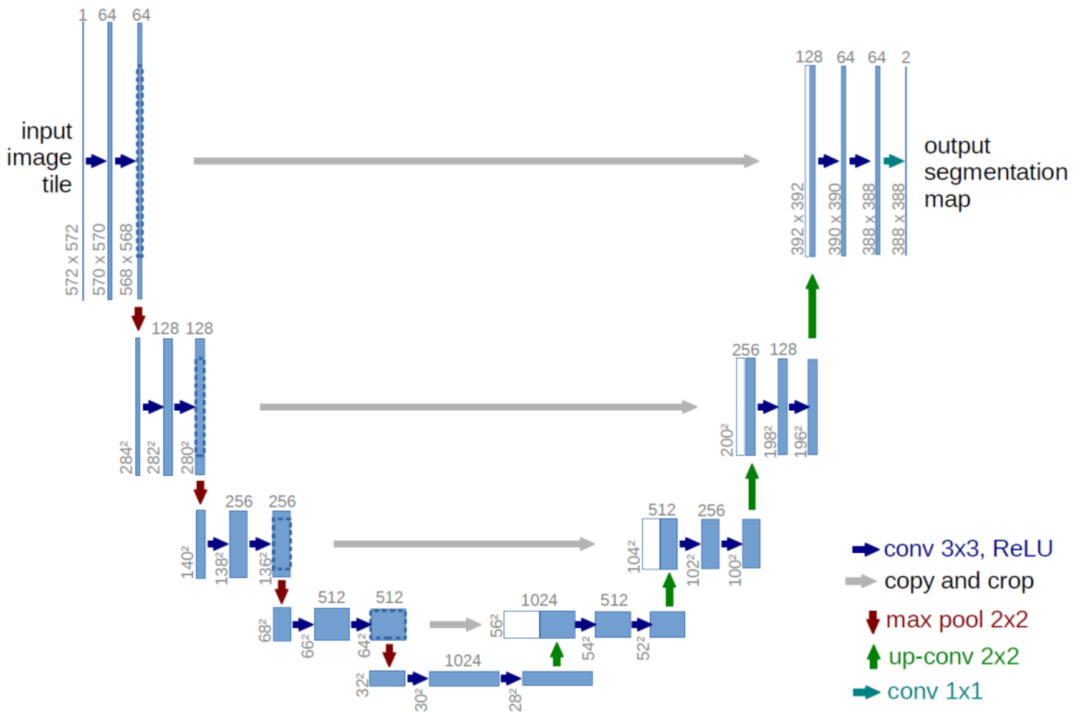
- A good model assigns high uncertainties to the OOD samples.



Segmentation Architecture and Performance Metrics

U-net Architecture:

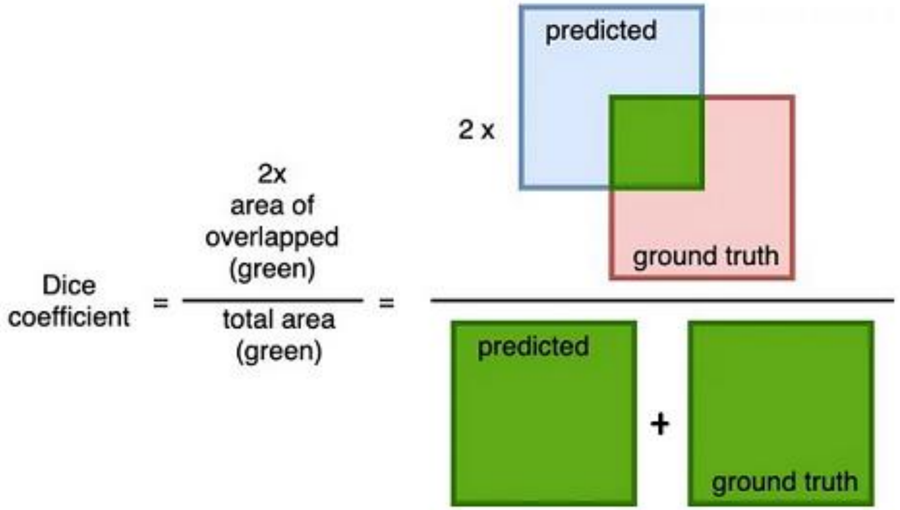
- The architecture has two paths, first the downsampling path to learn the spatially high-resolution features, then an upsampling path that learns high resolution features to construct the final segmentation map and assign the class labels.
- Skip connection propagates the information of local features from the earlier layers of the downsampling path to the higher resolution layers of the upsampling path.



Segmentation Architecture and Performance Metrics

Segmentation Performance Metrics:

- ❑ Dice Score can be defined as the pixel-wise measurement between the ground truth mask *y* and its predicted segmentation mask *ŷ*.
- ❑ It can be defined as two times the area of the union divided by the total number of voxels of both images.



Research questions that we want to answer

- Among different family of UQ methods, do the ones that capture multimodal posterior distribution, estimate uncertainty better?

- Does calibration in the ID setting translate to calibration under dataset shift?

- How do uncertainty and performance of different methods co-vary under dataset shift?

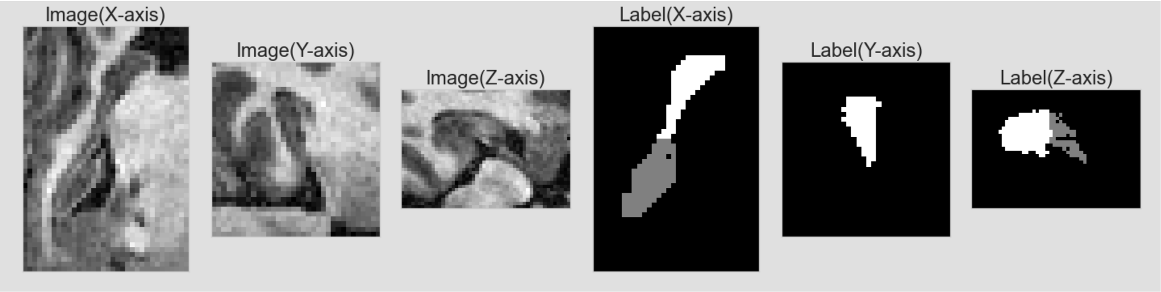
Experimental Setting

- **Data**:
  - Hippocampus
  - AMOS2022
  - KiTS21
- 3D images in Nifty format
- **Task**: Segmentation
- **Models**:
  - MC-Dropout
  - Deep Ensemble
  - SG-HMC
  - Vanilla

