Propagating Uncertainty Across Cascaded Medical Imaging Tasks for Improved Deep Learning Inference

Raghav Mehta, Thomas Christinck, Tanya Nair, Aurélie Bussy, Swapna Premasiri, Manuela Costantino, M. Mallar Chakravarthy, Douglas L. Arnold, Yarin Gal, and Tal Arbel, *Member, IEEE*, *for the Alzheimer's Disease Neuroimaging Initiative*

***Abstract*— Although deep networks have been shown to perform very well on a variety of medical imaging tasks, inference in the presence of pathology presents several challenges to common models. These challenges impede the integration of deep learning models into real clinical workflows, where the customary process of cascading deterministic outputs from a sequence of image-based inference steps (e.g. registration, segmentation) generally leads to an accumulation of errors that impacts the accuracy of downstream inference tasks. In this paper, we propose that by embedding uncertainty estimates across cascaded inference tasks, performance on the downstream inference tasks should be improved. We demonstrate the effectiveness of the proposed approach in three different clinical contexts: (i) We demonstrate that by propagating T2 weighted lesion segmentation results and their associated uncertainties, subsequent T2 lesion detection performance is improved when evaluated on a proprietary large-scale, multi-site, clinical trial dataset acquired from patients with Multiple Sclerosis. (ii) We show an improvement in brain tumour segmentation performance when the uncertainty map associated with a synthesised missing MR volume is provided as an additional input to a follow-up brain tumour segmentation network, when evaluated on the publicly available BraTS-2018 dataset. (iii) We show that by propagating uncertainties from a voxel-level hippocampus segmentation task, the subsequent regression of the Alzheimer's disease clinical score is improved.**

***Index Terms*— Bayesian deep learning, uncertainty, brain tumour, multiple sclerosis, Alzheimer's, segmentation, detection, synthesis, classification.**

Manuscript received July 29, 2021; revised September 10, 2021; accepted September 12, 2021. Date of publication September 20, 2021; date of current version February 2, 2022. This work was supported in part by the Canadian Natural Science and Engineering Research Council (NSERC) Collaborative Research and Development under Grant CRDPJ 505357-16, in part by Synaptive Medical, in part by the Canadian NSERC Discovery and CREATE Grants, and in part by the International Progressive MS Alliance under Grant PA-1603-08175. The work of Aurélie Bussy was supported by the Alzheimer Society of Canada. The work of Swapna Premasiri was supported by the Fond de Recherche du Quebec–Santé. The work of M. Mallar Chakravarthy was supported in part by salary from the Fond de Recherche du Quebec–Santé and in part by the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada, and the Healthy Brains for Healthy Lives (Canada First Research Excellence Fund). The work of Tal Arbel was supported by Canada Institute for Advanced Research (CIFAR) Artificial Intelligence Chairs program, MILA. *(Corresponding author: Raghav Mehta.)*

Raghav Mehta and Tanya Nair are with the Centre for Intelligent Machines (CIM), McGill University, Montreal, QC H3A 0G4, Canada (e-mail: raghav@cim.mcgill.ca).

Thomas Christinck is with the Centre for Intelligent Machines (CIM), McGill University, Montreal, QC H3A 0G4, Canada, and also with the Integrated Program in Neuroscience, McGill University, Montreal, QC H3A 0G4, Canada.

Aurélie Bussy, Swapna Premasiri, Manuela Costantino, and M. Mallar Chakravarthy are with the Computional Brain Anatomy (CoBrA) Laboratory, Cerebral Imaging Centre, Douglas Mental Health University Institute, Montreal, QC H4H 1R3, Canada.

Douglas L. Arnold is with Montreal Neurological Institute, McGill University, Montreal, QC H3A 0G4, Canada, and also with NeuroRx Research, Montreal, QC H3A 0G4, Canada.

Yarin Gal is with OATML, Department of Computer Science, University of Oxford, Oxford OX1 2JD, U.K.

Tal Arbel is with the Centre for Intelligent Machines (CIM), McGill University, Montreal, QC H3A 0G4, Canada, and also with Montreal Institute for Learning Algorithms (MILA), Montreal, QC H2S 3H1, Canada.

This article has supplementary downloadable material available at https://doi.org/10.1109/TMI.2021.3114097, provided by the authors.

Digital Object Identifier 10.1109/TMI.2021.3114097

I. Introduction

DEEP learning methods have been shown to outperform classical computer vision methods on a variety of medical imaging inference tasks [1]–[6]. However, challenges remain in applying deep networks to medical imaging tasks in the presence of pathologies, including the limited size of publicly available datasets, the lack of reliable *ground truth* labels, the small and sometimes subtle pathological structures of interest, among others. These challenges can lead to errors in the results, impeding the integration of deep learning models into real clinical workflows. Furthermore, in a real clinical context, a typical medical image analysis pipeline [7], [8] consists of a sequence of image-based inference steps (e.g., multi-modal registration, intensity normalization, pathology segmentation). Recent trends indicate that deep learning models [9]–[11] are increasingly used at each of these steps, where their deterministic outputs are propagated from one inference step to the next. Given the additional challenges introduced by the presence of pathological structures, errors in each of these steps can accumulate and hinder performance on the downstream clinical task of interest (e.g. survival prediction). For example, networks that synthesize missing MRI sequences (e.g. FLAIR) have significantly lower fidelity in the presence of tumours [12]. Recent work has shown that these poorly synthesized images negatively affect the downstream tasks that include tumour classification, staging, and sub-type segmentation [2], [13]. In this paper, we hypothesize that the performance of the downstream tasks in a medical image analysis pipeline should improve if, in addition to mean output predictions, the uncertainty estimates are propagated across cascaded inference tasks.

