

---

# Disentangled Uncertainty and Out of Distribution Detection in Medical Generative Models

---

Kumud Lakara<sup>1</sup>, Matias Valdenegro-Toro<sup>2</sup>

<sup>1</sup> Department of CSE, Manipal Institute of Technology, 576104 Manipal, India.

<sup>2</sup> Department of AI, University of Groningen, 9747AG Groningen, The Netherlands.

lakara.kumud@gmail.com, m.a.valdenegro.toro@rug.nl

## Abstract

Trusting the predictions of deep learning models in safety critical settings such as the medical domain is still not a viable option. Distentangled uncertainty quantification in the field of medical imaging has received little attention. In this paper, we study disentangled uncertainties in image to image translation tasks in the medical domain. We compare multiple uncertainty quantification methods, namely Ensembles, Flipout, Dropout, and DropConnect, while using CycleGAN to convert T1-weighted brain MRI scans to T2-weighted brain MRI scans. We further evaluate uncertainty behavior in the presence of out of distribution data (Brain CT and RGB Face Images), showing that epistemic uncertainty can be used to detect out of distribution inputs, which should increase reliability of model outputs.

## 1 Introduction

Translating an image from one domain to another is a challenging task as there may not necessarily be a one-one mapping between the two domains. Generative models are often used to learn this one-to-one mapping in an unsupervised manner with an additional constraint on the image or feature space. This constraint imposes structure on the underlying joint distribution of the two images being translated from different domains [34; 30; 52; 37]. In the medical domain image-to-image translation is very significant. It has a myriad of applications including denoising [14], motion correction [3] and MRI scan translation. Generative adversarial networks have achieved state of the art performance on these tasks. However, they usually do not provide uncertainty estimates for their predictions. In addition to this, existing models usually learn a deterministic mapping between the two domains being translated. In the medical domain such mapping cannot quantify the model uncertainty to predict trustworthy outputs. Moreover, most state-of-the-art methods do not account for out of distribution (OOD) data and often fail under such conditions thus making them unreliable and incapable of real world deployment. Incorporation of uncertainty into image to image translation tasks is of prime importance, especially in the medical domain in order to make informed medical decisions.

When considering uncertainty, aleatoric uncertainty is caused by inherent noise and randomness in the data itself and is hence irreducible [7; 26], while epistemic uncertainty on the other hand is due to the model's inability to learn effectively from the data and can be reduced by training with more data. These values are usually predicted as a single value called predictive uncertainty [13]. Disentangling epistemic and aleatoric uncertainties is of great value when it comes to out of distribution detection [17]. This is of particular importance in the case of critical applications including healthcare and medical diagnosis [10; 11].

This paper shows how a CycleGAN [52] generative model trained on medical images for image to image translation (Brain MRI to Brain CT) can be combined with uncertainty quantification

methods to produce disentangled estimates of aleatoric and epistemic uncertainty, while at the same time performing out of distribution detection. The contributions are a simple formulation for uncertainty disentanglement in CycleGAN, and a comparison of different uncertainty quantification methods (Flipout, Dropout, DropConnect, Ensembles) for both image translation with (disentangled) uncertainty and out of distribution detection.

## 2 Related Work

Past work on image-to-image translation in the medical domain has focussed on convolutional neural networks [5; 9; 16; 25; 29] and generative adversarial networks [1; 2; 22; 33; 38; 39; 47; 48; 50]. Some machine learning techniques also rely on explicit feature representations [21; 27; 35]. The existing methods usually work by enforcing additional constraints on the joint distribution of the images from different domains. [26; 40; 43] explore the shortcomings of using conditional GAN architectures with deterministic outputs. Though state-of-the-art models are capable of synthesizing high quality output images, the image is often very different from the ground-truth. This could result in various problems including over-confidence and mis-interpretation which could in turn lead to wrong diagnoses, misinformed decisions, delayed treatment and other severe consequences. Hence, in the medical domain confidence and certainty in the model’s predictions are of utmost value. Moreover, being able to evaluate uncertainty as two separate constituents further enhances the quality of confidence in the model’s predictions. Previous work [26; 6; 20] discusses methods to calculate the total uncertainty as well as individual epistemic and aleatoric uncertainty. Epistemic uncertainty arises in the model parameters due to training with finite data and aleatoric uncertainty arises due to the noise or uncertainty that is inherently present in the data itself [23; 26]. There have been numerous attempts to quantify aleatoric and epistemic uncertainty in medical imaging tasks including classification, segmentation and more [32; 36; 43; 42; 44] however, uncertainty quantification for image-to-image translation tasks still remains largely unexplored.

Image-to-image translation tasks in the medical domain have recently come to the fore-front for computer vision based research in the medical field. [49; 2; 24; 45; 15]. Uncertainty aware image-to-image translation models have also been proposed with special emphasis on CT and MRI images. Upadhyay et. al[40] propose a probabilistic method that models the per-pixel residual by generalized Gaussian distribution which helps increase the model’s robustness to OOD data. [41] propose an uncertainty guided progressive learning scheme for image-to-image translation. However, none of the current methods focus on the individual behavior of epistemic and aleatoric uncertainties for image-to-image translation tasks in the medical domain and that is what we address in this work.