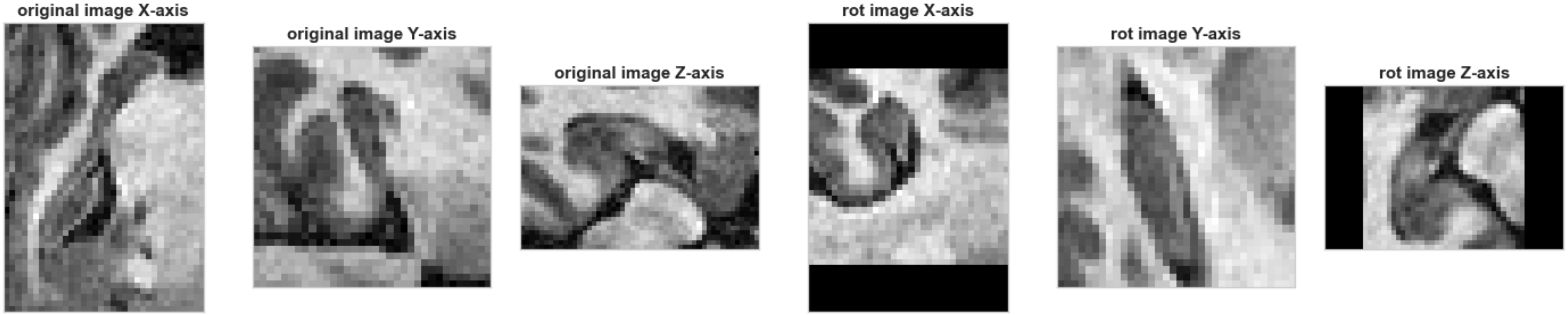
- **Distribution Shift:**
  - Rotation
  - Adding Gaussian blur
  - Different modalities
  - Corrupted version of the In-distribution test dataset

Datasets: Hippocampus

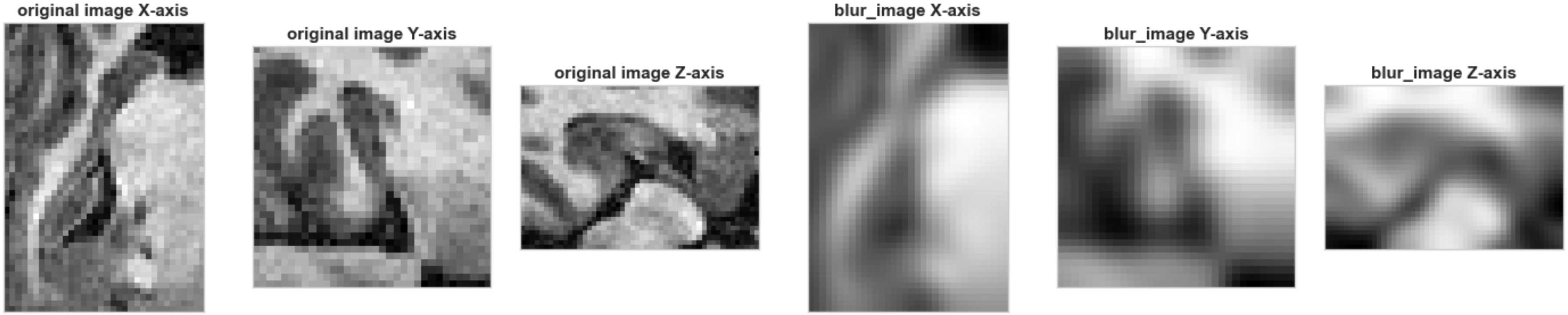
- ❑ Size: 260 3D labelled MRI
- ❑ Task: Segmentation of the left and right parts of the hippocampus.
- ❑ Distribution shift: Adding noise and rotated the test samples.