Recently, Bayesian machine learning approaches have begun to address the limitations of deterministic deep learning methods by providing uncertainties associated with each prediction. Gal and Ghahramani [14] showed that by training a neural network with dropout regularization [15] and taking Monte Carlo (MC) samples of the prediction using dropout at test time, one could estimate the uncertainties associated with the outputs of deep learning models. Other popular uncertainty estimation methods based on Bayesian Neural Networks include Dropout Ensemble [16], Mean-Field Variational Inference [17], and Laplace Approximation [18]. Uncertainty estimation methods based on ensembling include Stochastic-Weight Averaging - Gaussian (SWAG) [19], Batch Ensemble [20], Snapshot Ensemble [21] and Deep Ensemble [22]. These methods allow us to generate combinations of aleatoric (data) and epistemic (model) uncertainties, given one set of provided "ground truth" labels. [23].

Several recent Bayesian machine learning approaches [24]–[26] address an additional type of uncertainty in medical image analysis caused by the fact that a unique label cannot necessarily be attained in some regions of an image (e.g., at boundaries between tumour and healthy tissue in MRI). These papers focus on the context where different annotators might systematically label things differently and where *multiple* annotations are available. They then model these inherent uncertainties (in various ways) using label variability as a proxy. Given the requirement of having access to multiple annotations, a context that is not common in practice due to the expenses incurred in attaining them, we do not focus on this type of ambiguity directly.

Recently, MC-Dropout based uncertainty estimation has been applied to a variety of medical imaging problems [27]–[29], ranging from modality synthesis [2] to lung nodule detection [27] and brain lesion detection and segmentation [5]. Many of these papers report that uncertainty can be used to estimate regions of an image where the network is prone to error, enabling the triage of highly uncertain cases for further review [2], [5], [28], [29]. Uncertainty estimation is also used in semi-supervised scenarios for improved segmentation of left atrium from chest MRI [30] and retinal layers from OCT images [31]. Recent papers [28], [32]–[34] show that estimated model confidence and model performance are correlated for a variety of medical imaging tasks. Uncertainty estimation-based active learning [35], [36] and omni-learning [37] methods try to address data scarcity problems in medical imaging.

While these approaches illustrate how estimating uncertainty in medical imaging tasks is helpful in a clinical scenario, they do not show how uncertainty can be used to inform or improve network performance on a downstream task. Recent work in medical imaging has demonstrated how uncertainty estimates can be used to improve model performance [27], [38]. In [27], the focus is limited to a single 2D application and a single uncertainty measure, the sample variance. Recent medical imaging papers [5], [38] have shown that different measures derived from MC-Dropout capture different types of uncertainties. Other work [39], [40] shows that Deep Ensemble [22] and Dropout Ensemble [16] provide better uncertainties compared to MC-Dropout. Furthermore, existing works [27], [38] only explore uncertainty propagation when both inference steps are similar to each other. It remains an open problem to explore and validate whether the propagation of uncertainty maps from an upstream task can improve performance on a related but dissimilar task.

This paper presents a general framework for propagating uncertainties across different classes of inference steps. This manuscript extends previous work [41] where a deep learning framework was developed to propagate sample variance derived from MC-Dropout at inference across cascading tasks for the contexts of Multiple Sclerosis (MS) lesion detection and brain tumour segmentation. This work presents a unified analysis of a variety of popular uncertainty generation methods (MC-Dropout, Deep Ensembles, Dropout Ensemble), uncertainty measures (e.g., entropy, sample variance, mutual information), and propagation techniques (summary statistics, random sampling) across three distinct contexts: (i) voxel-level binary MS T2 lesion segmentation to lesion detection, (ii) voxel-level MR modality synthesis to voxel-level multi-class brain tumour segmentation, and (iii) voxel-level hippocampus binary segmentation to volume-level Alzheimer's Disease clinical score regression.

Extensive experimentation shows that uncertainty propagation from a previous task to a downstream task of interest results in performance improvements in all three contexts (1-5%) and for all three model sampling methods, with Deep Ensemble and Dropout Ensemble achieving significant performance improvements over MC-Dropout (1-5%). The maximum increase in performance gain with uncertainty propagation (2-5%) is achieved when the entire set of different uncertainty measures are propagated together to the downstream task of interest, indicating that they provide helpful complementary information. However, the quantitative results only tell part of the story. The qualitative results illustrate that uncertainty propagation does indeed assist in correcting clinically relevant errors even when improvement in terms of absolute numbers are small. Finally, experiments indicate that, should the clinical context permit that the multiple samples resulting from the first inference task themselves be available to the downstream task, rather than just the uncertainty information in the form of summary statistics (e.g., entropy, variance), comparable performance improvements on the downstream task of interest result. This might be helpful for other tasks where more complex distributions prevail.

II. Methodology: Propagating Uncertainty Across Inference Tasks

In this paper, we consider a general medical imaging pipeline (see Fig. 1), where input images, $\mathbf{x}_i$, are passed through a sequence of inference tasks (Task-1, Task-2,..., Task-K) before producing the downstream output of interest (see Freesurfer [7] or ANTs [8].) The model is general, but here the context explored is one where the images may reflect some patient pathology (e.g. tumour, lesion), leading to additional challenges. The framework follows a protocol where each task is performed by a separate deep learning model sequentially. This is typical for most clinical contexts, where access to the individual training label sets for each of the tasks (e.g. reconstruction, segmentation), ($\mathbf{y}_i^1$, $\mathbf{y}_i^2, \ldots, \mathbf{y}_i^K$), is not typically available for the same input