## 3 CycleGAN and Uncertainty Disentanglement

CycleGAN provides a method to translate an image from a source domain (say  $P$ ) to a target domain (say  $Q$ ) in the absence of paired examples. The basic idea is to learn a mapping  $G : P \rightarrow Q$  in a manner such that the distribution of resultant images ( $G(P)$ ) is identical to the distribution of  $Q$ . This is achieved by enforcing an additional structure on the joint distribution. The initial  $G : P \rightarrow Q$  mapping is coupled with an inverse  $F : Q \rightarrow P$  mapping and a *cycle consistency* loss is used.

$$L_{\text{cyc}} = \mathbb{E}_{x \sim p_{\text{data}(x)}} [\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}(y)}} [\|G(F(y)) - y\|_1] \quad (1)$$

To disentangle the uncertainty value into epistemic and aleatoric uncertainty we consider the mean ( $\mu$ ) and variance ( $\sigma^2$ ) output from the model. The variance head of the model is trained using the Gaussian negative log-likelihood loss:

$$L_{NLL}(y_n, \mathbf{x}_n) = \frac{\log \sigma_i^2(\mathbf{x}_n)}{2} + \frac{(\mu_i(\mathbf{x}_n) - y_n)^2}{2\sigma_i^2(\mathbf{x}_n)}. \quad (2)$$

This replaces the L1 distance in the cycle consistency loss, which now becomes:

$$L_{\text{cyc}} = \mathbb{E}_{x \sim p_{\text{data}(x)}} [L_{NLL}(F(G(x)), x)] + \mathbb{E}_{y \sim p_{\text{data}(y)}} [L_{NLL}(G(F(y)), y)] \quad (3)$$

During inference, we sample weights from a distribution  $\theta \sim p(\theta|x, y)$ , which produce different predictions for  $\mu$  and  $\sigma^2$  which correspond to the approximate predictive posterior distribution of the model. These obtained samples are then combined to obtain the Gaussian mixture distribution:

$$\mu_*(x) = M^{-1} \sum_i \mu_i(x) \quad \sigma_*^2(x) = M^{-1} \sum_i (\sigma_i^2(x) + \mu_i^2(x)) - \mu_*^2(x) \quad (4)$$

The predictive variance can further be decomposed as follows:

$$\begin{aligned}\sigma_*^2(x) &= M^{-1} \sum_i \mu_i^2(x) - \mu_*^2(x) = \mathbb{E}_i[\sigma_i^2(x)] + \mathbb{E}_i[\mu_i^2(x)] - \mathbb{E}_i[\mu_i(x)]^2 \\ &= \underbrace{\mathbb{E}_i[\sigma_i^2(x)]}_{\text{Aleatoric Uncertainty}} + \underbrace{\text{Var}_i[\mu_i(x)]}_{\text{Epistemic Uncertainty}}\end{aligned}\quad (5)$$

From the above derivation, we posit that the mean of the variances corresponds to aleatoric uncertainty where as the variance of the means gives us the epistemic uncertainty. The common intuition here is that using the Gaussian NLL loss, each variance head estimates aleatoric uncertainty, so taking the mean of these variance heads, produces a combined aleatoric uncertainty, while taking the variance through ensemble members or samples from a forward pass corresponds to disagreement between ensembles/samples, large disagreement indicates that the whole ensemble/samples are not sure of the prediction, while a low disagreement indicates a large confidence in the prediction, both of these concepts measure epistemic uncertainty.

## 4 Experiments and Results

We provide a comprehensive overview of different uncertainty quantification methods as described in section A in their ability to disentangle epistemic uncertainty from aleatoric uncertainty. For the purpose of our experiments we take 3 datasets: IXI Brain MRI Scans [8], Computed Tomography Images for Intracranial Hemorrhage Detection and Segmentation version 1.3.1( referred to as Brain CT Scans dataset hereafter) [19; 18], UTKFace dataset [51]. The IXI Brain MRI Scans dataset is used for training and the CT scans and UTKFace datasets are used for out of distribution analyses of the model. The uncertainty values are normalized using min-max normalization for the purpose of analyses.

Figures 1 through 4 show the respective histograms for the various method-dataset combinations. We notice that for most cases, when a model is provided out of distribution data, the epistemic uncertainty increases (sometimes very drastically as seen in figure 3) where as the aleatoric uncertainty remains in a relatively similar range compared to the non-OOD data counterpart. This is also clearly visible in figures 5 through 6 where the uncertainty histograms for the in-distribution (ID) dataset (IXI brain scans) and the OOD datasets are plotted together. For most cases a clear distinction can be made between the plots for in-distribution and OOD data. Moreover, epistemic uncertainty can be seen to increase in the case of OOD data. This can be used to detect outliers and further improves model robustness to OOD data.

Qualitative results are presented in tables 2 through 3. Table 4 shows the ROC curves for aleatoric and epistemic uncertainty for various methods. Epistemic uncertainty is more reliable for OOD detection in most but not all methods, while aleatoric uncertainty fails to separate ID from OOD data in all but one case (Flipout).