Original images and their correspondings rotated images

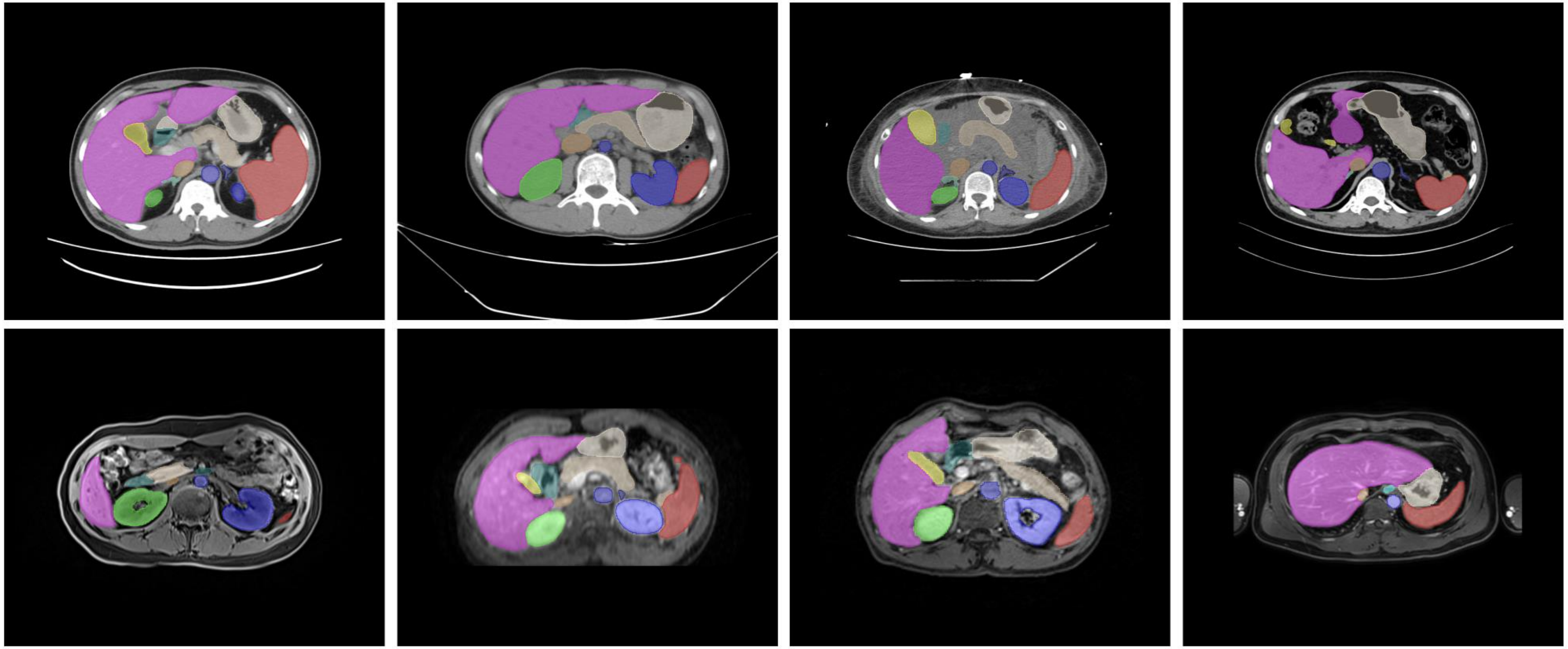


Original images and their correspondings blurred images



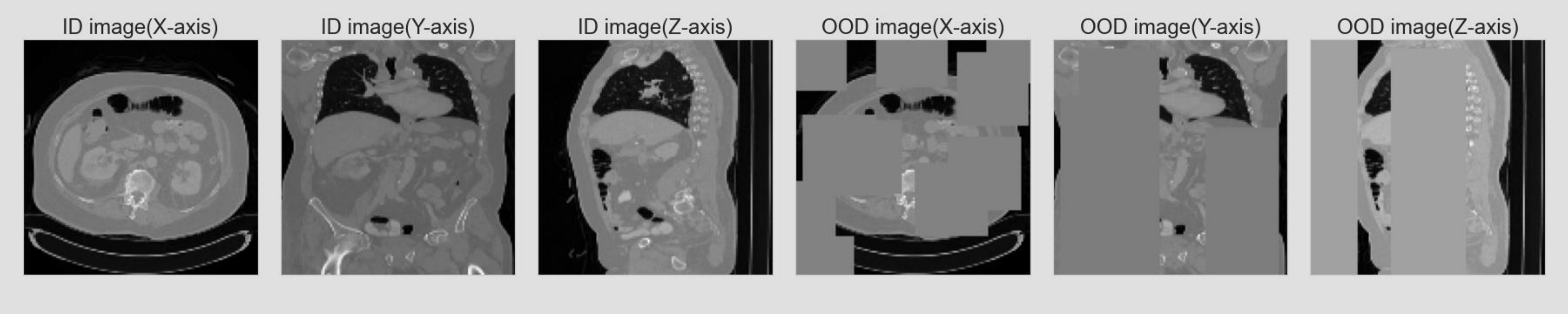
Datasets: AOMS

- ❖ Size: 200 CT and 40 MRI
- ❖ Task: segmentation of 15 abdominal organs
- ❖ Distribution shift: Different modalities
- ❖ ID : CT (top row), OOD: MRI (bottom row)



Datasets: KITS

- ❑ Size: 300 3D images
- ❑ Task: Tumors, and kidneys segmentation
- ❑ Distribution shift: Corrupted version of the test images.
- ❑ Adding some rectangles filled with zeroes to the random position of the ID test images

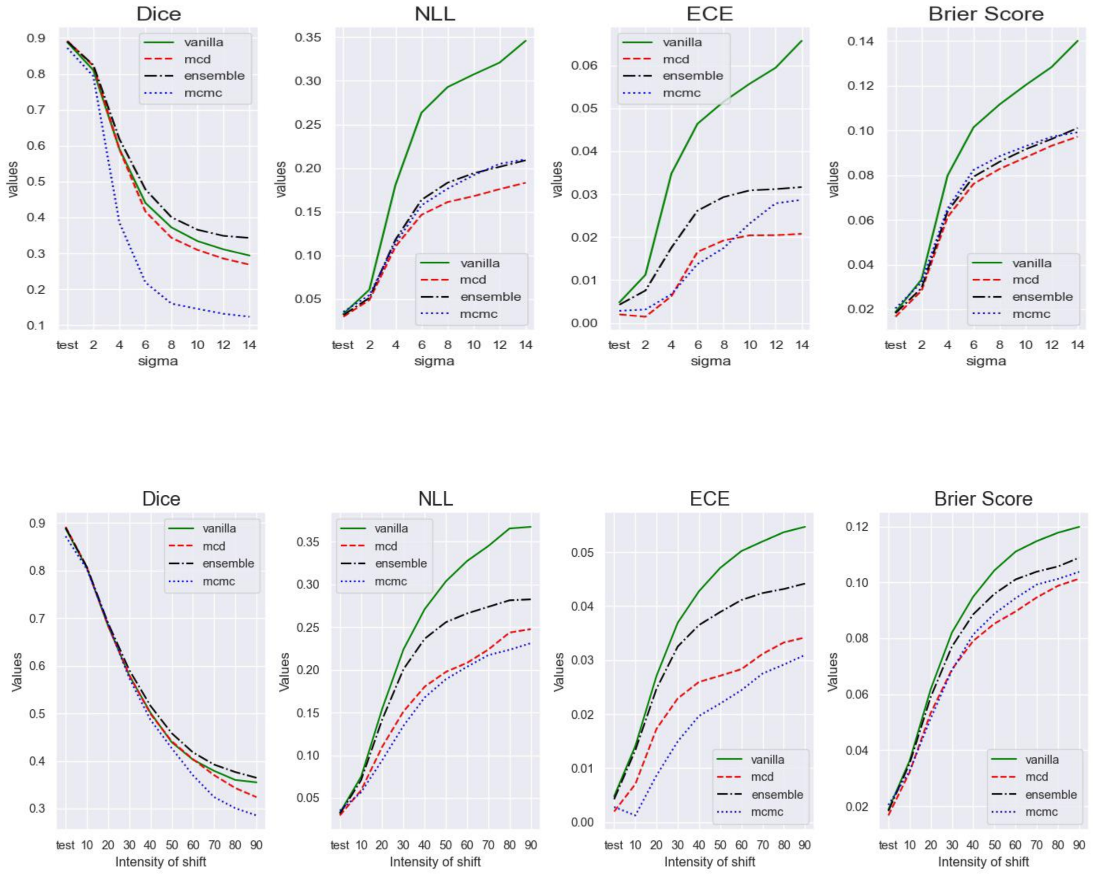


Results: Hippocampus Data

The Representation of the segmentation performance and calibration metrics for all models of blurred data (top) and the rotated data(bottom).