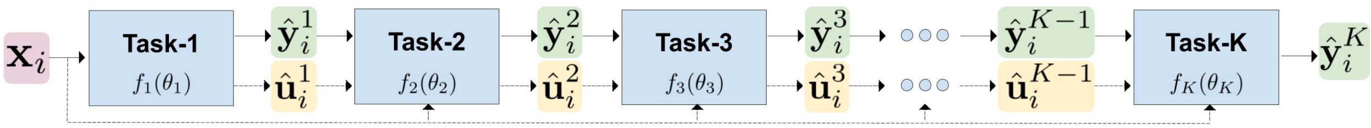

Fig. 1. An example of a medical image analysis pipeline. During inference, the input image $\mathbf{x}_i$ (and output of previous task, $\hat{\mathbf{y}}_i^k$) is passed through a cascade of inference tasks (1, 2, ..., K). The neural network for any task, Task-k, is parameterized by θ_k. The output for Task-k is defined as $\hat{\mathbf{y}}_i^k = f_k(\theta_k; \mathbf{x}_i, \hat{\mathbf{y}}_i^{k-1})$. In the proposed framework, we also estimate uncertainties ($\hat{\mathbf{u}}_i^k$) associated with output ($\hat{\mathbf{y}}_i^k$) for each task. These uncertainties are used as an additional input to the subsequent task ($\hat{\mathbf{y}}_i^k = f_k(\theta_k; \mathbf{x}_i, \hat{\mathbf{y}}_i^{k-1}, \hat{\mathbf{u}}_i^{k-1})$). Here, Task-K represents the final downstream task of interest.

images, $\mathbf{x}_i$. This hinders end-to-end training of the whole medical image analysis pipeline. Each task model is parameterized by its corresponding parameters ($\theta_1, \theta_2, \ldots, \theta_K$) such that $\hat{\mathbf{y}}_i^k = f_k(\theta_k; \mathbf{x}_i, \hat{\mathbf{y}}_i^{k-1})$.

We adopt a Bayesian deep learning [14], [16], [22] framework, whereby model predictions ($\hat{\mathbf{y}}_i^k$), as well as uncertainties ($\hat{\mathbf{u}}_i^k$) associated with these predictions can be generated for each task. These uncertainties are estimated by acquiring multiple output samples ($\hat{\mathbf{y}}_{i\ (t)}^k$) for the same input images (Sec. III-A). The model prediction becomes the mean of the samples ($\hat{\mathbf{y}}_i^k$), and the uncertainties ($\hat{\mathbf{u}}_i^k$) are derived from statistics across the samples (Sec. III-B).

In the proposed framework, depicted in Fig. 1, in addition to passing the model predictions ($\hat{\mathbf{y}}_i^k$) from each preceding task to its subsequent task, uncertainties ($\hat{\mathbf{u}}_i^k$) are also passed onto the subsequent tasks. The hypothesis is that this would lead to better performance for the downstream task of interest. We also explore a premise where instead of passing the mean prediction and its associated uncertainties from the previous task to the subsequent task, the samples ($\hat{\mathbf{y}}_{i\ (t)}^k$) themselves (should they be available) are passed individually to the next task. Direct sample propagation would help in scenarios where the output distribution might be multi-modal, for example, and not well represented by a single statistic (e.g. variance). It should be noted that this comes at the cost of increased storage requirements and substantial increases in inference time.

In order to prove the generality of the proposed framework, experiments are performed for three different clinical contexts with diverse inference steps: (i) T2 weighted MS lesion segmentation and detection, (ii) Brain tumour segmentation, and (iii) Alzheimer's (AD) clinical score prediction. Here, pipelines include two different sequential inference tasks, as depicted in Fig. 2. Note that the uncertainties produced on training cases would not properly reflect the uncertainties on unseen test cases [14], [22], [39]. In the proposed framework, the Task-1 network and the Task-2 network are trained separately to provide the Task-2 network with meaningful Task-1 uncertainties as input.

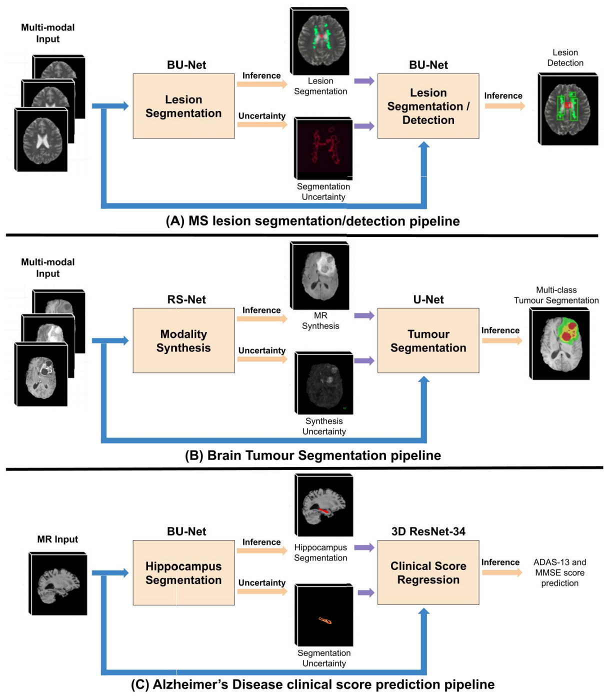

Fig. 2. Overview of the proposed general framework for propagating inference results and their associated uncertainties across sequential tasks in medical image analysis. (A) MS T2 lesion segmentation, (B) MR synthesis - brain tumour segmentation, and (C) Alzheimer's disease clinical score prediction.

A. MS T2 Lesion Segmentation

One of the hallmarks of Multiple Sclerosis (MS) is the presence of multiple hyperintense lesions visible on T2-weighted MRI (i.e. T2 lesions). The detection and segmentation of T2 lesions in MRI is therefore important to monitor disease activity and treatment efficacy. However, T2 lesions can be very small (3-10 voxels) and difficult to detect. Popular neural networks, including U-Nets, have not yet proven to be effective at the detection and segmentation of small MS lesions in MRI when deployed with commonly used settings [5]. However, uncertainties based on MC-Dropout have been shown to correlate well with network errors in the context of MS lesion segmentation [5]. In this work, we propose to first segment T2 lesions from multi-sequence MRI ($\mathbf{x}_i$) acquired from patients with MS using a Bayesian U-Net [5] (Task-1). The resulting mean T2 lesion segmentation map ($\hat{\mathbf{y}}_i^1$) and its associated voxel-level uncertainties ($\hat{\mathbf{u}}_i^1$), along with the original MRI patient sequences ($\mathbf{x}_i$), are then provided as inputs to a second T2 lesion segmentation U-Net (Task-2). The conjecture is that the second network will learn to improve the lesion segmentation/detection ($\hat{\mathbf{y}}_i^2$) performance by learning to interpret the predictions and associated uncertainties from the first network (see Fig. 2(A)). This includes learning, for example, which regions with high uncertainties should indeed be labeled as lesions and which should not, thus assisting in detecting and segmenting subtle lesions.