We believe these results show that out of distribution inputs can be successfully detected using a image to image translation generative model such as CycleGAN, when combining it with a uncertainty quantification method, and that disentangled estimates of aleatoric and epistemic uncertainty allow for separate modeling of model errors and noise in the data, which we believe can improve interpretation of visual results, which is generally not possible with standard predictive uncertainty quantification.

## 5 Conclusions and Future Work

In this paper we presented a novel method for uncertainty disentanglement using CycleGAN for image-to-image translation in the medical domain. We analyzed various uncertainty quantification methods and their ability to disentangle predictive uncertainty into epistemic and aleatoric uncertainty estimates.

We drew insights related to uncertainty values/distributions and out of distribution data. Our results show the differences between different uncertainty quantification methods and how they behave when exposed to out of distribution data. In particular when combining disentangled uncertainties with a generative model, out of distribution inputs can be detected by looking at the epistemic uncertainty

of the output image, while aleatoric uncertainty does not produce relevant information about out of distribution inputs, but does provide insights about noise in the input data.

We believe these results show the potential of disentangled uncertainties in image to image generative models for medical data. Without uncertainty quantification, the medical worker that has to interpret these images and obtain useful information for diagnosis, would not be aware if the model is unsure of its output, if out of distribution inputs were provided, or if the data is too noisy, which would propagate through the model (garbage in, garbage out).

We see multiple avenues of future research including fast prediction of uncertainty estimates, further evaluation of the quality of uncertainty, and evaluating these methods on more realistic medical imaging settings. We hope our work inspires further research in the field of uncertainty and out of distribution detection particularly in the field of medical imaging.

## References

- [1] K. Armanious, C. Jiang, S. Abdulatif, T. Küstner, S. Gatidis, and B. Yang. Unsupervised medical image translation using cycle-medgan. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.
- [2] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang. Medgan: Medical image translation using gans. *Computerized medical imaging and graphics*, 79:101684, 2020.
- [3] K. Armanious, A. Tanwar, S. Abdulatif, T. Küstner, S. Gatidis, and B. Yang. Unsupervised adversarial correction of rigid mr motion artifacts. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1494–1498. IEEE, 2020.
- [4] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.
- [5] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang. Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE transactions on medical imaging*, 36(12):2524–2535, 2017.
- [6] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR, 2018.
- [7] A. Der Kiureghian and O. Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.
- [8] B. Development. Ixi dataset. Available at <http://brain-development.org/ixi-dataset> (2022/09/28).
- [9] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng. Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection. *IEEE Transactions on Biomedical Engineering*, 64(7):1558–1567, 2016.
- [10] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.
- [11] A. Filos, S. Farquhar, A. N. Gomez, T. G. Rudner, Z. Kenton, L. Smith, M. Alizadeh, A. De Kroon, and Y. Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*, 2019.
- [12] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
- [13] Y. Gal et al. Uncertainty in deep learning. 2016.

- [14] J. Gu and J. C. Ye. Adain-based tunable cyclegan for efficient unsupervised low-dose ct denoising. *IEEE Transactions on Computational Imaging*, 7:73–85, 2021.
- [15] C. Han, L. Rundo, R. Araki, Y. Nagano, Y. Furukawa, G. Mauri, H. Nakayama, and H. Hayashi. Combining noise-to-image and image-to-image gans: Brain mr image augmentation for tumor detection. *Ieee Access*, 7:156966–156977, 2019.
- [16] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.
- [17] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
- [18] M. D. Hssayeni, M. S. Croock, A. D. Salman, H. F. Al-khafaji, Z. A. Yahya, and B. Ghoraani. Computed tomography images for intracranial hemorrhage detection and segmentation (version 1.3.1). URL <https://doi.org/10.13026/4nae-zg36>.
- [19] M. D. Hssayeni, M. S. Croock, A. D. Salman, H. F. Al-khafaji, Z. A. Yahya, and B. Ghoraani. Intracranial hemorrhage segmentation using a deep convolutional model. *Data*, 5(1):14, 2020.
- [20] E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.
- [21] T. Huynh, Y. Gao, J. Kang, L. Wang, P. Zhang, J. Lian, and D. Shen. Estimating ct image from mri data using structured random forest and auto-context model. *IEEE transactions on medical imaging*, 35(1):174–183, 2015.
- [22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [23] H. D. Kabir, A. Khosravi, M. A. Hosen, and S. Nahavandi. Neural network-based uncertainty quantification: A survey of methodologies and applications. *IEEE access*, 6:36218–36234, 2018.
- [24] S. Kaji and S. Kida. Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging. *Radiological physics and technology*, 12(3):235–248, 2019.
- [25] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.
- [26] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.
- [27] T. Küstner, M. Schwartz, P. Martirosian, S. Gatidis, F. Seith, C. Gilliam, T. Blu, H. Fayad, D. Visvikis, F. Schick, et al. Mr-based respiratory and cardiac motion correction for pet imaging. *Medical image analysis*, 42:129–144, 2017.
- [28] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.
- [29] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
- [30] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.
- [31] A. Mobiny, P. Yuan, S. K. Moulik, N. Garg, C. C. Wu, and H. Van Nguyen. Dropconnect is effective in modeling uncertainty of bayesian deep networks. *Scientific reports*, 11(1):1–14, 2021.