The dice score of the shifted test sets deteriorate as the intensity of the shift increases.

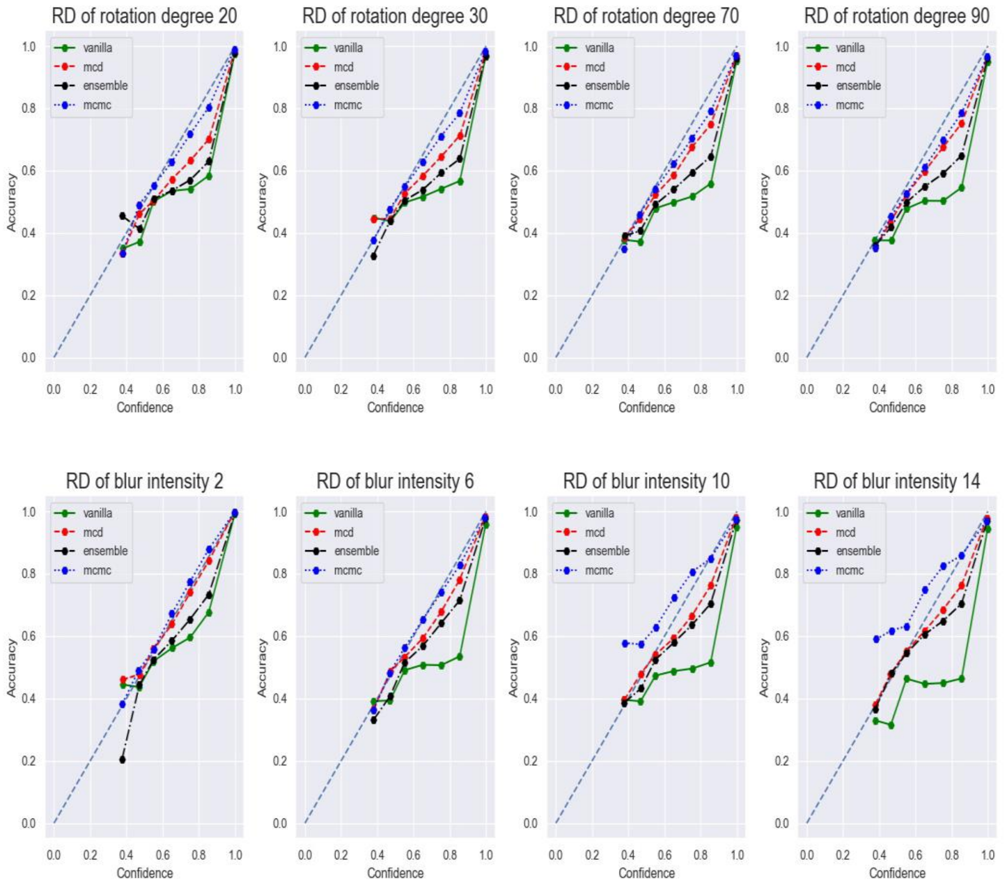
The mis-calibration gradually increases as the shift increases.



Results: Hippocampus

Reliability Diagram of the blurred (top) and rotated (bottom) OOD datasets.

Those models are closer to the blue horizontal line, are the well calibrated model on OOD test sets.