B. Brain Tumour Segmentation

The accuracy of detecting and segmenting brain tumours increases significantly should several MRI channels be available. Different contrasts generally assist in differentiating healthy tissues from focal pathologies (e.g., T1, T1c, T2, FLAIR) [12], [42]. However, in real clinical practice, the availability of all sequences is not guaranteed for each patient for various reasons, including cost or time constraints, and corruption from noise or patient motion. As such, accurate synthesis of one or more of the missing 3D MRI volumes based on those acquired would be beneficial to both clinical practice [43] and automatic downstream segmentation techniques [13], [44], [45]. Synthesizing high-resolution volumes in the presence of pathological structures presents significant challenges to current machine learning methods. As a result, any resulting synthesized MR volumes may not be reliable on their own. In this context, voxel-level uncertainties associated with the synthesized volume can be helpful to guide a clinician towards regions of lower confidence where further inspection is needed [2] or towards detecting an anomaly in a synthesized volume [46].

In this work, we suggest that by propagating the uncertainties associated with the synthesized missing MRI sequence provided by the synthesis network (Task-1) to a downstream tumour segmentation network (Task-2), the final results should improve. Details are shown in Fig. 2(B). The Task-1 network is a synthesis network, which takes multi-modal MR sequences acquired from a brain tumour patient as inputs. It regresses a full, synthesized image volume for the mean missing MR sequence ($\hat{\mathbf{y}}_i^1$) as well as the uncertainties ($\hat{\mathbf{u}}_i^1$) associated with the synthesis at each voxel. The synthesis network chosen here is the multi-task Regression-Segmentation Network (RS-Net) proposed in [2]. The Task-2 network is a multi-class tumour segmentation network that takes the original MRI sequences ($\mathbf{x}_i$), and the synthesized (mean) missing sequence volume ($\hat{\mathbf{y}}_i^1$) and associated uncertainties ($\hat{\mathbf{u}}_i^1$) produced from Task-1 as inputs, and produces multi-class tumour labels ($\hat{\mathbf{y}}_i^2$) at each voxel. The network is a U-Net [47] with instance normalization [48] added in order to improve performance on small batch sizes.

C. Alzheimer's Disease Clinical Score Prediction

Alzheimer's disease (AD) is the most common form of neurodegenerative disorder in elderly people [49]. Machine learning methods have been shown to perform well in providing an AD diagnosis (i.e., a classification task) [50], [51]. However, clinicians are more likely to treat symptoms based on structured clinical assessments (e.g., Alzheimer's Disease Assessment Scale – ADAS-13, Mini-Mental State Examination – MMSE) than on a specific diagnosis [52]. In this work, the objective is to develop an accurate model to estimate clinical disease severity scores, specifically the commonly used ADAS13 [53] and MMSE [54], directly from neuroimaging data (i.e., T1 MR image) [55]. A recognized biomarker for AD is the presence of reduced hippocampal volume as measured from a single time point, high-resolution T1-weighted MR image [56]. As such, automatic hippocampal segmentation has previously been shown to effectively diagnose AD [57], [58].

In this work, we hypothesize that a downstream clinical score prediction network's accuracy can be increased by propagating the estimated uncertainty maps from a preceding hippocampus segmentation network. Details are shown in Fig. 2(C). The hippocampal segmentation network (Task-1) is a BU-Net, which takes a T1 MR image ($\mathbf{x}_i$) as input and produces a mean segmentation of the hippocampus ($\hat{\mathbf{y}}_i^1$), as well as an estimate of its associated segmentation uncertainty map ($\hat{\mathbf{u}}_i^1$). The two outputs ($\hat{\mathbf{y}}_i^1$ and $\hat{\mathbf{u}}_i^1$), along with the original T1 MR image ($\mathbf{x}_i$), are then provided to a downstream deep network (3D ResNet-34 [59]) which regresses two clinical scores, ADAS-13 and MMSE ($\hat{\mathbf{y}}_i^2$).

III. Background: Uncertainty Estimation and Common Uncertainty Measures

This section provides background on sample-based uncertainty measures and various uncertainty estimation methods.

A. Uncertainty Estimation Methods

In this work, we focus on MC Dropout [14], Deep Ensembles [22], and Dropout Ensemble [16], which are the most widely used methods for uncertainty estimation. However, we expect that the method can be generalized to any method that estimates uncertainty based on multiple predicted samples.

1) MC-Dropout: Bayesian Neural Networks can estimate the uncertainty associated with model outputs. Uncertainty is estimated by placing a prior distribution over the neural network weights, though exact inference is computationally expensive and many times intractable [60], [61]. In [14], authors proposed a method known as Monte-Carlo Dropout (MC-Dropout) which uses a commonly used regularization method (Dropout [15]) at test time to approximate uncertainties associated with neural network outputs. The same input is passed through the neural network multiple times, leading to a collection of T different samples. Uncertainty is estimated using statistics computed across these samples.

2) Deep Ensemble: Deep Ensemble methods [20]–[22] have become popular due to their demonstrated reliability in predicting uncertainties. The idea behind the method is that T neural networks are trained independently, and T deterministic predictions are collected from each. These predictions then form an ensembled prediction that can be used to estimate uncertainties. While their theoretical connection to Bayesian posteriors is still a topic of active research [62], deep ensembles have been shown to work well empirically in many domains [20]–[22].