- [32] T. Nair, D. Precup, D. L. Arnold, and T. Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557, 2020.
- [33] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen. Medical image synthesis with deep convolutional adversarial networks. *IEEE Transactions on Biomedical Engineering*, 65(12):2720–2730, 2018.
- [34] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer, 2020.
- [35] A. Rueda, N. Malpica, and E. Romero. Single-image super-resolution of brain mr images using overcomplete dictionaries. *Medical image analysis*, 17(1):113–132, 2013.
- [36] R. Tanno, D. E. Worrall, A. Ghosh, E. Kaden, S. N. Sotiropoulos, A. Criminisi, and D. C. Alexander. Bayesian image quality transfer with cnns: exploring uncertainty in dmri super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 611–619. Springer, 2017.
- [37] N.-T. Tran, T.-A. Bui, and N.-M. Cheung. Dist-gan: An improved gan using distance constraints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–385, 2018.
- [38] U. Upadhyay and S. P. Awate. A mixed-supervision multilevel gan framework for image quality enhancement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 556–564. Springer, 2019.
- [39] U. Upadhyay and S. P. Awate. Robust super-resolution gan, with manifold-based and perception loss. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1372–1376. IEEE, 2019.
- [40] U. Upadhyay, Y. Chen, and Z. Akata. Uncertainty-aware generalized adaptive cyclegan. *arXiv preprint arXiv:2102.11747*, 2021.
- [41] U. Upadhyay, Y. Chen, T. Hepp, S. Gatidis, and Z. Akata. Uncertainty-guided progressive gans for medical image translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 614–624. Springer, 2021.
- [42] G. Wang, W. Li, S. Ourselin, and T. Vercauteren. Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In *International MICCAI Brainlesion Workshop*, pages 61–72. Springer, 2018.
- [43] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.
- [44] G. Wang, W. Li, S. Ourselin, and T. Vercauteren. Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation. *Frontiers in computational neuroscience*, 13:56, 2019.
- [45] P. Welander, S. Karlsson, and A. Eklund. Generative adversarial networks for image-to-image translation on multi-contrast mr images-a comparison of cyclegan and unit. *arXiv preprint arXiv:1806.07777*, 2018.
- [46] Y. Wen, P. Vicol, J. Ba, D. Tran, and R. Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*, 2018.
- [47] J. M. Wolterink, A. M. Dinkla, M. H. Savenije, P. R. Seevinck, C. A. van den Berg, and I. Išgum. Deep mr to ct synthesis using unpaired data. In *International workshop on simulation and synthesis in medical imaging*, pages 14–23. Springer, 2017.
- [48] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*, 37(6):1348–1357, 2018.



- [49] Q. Yang, N. Li, Z. Zhao, X. Fan, E. I. Chang, Y. Xu, et al. Mri cross-modality image-to-image translation. *Scientific reports*, 10(1):1–18, 2020.
- [50] H. Zhang, V. Sindagi, and V. M. Patel. Image de-raining using a conditional generative adversarial network. *IEEE transactions on circuits and systems for video technology*, 30(11): 3943–3956, 2019.
- [51] S. Y. Zhang, Zhifei and H. Qi. Age progression or regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.
- [52] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

## A Uncertainty Quantification Methods

We employ multiple uncertainty quantification methods to achieve a more holistic view of their ability to disentangle predictive uncertainty into epistemic and aleatoric uncertainty.

**MC-Dropout** This method acts as a regularizer which is applied at both training and testing time. It zeros out random activations in a layer. Since MC-Dropout[12] is active during test time as well, the model becomes stochastic. Hence, each forward pass produces a unique sample from the Bayesian posterior distribution. For our experiments we take the drop probability as  $p = 0.5$

**Flipout** This is a variational inference method that models weights as an approximate Gaussian distribution [4]. Flipout [46] is employed to reduce training process variance and to improve learning stability and performance. The resulting model that uses Flipout is stochastic in nature. For the purpose of our experiments, we use Flipout in multiple layers with a disabled prior.

**Ensembles** This method[28] involves multiple training runs of the same model with different random weight initializations each time. The final output is predicted by considering the individual outputs from each of the member models of the ensemble. We make use of an ensemble of 5 member models for our experiments.

**MC-DropConnect** This method[31] sets random weights to zero. It is in a sense similar to MC-Dropout[12] and has a similar regularization effect. This method allows weights to be zeroed out at inference time which produces samples from the Bayesian posterior distribution as in the case of MS-Dropout[12]. We use DropConnect layers with a drop probability of  $p = 0.05$

## B Potential Negative Societal Impact

While the sole purpose of this work is to improve the quality of predictions made by medical generative models and make them more trustworthy, we can foresee some negative societal impacts as well.

A model is only as good as the data it is trained on hence unknown correlations within the data can reflect on the model. For uncertainty in particular, all existing uncertainty quantification methods are approximations and thus do not provide the true predictive posterior distribution which in turn brings into question the trustworthiness of predicted uncertainty values. Such methods could at times prove to be misleading and thus lead to wrong medical diagnoses.

In addition to this, it ultimately comes down to the medical professionals to consider the uncertainty predictions from the model while making their diagnoses. If medical professionals decide not to consider model uncertainty, it could lead to wrong diagnoses and further complications in treatment. Another important point of concern is privacy and anonymity. In order to make useful predictions, the model needs to train on patient data and learn from it and this can put personal medical records in a vulnerable position.