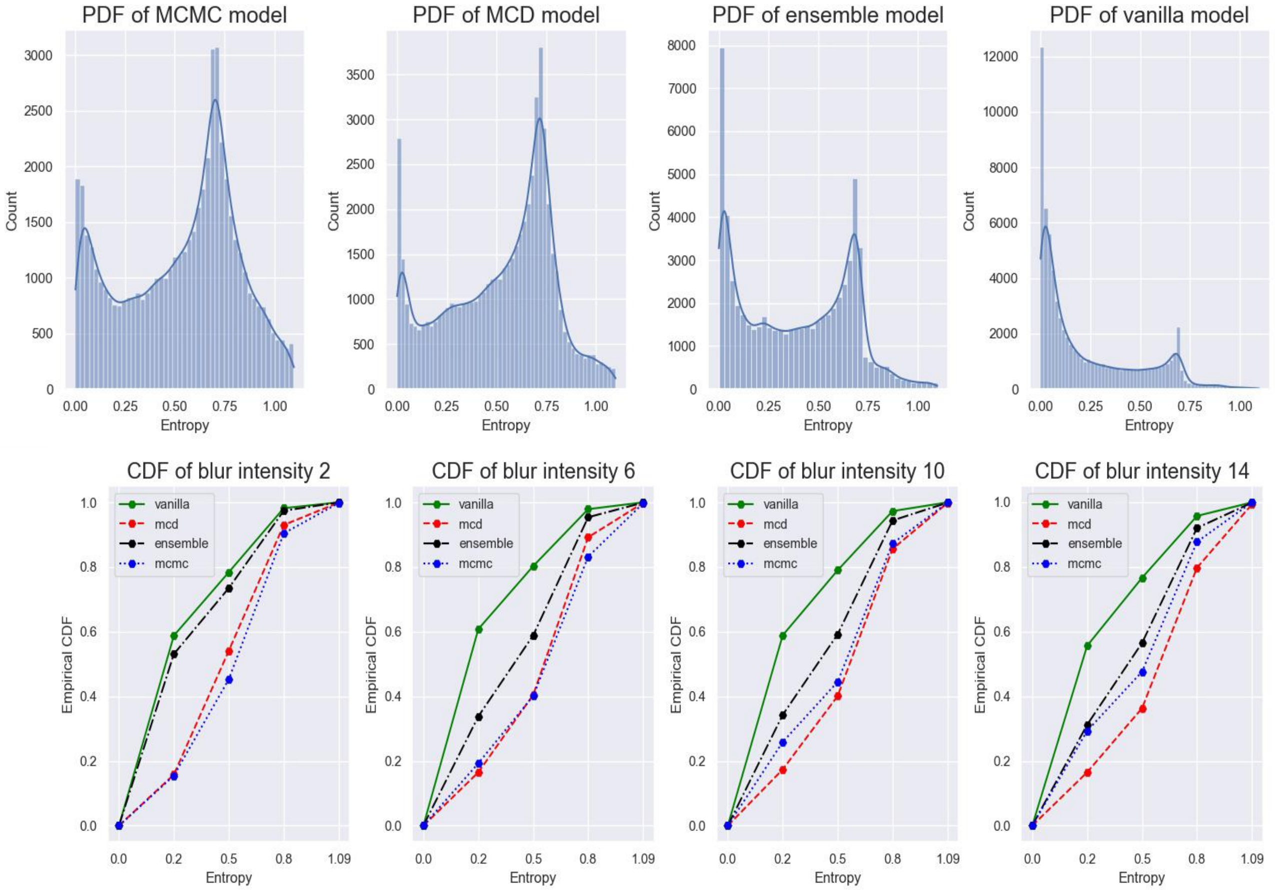


Results: Hippocampus

- ❖ PDF plot of all models for the blurred dataset of intensity 6 (top).

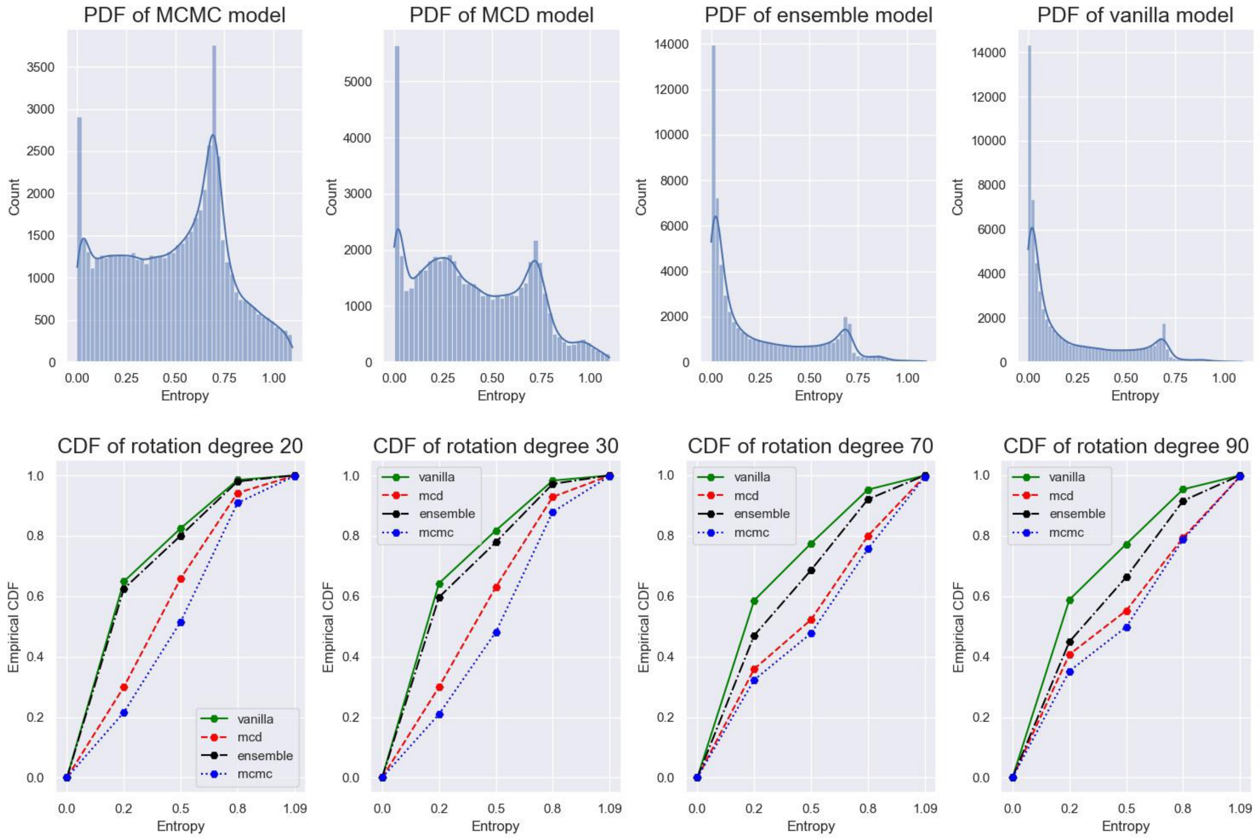
- ❖ Empirical CDF for the entropy of predictive distribution of blurred dataset (bottom).

- ❖ For different intensities, MCMC and MCD capture uncertainties better than other models.