3) Dropout Ensemble: MC-Dropout captures local variability across a single network and, in turn, captures how uncertain a single network is about its prediction. Deep Ensemble captures global variability in the prediction across different networks in an ensemble and uncertainty associated with this variability [40]. Dropout Ensemble [16] combines both MC-Dropout and Deep Ensemble by training N independent networks in an ensemble and using dropout at test time for each of these networks to collect M different samples for each network. This results in a total of $T = M * N$ sample outputs across these networks.

B. Uncertainty Measures

We can use the samples generated using the above mentioned methods as a proxy for uncertainty captured by the models, or we can calculate statistics across these samples (e.g. sample variance) and consider these statistics as measures of uncertainty associated with the model output. The predicted output ($\hat{y}_i$) is the mean value across samples ($\hat{y}_i = \frac{1}{T} \sum_{t=1}^{T} \hat{y}_{i\,(t)}$).

In this section, we give details about three popular uncertainty measures: sample variance, predictive entropy, and mutual information.

1) Sample Variance: The simplest uncertainty measure, sample variance, is estimated by computing the variance across the T samples collected using either Bayesian Neural Networks (ex. [14]) or Ensembles (ex. [22]). For a regression task, such an image sequence synthesis (Sec.II-B), the variance in the output $\hat{y}_i$ for any input x_i, is defined as follows:

$$\mathrm{Var}(\hat{y}_i) = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_{i(t)}^2 - \left(\frac{1}{T}\sum_{t=1}^{T}\hat{y}_{i(t)}\right)^2. \tag{1}$$

where $\hat{y}_{i(t)}$ is a prediction for sample t.

For the segmentation tasks with C classes considered here, whether it is MS lesion segmentation (Sec.II-A), or hippocampus segmentation (Sec.II-C), the variance in the output $\hat{y}_i$ is defined as follows for any input x_i:

$$\begin{aligned}&\mathrm{Var}(\hat{y}_i)\\&=\frac{1}{C}\sum_{c=1}^{C}\left(\frac{1}{T}\sum_{t=1}^{T}p(\hat{y}_{i(t)}=c|x_i)^2-\left(\frac{1}{T}\sum_{t=1}^{T}p(\hat{y}_{i(t)}=c|x_i)\right)^2\right).\end{aligned} \tag{2}$$

Here, $p(\hat{y}_{i(t)} = c|x_i)$ denotes output softmax probability for class c for a sample t. Sample variance can be more simply interpreted as a measure of model output consistency across different samples.

2) Predictive Entropy: The predictive entropy is a measure of the informativeness of the model's predictive density function for each model output $\hat{y}_i$. It is defined as:

$$\begin{aligned}H[\hat{y}_i|x_i] &= -\sum_{c=1}^{C}p(\hat{y}_i = c|x_i)\log\Big(p(\hat{y}_i = c|x_i)\Big)\\&\approx -\sum_{c=1}^{C}\left(\frac{1}{T}\sum_{t=1}^{T}p(\hat{y}_{i(t)} = c|x_i)\right)\\&\qquad\log\left(\frac{1}{T}\sum_{t=1}^{T}p(\hat{y}_{i(t)} = c|x_i)\right).\end{aligned} \tag{3}$$

where C is the total number of class labels, and $p(\hat{y}_{i(t)} = c|x_i)$ denotes output softmax probability for class c for sample t [14], [16], [22]. High entropy implies a flatter probability distribution across classes, while low entropy implies a more peaky probability distribution. Lower entropy shows that model is more confident in its prediction of the output class. Predictive entropy measures both epistemic and aleatoric uncertainties (which will be high whenever either epistemic is high or aleatoric is high) [63], [64]. Here we only consider entropy for a segmentation task. The calculation of entropy for a regression task (e.g., modality synthesis) requires calculating a normalized histogram, a computationally intensive process.

3) Mutual Information: The mutual information (MI) captures how much information we gain about the model parameters by knowing the label for input x_i. Similar to sample variance, mutual information also captures the variability in model predictions. MI is calculated as the difference between the entropy of the average model prediction ($\hat{y}_i$) and the average of the entropies of each model prediction ($\hat{y}_{i(t)}$) [64]:

$$MI[\hat{y}_i, x_i] \approx H[\hat{y}_i|x_i] - \frac{1}{T}\sum_{t=1}^{T}H[\hat{y}_{i(t)}|x_i]. \tag{4}$$

MI measures a difference between predictive entropy and aleatoric uncertainty [63], [64]. Propagating both predictive entropy and MI together could allow the network to isolate aleatoric uncertainty component through a simple subtraction if needed [63]. Like entropy, MI is also only considered for a segmentation task as extending it to the regression task is non-trivial.

IV. Implementation Details, Datasets, and Evaluation Metrics

A. Task Specific Details[1]

1) MS T2 Lesion Segmentation: As depicted in Fig. 2(A), both the MS T2 lesion labels and their associated uncertainties produced from a Bayesian U-Net are propagated to a second T2 lesion segmentation U-Net. A large proprietary dataset of multi-modal MRI sequences acquired from a total of 1073 patients with relapsing-remitting MS (RRMS) at different stages of the disease was used for training and testing. The dataset consists of over 2700 multi-modal MRI sequences (T1, T2, Fluid Attenuated Inverse Recovery – FLAIR, and Proton Density – PD) federated from three different multi-site, multi-scanner clinical trials. The majority of the patients were scanned annually or bi-annually over 24 months. MRI sequences were acquired at 1mm × 1mm × 3mm resolution. T2 lesion labels were provided with the dataset and were produced through an external process where trained expert human annotators manually corrected a proprietary automated segmentation method. The dataset was split as follows: 40% of the available data was used for training/validating the first network, with a 90/10 training/validation split. Another 40% was used for training/validating the second network, again with a 90/10 training/validation split. The final 20% of the available data was used for testing the second network. The dataset was carefully divided this way to provide the second network with consistent and meaningful uncertainties reflective of unseen test cases.