## C Qualitative Result Comparison

This section presents visual comparison of predicted image translation results, including disentangled aleatoric and epistemic uncertainty. We present outputs produced using three different inputs, one is in-distribution, and two are out of distribution.

Table 1: Qualitative results for IXI Brain Scans: Since the these results are for ID data we note the output is more defined compared to the other results. Moreover, the epistemic and aleatoric uncertainty images also have a definite structure to them which is not the case for the OOD data results which follow.

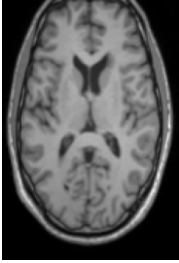
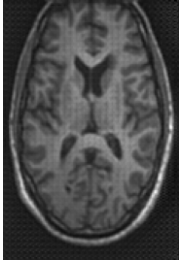

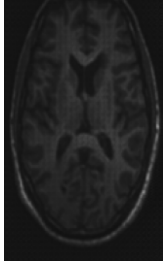


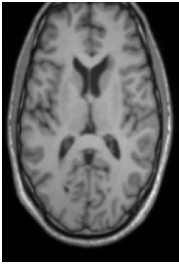
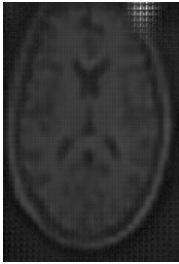
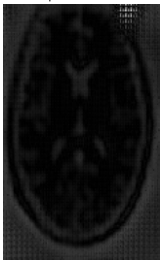
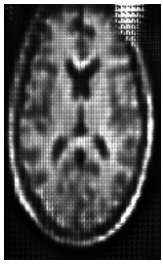
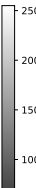
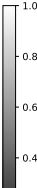
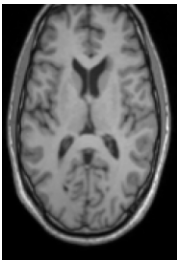
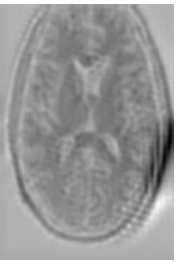

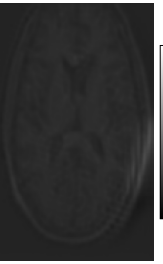

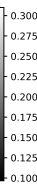
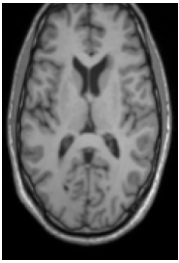
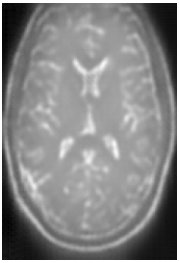

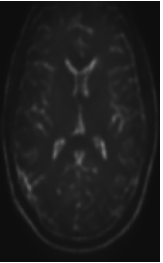


Method	Results			
MC-Dropout	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
				
Flipout	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
				
Ensemble	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
				
MC-DropConnect	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
				



Table 2: Qualitative results for Brain CT Scans: The model predictions make sense to some extent however since the CT data is OOD we can notice strange patterns in the images for epistemic and aleatoric uncertainties.

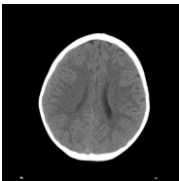
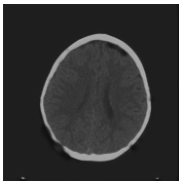
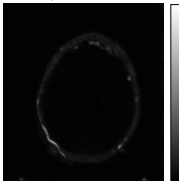
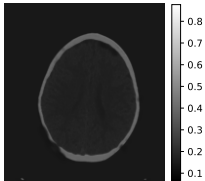
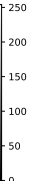
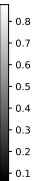
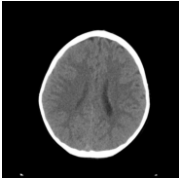
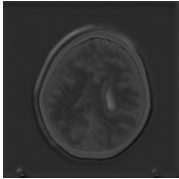
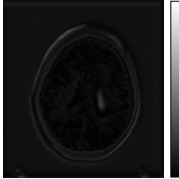
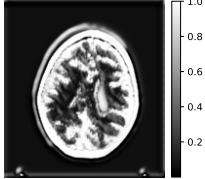
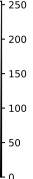

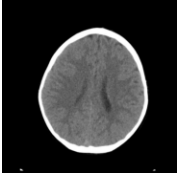
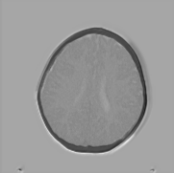
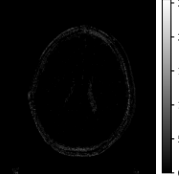
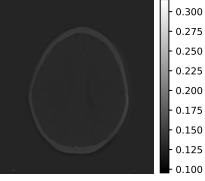
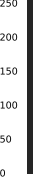
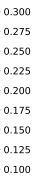
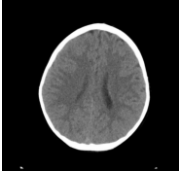
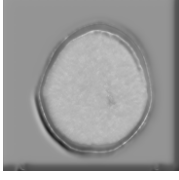
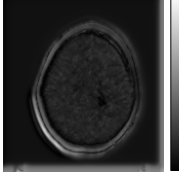
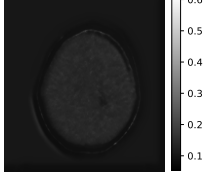
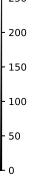