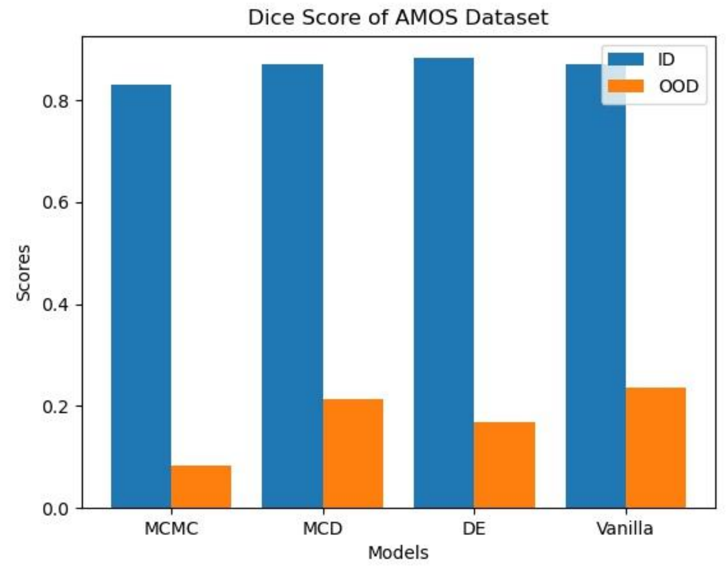
Results: Hippocampus

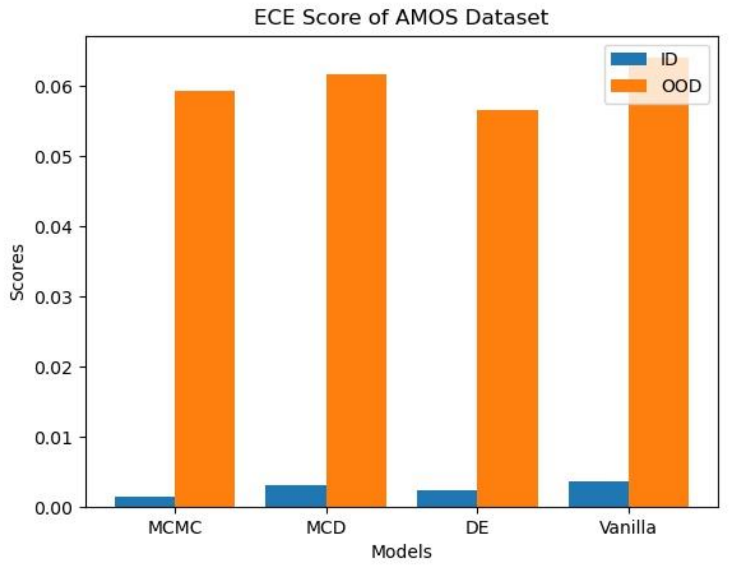
- PDF plot of all models for the rotated dataset with rotated degree of 30.
- The Empirical CDF of the entropy for several rotated datasets with rotation angle 20, 30, 70, and 90 degrees.
- For all rotation shifts, the MCMC model is estimating uncertainty better than other models.



Results: AMOS Data

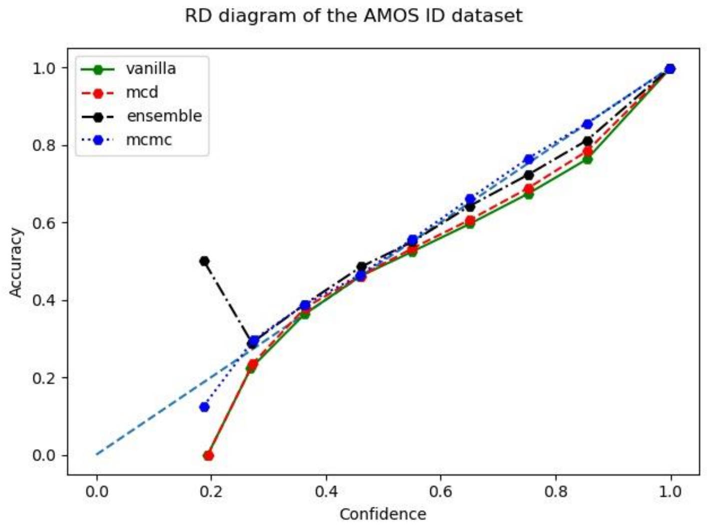
- For the shifted test samples, the Dice score has significantly degraded.
- Deep ensemble and MCMC model is providing the lowest ECE score than others model.

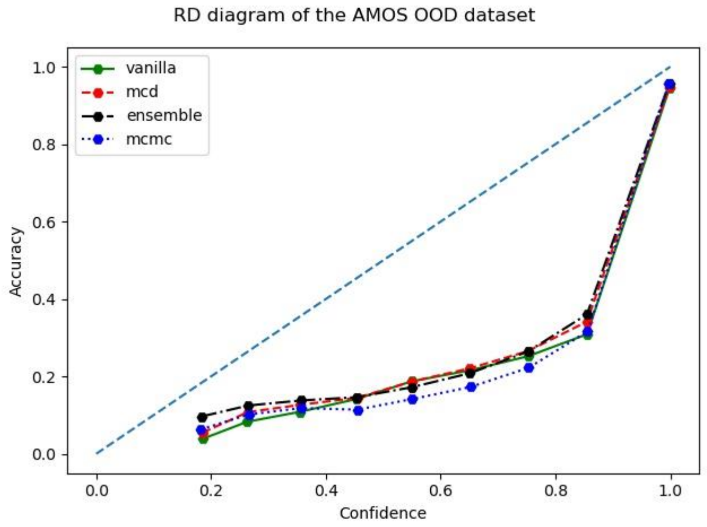




Results: AMOS Data

- For ID test samples, the MCMC model is well calibrated.
- For OOD dataset, the models are not well-calibrated.