The downstream outcome of interest is accurate detection of T2 lesions. Therefore, the performance is evaluated based on lesion-level detection metrics. A connected component analysis is performed on the voxel-based segmentation provided by the network to group lesion voxels in an 18-connected neighbourhood [5]. The detection level metrics, namely True Positive Rate (TPR) vs. False Detection Rate (FDR), are calculated at the lesion level and are used to plot receiver operating characteristic (ROC)-like curves. Given that MS lesions vary significantly in size, lesions are grouped into three sized bins for performance evaluation: small (3-10 vox), medium (11-50 vox), and large (51+ vox). Given that the detection of small lesions is particularly challenging and 40% of the lesions in the dataset are small, we mainly focus on the overall detection performance for all the lesions and show the performance on only the small lesions separately. We calculate

[1]Network architecture and training details specific to each pipeline is provided in Appendix:A.

the area under the curve (AUC) for ROC-like curves and use it as a quantitative measure of the network performance.

2) Brain Tumour Segmentation: RS-Net (Task-1 network) [2] was developed to take in 3 real MRI sequences and synthesize the missing fourth sequence. This paper focuses on the synthesis of T1 post-contrast (T1ce) and FLAIR MRIs as previous work [2], [13] has shown that their absence significantly decreases brain tumour segmentation performance compared to either T1 or T2 sequences. T1ce is the most challenging sequence to synthesize, as it is the only MR sequence that indicates enhancement within the tumour post-injection with a contrast agent, providing a signal of new disease activity. T1, T2, and FLAIR sequences are presented to RS-Net to synthesize the T1ce MRI, and T1, T1ce, and T2 MRI sequences are used as inputs to synthesize the FLAIR MRI.

This pipeline is evaluated using the 2018 MICCAI BraTS [42] dataset. The BraTS training dataset comprises 210 HGG and 75 LGG patients with T1, T1ce, T2, and FLAIR MRI sequences. Ground truth tumour labels were provided by expert human annotators and consist of 3 classes: edema, necrotic/non-enhancing core, and enhancing tumor core. 228 patients were randomly selected for training the network and another remaining 57 for network validation. A separate BraTS 2018 validation dataset was used to test the segmentation performance. This dataset contains 66 patient multi-channel MRI. The BraTS challenge provides pre-processed volumes that were skull-stripped, co-aligned, and resampled to isotropic (1mm $\times$ 1mm $\times$ 1mm) resolution. As we mentioned before, uncertainties on a training dataset would not reflect uncertainties on an unseen dataset. The RS-Net was trained in two folds, with each fold comprised of 114 volumes. This training strategy allows us to generate uncertainties on the whole training dataset in two folds, and should reflect uncertainties on an unseen dataset. The downstream segmentation U-Net was trained using all 228 volumes in a single fold.

In line with the BraTS challenge [42], the brain tumour segmentation performance is evaluated by calculating Dice scores for three different tumour sub-types: enhancing tumor, whole tumor, and tumour core. Quantitative assessment was generated by uploading the segmentation results on the challenge portal as there are no ground-truth labels available for the validation set.

3) Alzheimer's Disease Clinical Score Prediction: As depicted in Fig. 2, a BU-Net [5] is used for hippocampus segmentation with T1 MRI as the input (Task-1). The segmentation maps and their associated voxel-wise uncertainties are propagated to a volume-level clinical score regression network (Task-2), which produces values for MMSE and ADAS-13 scores. A 3D ResNet-34 [59] network was used for clinical score regression. MMSE is one of the most widely used cognitive assessments for diagnosing Alzheimer's disease and related dementias. The scores range from 0 to 30, with lower scores indicating greater cognitive impairment. The ADAS-13 is a modified version of the ADAS-cog assessment, and it has a maximum score of 85. In contrast to MMSE, higher scores on the ADAS-13 indicate greater cognitive impairment.

The EADC-ADNI/HARP dataset [65] is used for training the hippocampus segmentation network. This dataset consists of a subset of 135 volumes selected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, with expert manual 3D segmentations of the hippocampus. There are 45 AD, 46 Mild Cognitive Impairment (MCI), and 44 Cognitive Normal (CN) patients in this dataset. All volumes were in isotropic resolution, brain-extracted, and linearly registered to MNI152 space. We divide this dataset into an 80/20 training/validation split. The clinical score regression network (3D ResNet-34) is trained and tested using the ADNI [66] dataset. Specifically, we used baseline data from participants in the ADNIGO ($n = 69$), ADNI1 ($n = 442$) and ADNI2 ($n = 354$) databases. We divide this dataset into a training/validation/testing (70/10/20) split such that the ratio of AD/MCI/CN is maintained across the split. We perform 5-fold cross-validation on this dataset. Performance evaluation for both ADAS-13 and MMSE scores is based on the Pearson correlation (r), and root mean square error (RMSE) between true and predicted clinical scores.

B. Sampling for Uncertainty Estimation

The proposed framework requires producing uncertainties at the outputs of the Task-1 network along with the estimated predictions (e.g., voxel-based segmentation, regression). This is achieved by calculating various statistics (See Section III-B) across multiple samples generated using different uncertainty estimation methods (See Section III-A).

In this work, we also explore propagating the samples from the Task-1 network directly (if available) as inputs to the Task-2 networks. Details about sampling are now provided:

1) MC-Dropout: For all three clinical contexts, 20 samples are generated for the Task-1 network using dropout (dropout rate = 0.2) at test-time. We chose this as previous studies have shown that there is a marginal improvement in performance with more samples [67].