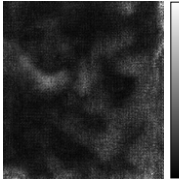
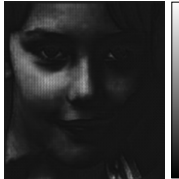






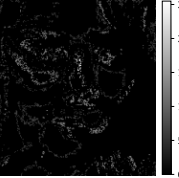




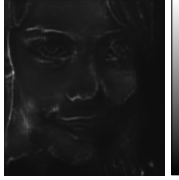





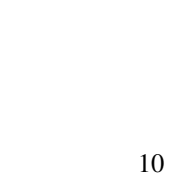






Method	Results			
MC-Dropout	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
				
Flipout	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
				
Ensemble	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
				
MC-DropConnect	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
				

Table 3: Qualitative results for UTKFace Dataset: The output images here do not make a lot sense and their quality is not as good as the ID resultant images. In addition to this, the epistemic and aleatoric uncertainty images in particular have strange patterns which are key to detecting out of distribution data.

Method	Results			
MC-Dropout	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
Flipout	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
Ensemble	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
MC-DropConnect	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC
				
	RealT1W	GenT2W	EpistemicUNC	AleatoricUNC

## D Out of Distribution Detection Results

This section presents all out of distribution detection results, including numerical AUC scores in Table 5, and ROC curves in Table 4. Each ID-OOD dataset combination is shown in the header of each table.

Table 4: ROC curves for IXI-CT (OOD) and IXI-UTKFace (OOD) dataset combinations

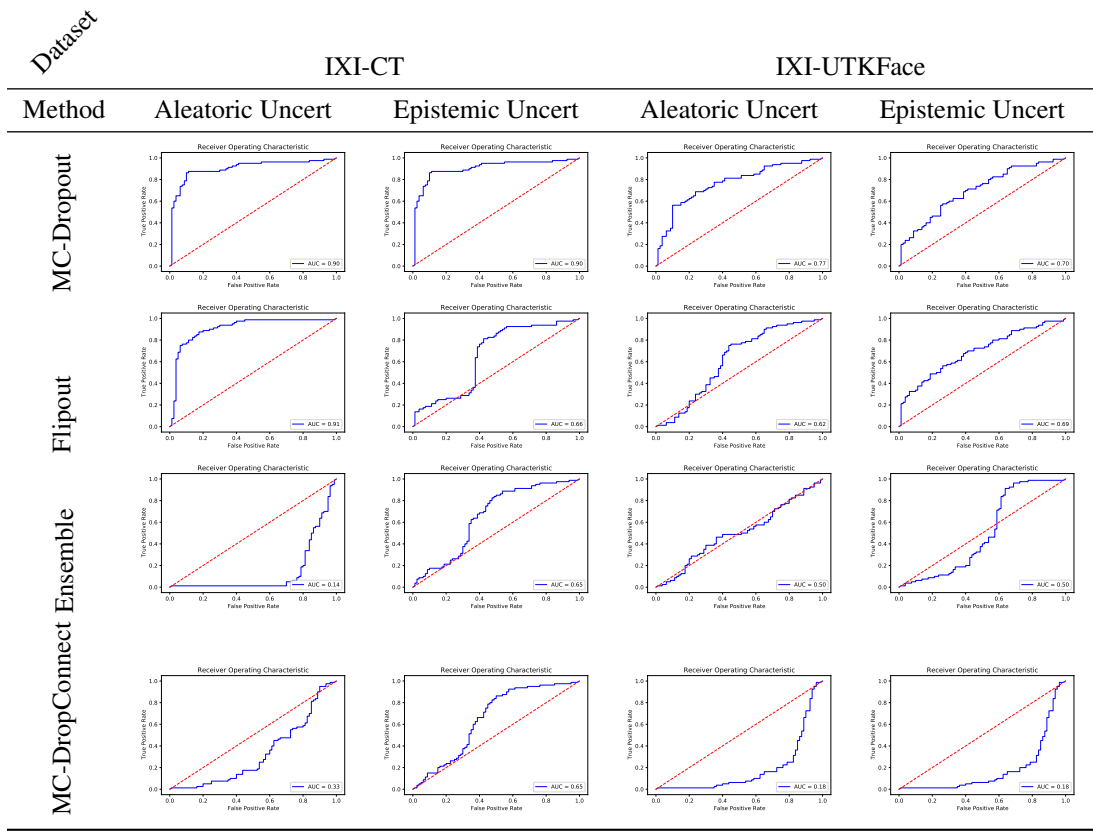


Table 5: AUC values for IXI-CT (OOD) and IXI-UTKFace (OOD) dataset combinations

Dataset	IXI-CT		IXI-UTKFace	
	Aleatoric Uncert	Epistemic Uncert	Aleatoric Uncert	Epistemic Uncert
MC-Dropout	0.90	0.90	0.77	0.70
Flipout	0.91	0.66	0.62	0.69
Ensemble	0.14	0.65	0.50	0.50
DropConnect	0.33	0.65	0.18	0.18