RD diagram of the AMOS ID dataset

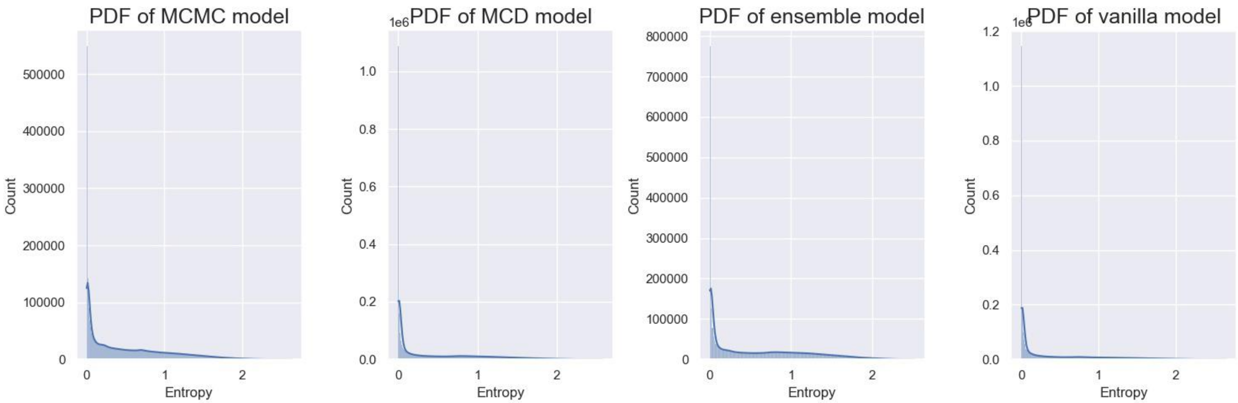

RD diagram of the AMOS OOD dataset

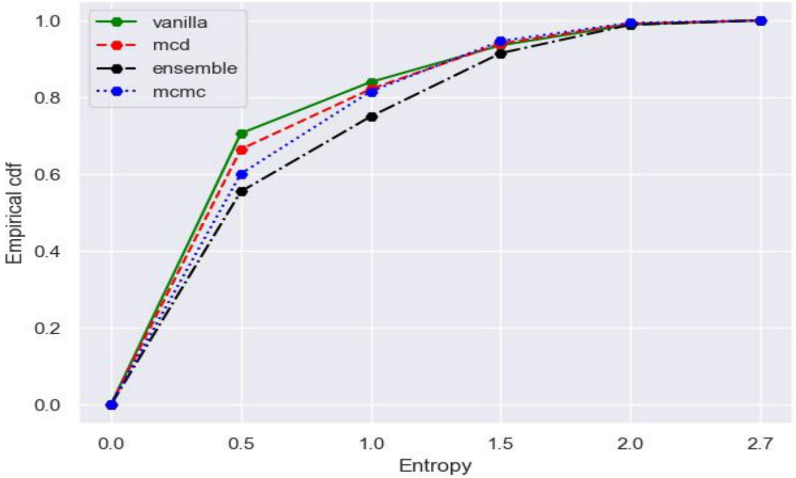
Results: AMOS Data

- The PDF shows that all models assign low entropy to the data with different modalities.

- That means the distribution of the data across different modalities does not change too much

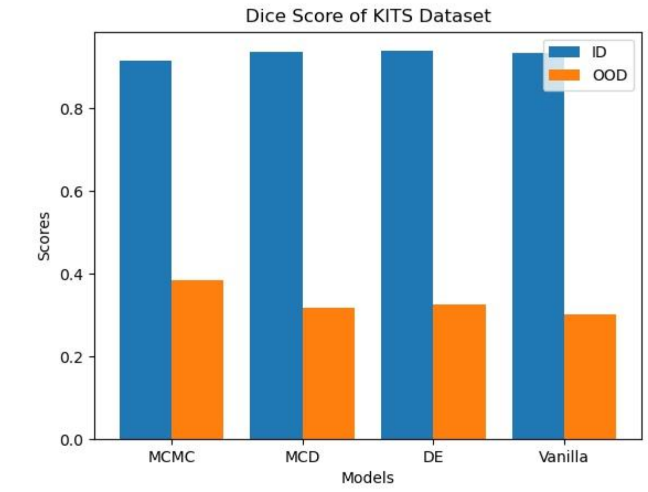
- Empirical CDF also provides the same results. Emphasizing MCMC and Deep Ensemble capture uncertainty better.

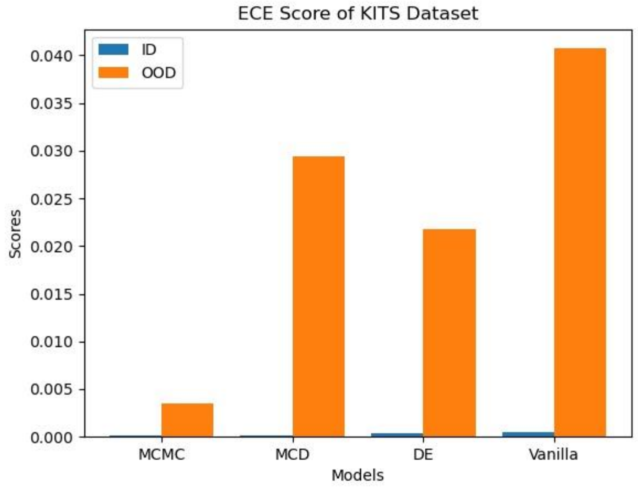




Results:
KITS Data

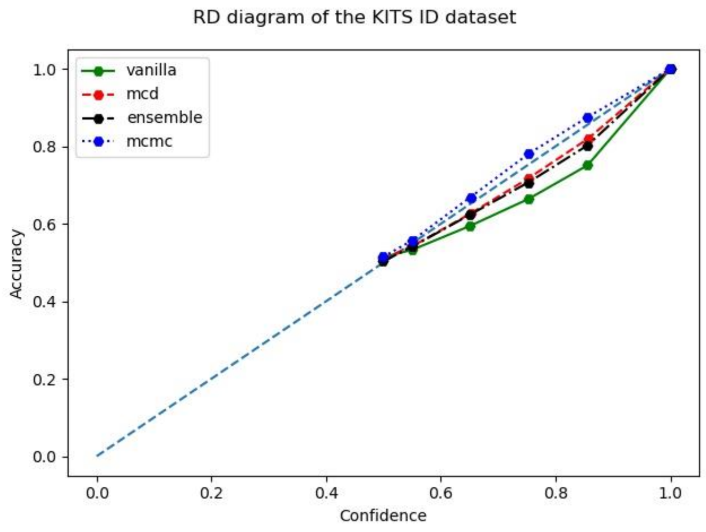
- For Corrupted test samples test samples, the Dice score has reduced drastically.
- MCMC is providing lowest ECE score.