2) Deep Ensemble: 5 different Task-1 networks are trained with different weight initializations on the same training set to get an ensemble of size 5 for each clinical pipeline. This choice is based on previous studies [22], [39] which showed that only marginal improvement was attained with ensembles with sizes larger than 5. During test time, the 5 networks provide 5 different samples for the same input.

3) Dropout Ensemble: Each of the 5 trained networks developed for the Deep Ensemble model generates 20 samples using dropout at test time. This results in a total 100 samples for Dropout Ensembles.

V. Experiments and Results

Several experiments were performed for each of the clinical pipelines. The goal was to evaluate the effectiveness of propagating the uncertainties from Task-1 to Task-2 in improving the final downstream results. Evaluations and comparisons were made based on (a) different uncertainty estimation methods: MC-Dropout [14], Deep Ensemble [22], and Dropout Ensemble [16], (b) different uncertainty measures: sample variance, entropy, MI, and finally (c) propagating the uncertainties derived from the samples (e.g., sample variance) against propagating the samples themselves.

A. Effectiveness of Uncertainty Propagation

The first set of experiments were designed to evaluate the effectiveness of propagating uncertainties from the Task-1

TABLE I

COMPARING OVERALL MS T2 LESION DETECTION PERFORMANCE USING AREA UNDER CURVE (AUC) OF ROC-LIKE CURVES, ILLUSTRATING TPR (TRUE POSITIVE RATE) VS. FDR (FALSE DETECTION RATE) ACROSS (A) ALL LESIONS, AND (B) SMALL LESIONS (3-10 VOXELS) WITH SEVERAL INPUT COMBINATIONS. THE INCLUSION OF THE ASSOCIATED UNCERTAINTIES WITH OUTPUTS FROM TASK-1, IN ADDITION TO TASK-1 OUTPUTS, AS INPUTS TO THE TASK-2 NETWORK RESULTS IN IMPROVED DETECTION PERFORMANCE. **BOLD** VALUES INDICATE THE BEST PERFORMANCE FOR EACH METHOD, WHILE UNDERLINED VALUES INDICATE OVERALL BEST PERFORMANCE ACROSS DIFFERENT METHODS. THE PERFORMANCE OF THE MS T2 LESION DETECTION FOR MEDIUM AND LARGE LESION IS PROVIDED IN TABLE IV IN APPENDIX: A

	Method	Input						AUC all lesions (↑)	AUC small lesions (↑)
		MR sequences (T1, T2, FLR, T1ce, PDw)	Mean Segm.	Uncertainties			Segm. Samples		
				Var.	Entr.	MI			
1	**Baseline-1**	✓						0.8425	0.6704
2	**MC-Dropout**	✓	✓					0.8465	0.6837
3		✓	✓	✓				0.8643	**0.7197**
4		✓	✓		✓			0.8479	0.6876
5		✓	✓			✓		0.8419	0.6853
6		✓	✓	✓	✓	✓		**0.8652**	0.7170
7		✓					✓	0.8591	0.7019
8	**Dropout Ensemble**	✓	✓					0.8613	0.7116
9		✓	✓	✓				0.8739	0.7312
10		✓	✓		✓			0.8650	0.7235
11		✓	✓			✓		0.8654	0.7131
12		✓	✓	✓	✓	✓		**0.8781**	**0.7409**
13		✓					✓	0.8771	0.7341
14	**Deep Ensemble**	✓	✓					0.8603	0.7113
15		✓	✓	✓				0.8735	0.7349
16		✓	✓		✓			0.8697	0.7225
17		✓	✓			✓		0.8649	0.7159
18		✓	✓	✓	✓	✓		**0.8792**	**0.7410**
19		✓					✓	0.8767	0.7369

network to the Task-2 network. To this end, we first examine the results of the proposed framework for all three clinical pipelines (Fig. 2) based on a set of fixed experimental parameters: using MC-Dropout [14] during inference to provide 20 samples from the Task-1 network, and estimating and propagating the sample mean and variance across these samples to the Task-2 network along with the original MRI.[2] Sample variance was chosen as it is the simplest and the most commonly used uncertainty measure [6], [27], [29], [37], [46], [67]. Results were compared against (1) *Baseline-1*: only passing the MR sequences to Task-2 and (2) *Baseline-2*: passing the MRIs and the sample mean outputs from the Task-1 network to Task-2. Comparisons between Baseline-1 and Baseline-2 indicate the effectiveness of cascading inference results in general. A comparison of the proposed method with Baseline-2 should reflect the effectiveness of additionally propagating uncertainties.

Tables I, II, and III illustrate the results for the MS lesion segmentation/detection, brain tumour segmentation, and AD clinical score prediction pipelines, respectively. We perform two-sided paired sample t-test to find statistical significant difference between methods which propagates uncertainty and the baseline method which doesn't consider uncertainty propagation.[3]

[2]Fig. 9 in the Appendix:A shows the effect of varying the number MC-Dropout of sample for uncertainty estimation on a downstream task of interest for MS lesion detection.

[3]We do not report the statistical significance test result for Table I as it would require us to run multiple different runs for the large MS dataset, where each training setup takes approximately four days to run, which is practically infeasible. In this case, we have kept the folds constant across different experiments throughout the paper (and even the random seeds for the neural network initialization), which gives a fair comparison without repeated runs.