## E Histograms for Aleatoric and Epistemic Uncertainty

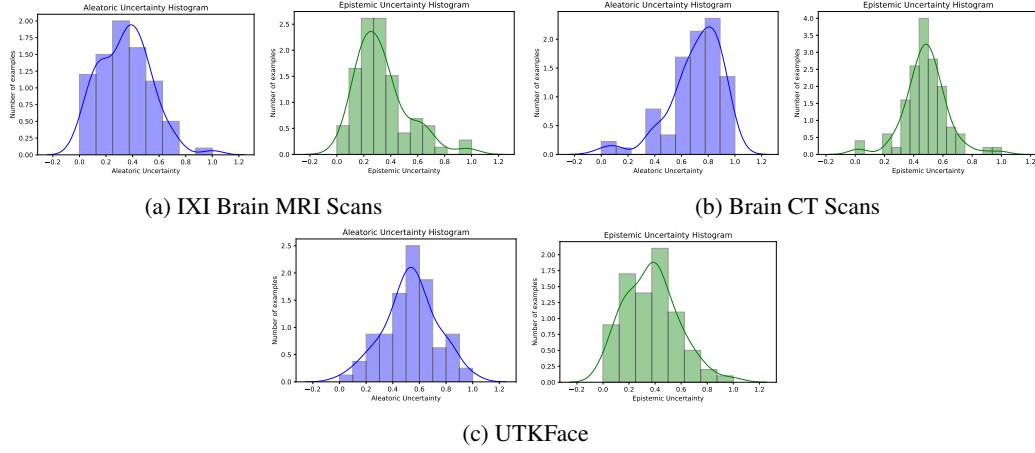


Figure 1: Aleatoric and epistemic uncertainty comparison of individual datasets with MC-Dropout

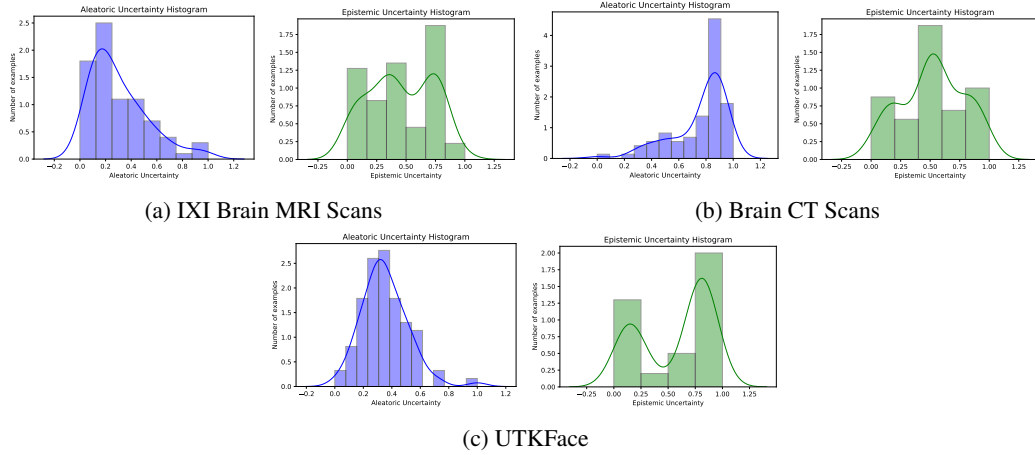


Figure 2: Aleatoric and epistemic uncertainty comparison of individual datasets with Flipout

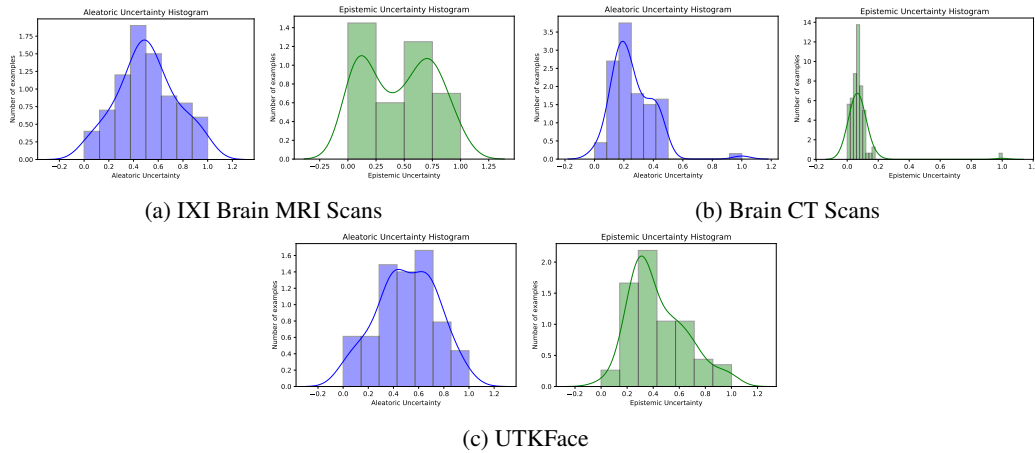


Figure 3: Aleatoric and epistemic uncertainty comparison of individual datasets with Ensembles