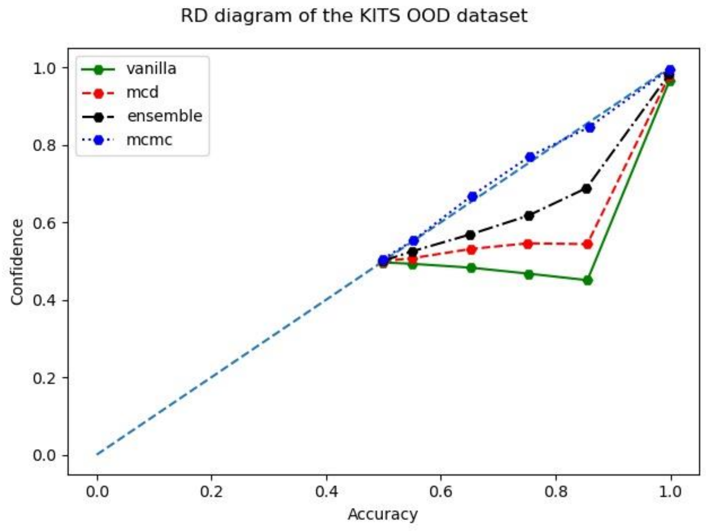




Results: KITS Data

- MCMC is well-calibrated model for both ID and shifted samples.

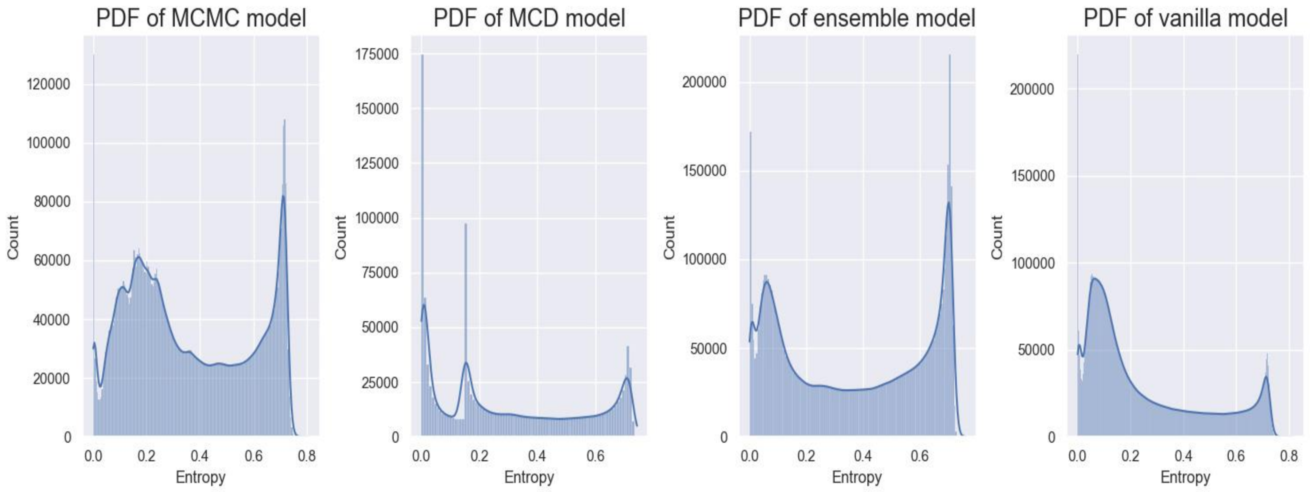


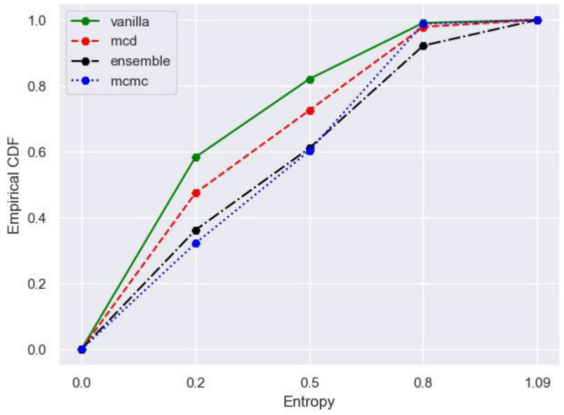


Results: KITS Data

- The PDF plot indicates that MCMC and the Deep Ensemble models assign higher uncertainty to corrupted images

- The same result can be seen in CDF plot too

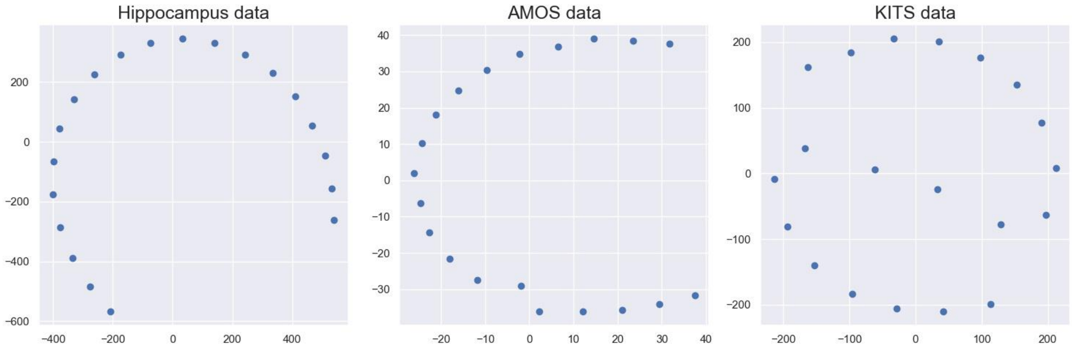




Conclusion

❖ Overall, the cSG-MCMC model provides well-calibrated uncertainty estimation in most cases.

❖ cSG-MCMC is able to explore, capture , and characterize the multimodal posterior distribution resulting in reliable predictive uncertainty.



❖ MCD, provides a good uncertainty estimation of the Hippocampus dataset, but for the other two datasets, it could not provide a reliable uncertainty estimation.

❖ Deep Ensemble also performs well on AMOS and KITS datasets, but it also fails to capture uncertainty on Hippocampus data set well. Moreover, training 20 models from scratch is way expensive than training a single MCMC model and collecting 20 samples from it

Conclusion

- Calibration in ID does not necessarily translate to the distributional shift. For Hippocampus and Amos datasets, the models are well calibrated for ID data, but they are mis-calibrated for shifted data with increasing in mis-calibration as distributional shift increases.
- For synthetic distribution shift we see that the performance degrades as shift increases and the reduction in performance coincides with increased entropy (Hippocampus and Kits). But for natural distribution shift, we see that the reduction in performance does not translate to high uncertainty (Amos).