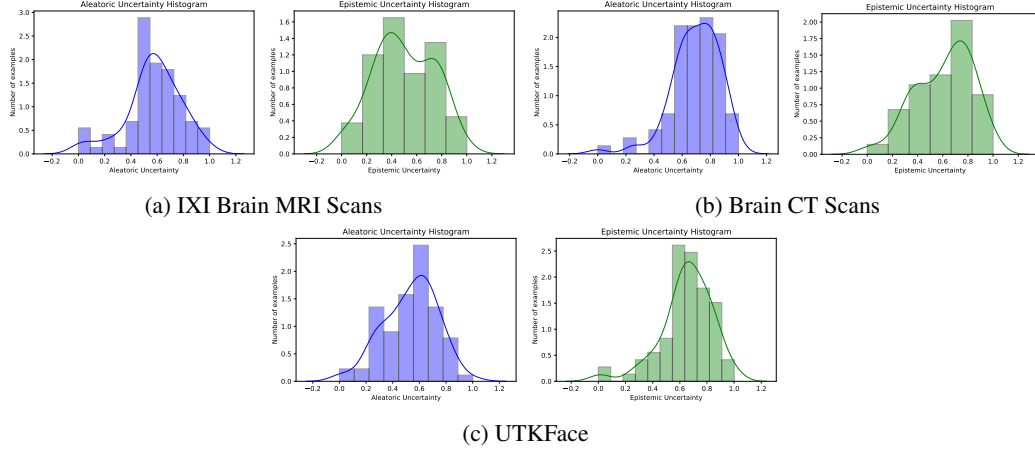


Figure 4: Aleatoric and epistemic uncertainty comparison of individual datasets with MC-DropConnect

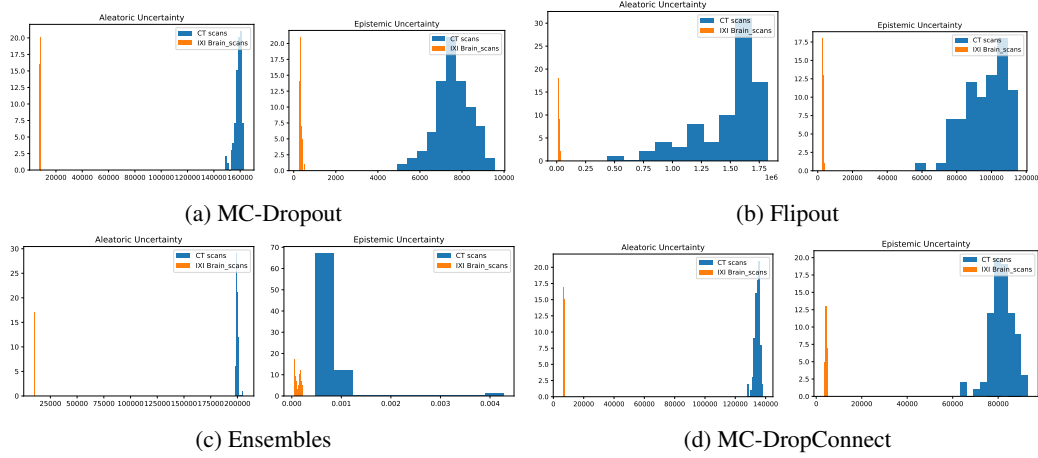


Figure 5: Aleatoric and epistemic uncertainty comparison between IXI Brain Scans and Brain CT Scans

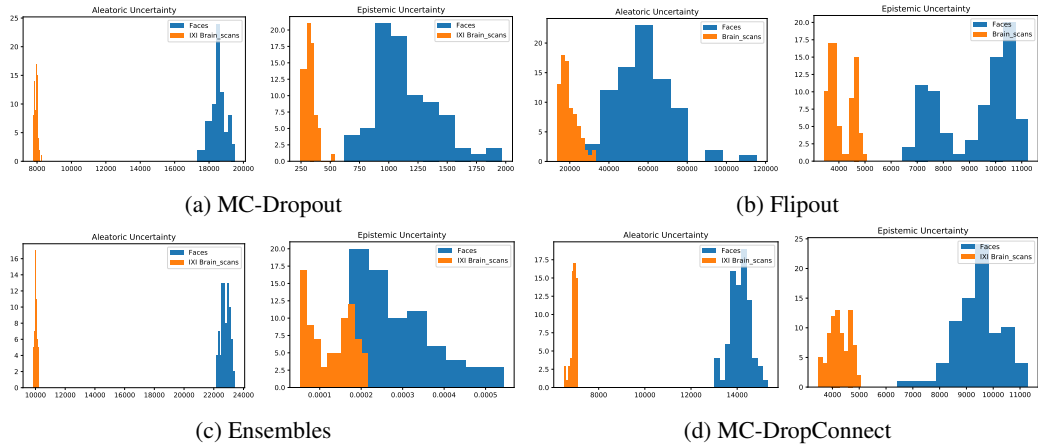


Figure 6: Aleatoric and epistemic uncertainty comparison between IXI Brain Scans and UTKFace