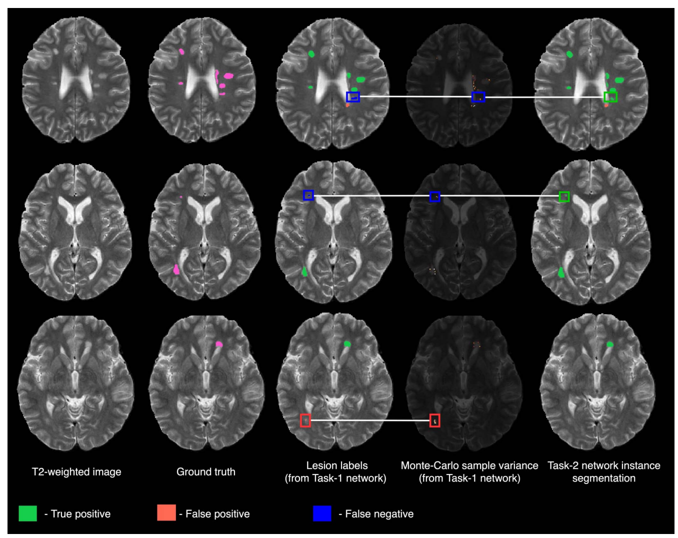

Fig. 3. Examples demonstrating the corrective effect of uncertainty propagation for MS lesion detection for three patient cases (Rows 1-3). From left to right: T2 weighted MRI input, expert T2 lesion labels (in magenta), T2 lesion labels produced by the Task-1 network, sample variance uncertainty estimates for the Task-1 network output, and the T2 lesion labels produced by the Task-2 network.

Row-1 to Row-3 in each of these tables illustrate that, for all three pipelines, the network for the downstream task of interest (Task-2) shows performance improvements of 0.5-4% when the Task-1 sample mean output is passed to the Task-2 network, relative to only passing MR sequences (*Baseline-1*). Propagating uncertainties leads to a further 2-12% performance improvement over only passing the Task-1 sample mean output to the Task-2 network (*Baseline-2*).

Although quantitative improvements are important, they do not tell the entire story. In some cases, the overall numerical improvements based on the standard performance metrics seem relatively small, however there still can be significant

TABLE II

COMPARISON OF MULTI-CLASS BRAIN TUMOUR SEGMENTATION PERFORMANCE ON THE BRATS VALIDATION DATASET. THE INCLUSION OF THE ASSOCIATED UNCERTAINTIES FROM THE SYNTHESIS NETWORK, IN ADDITION TO THE SYNTHESIS OUTPUT, AS INPUT TO THE SEGMENTATION NETWORK RESULTS IN IMPROVED PERFORMANCE. QUANTITATIVE RESULTS ARE BASED ON PERCENTAGE DICE COEFFICIENTS FOR ENHANCING TUMOR (DE), WHOLE TUMOR (DT), AND TUMOR CORE (DC). $*$ INDICATES STATISTICALLY SIGNIFICANT ($p \leq 0.05$) DIFFERENCES BETWEEN INCLUDING AND EXCLUDING UNCERTAINTY USING TWO-SIDED PAIRED SAMPLE T-TEST. **BOLD** VALUES INDICATE BEST PERFORMANCE FOR EACH METHOD, WHILE UNDERLINES INDICATE OVERALL BEST PERFORMANCE ACROSS DIFFERENT METHODS. TABLE WITH ACTUAL P-VALUES CAN BE FOUND IN SUPPLEMENTARY MATERIAL TABLE I

		Method	Input									Dice Coefficients (%)		
			Real MR sequ.				synth. MR sequ.		Var. Uncer.	synth. samples		DT (↑)	DC (↑)	DE (↑)
			T1	T2	T1ce	FLR	T1ce	FLR		T1ce	FLR			
FLR Synthesis	1	**Baseline-1**	✓	✓	✓							83.27	73.91	71.07
	2	**MC-Dropout**	✓	✓	✓			✓				84.56	76.72	72.89
	3		✓	✓	✓			✓	✓			**85.84** *	**79.25** *	**74.51** *
	4		✓	✓	✓						✓	84.83	78.43 *	73.84
	5	**Dropout Ensemble**	✓	✓	✓			✓				84.76	77.65	74.09
	6		✓	✓	✓			✓	✓			**86.45** *	**78.98**	**75.43**
	7		✓	✓	✓						✓	86.33	79.11 *	74.99
	8	**Deep Ensemble**	✓	✓	✓			✓				84.86	77.52	74.01
	9		✓	✓	✓			✓	✓			$\underline{\mathbf{86.51}}$ *	$\underline{\mathbf{79.98}}$ *	$\underline{\mathbf{75.24}}$
	10		✓	✓	✓						✓	86.03	79.19 *	74.99
T1ce Synthesis	1	**Baseline-1**	✓	✓		✓						87.17	50.25	26.89
	2	**MC-Dropout**	✓	✓		✓	✓					86.72	52.80	27.35
	3		✓	✓		✓	✓		✓			**88.20**	**57.29** *	**32.86** *
	4		✓	✓		✓				✓		87.91	56.71 *	31.95
	5	**Dropout Ensemble**	✓	✓		✓	✓					87.54	55.41	29.62
	6		✓	✓		✓	✓		✓			**88.38**	**58.99** *	**34.02** *
	7		✓	✓		✓				✓		88.01	58.09 *	32.91
	8	**Deep Ensemble**	✓	✓		✓	✓					87.45	55.68	29.62
	9		✓	✓		✓	✓		✓			$\underline{\mathbf{88.63}}$	$\underline{\mathbf{58.84}}$ *	$\underline{\mathbf{33.91}}$ *
	10		✓	✓		✓				✓		88.28	57.76 *	32.56

TABLE III

ADAS-13 AND MMSE SCORE PREDICTION PERFORMANCE COMPARISON ON THE ADNI TEST DATASET. THE INCLUSION OF THE ASSOCIATED UNCERTAINTIES FROM THE HIPPOCAMPUS SEGMENTATION NETWORK, IN ADDITION TO THE HIPPOCAMPUS SEGMENTATION OUTPUT, AS INPUT TO THE CLINICAL SCORE PREDICTION NETWORK IMPROVES BOTH ADAS-13 AND MMSE. QUANTITATIVE PREDICTION PERFORMANCE IS BASED ON ROOT MEAN SQUARED ERROR (RMSE) AND PEARSON CORRELATION COEFFICIENT (R). ($*$) INDICATES STATISTICALLY SIGNIFICANT ($p \leq 0.05$) DIFFERENCES BETWEEN INCLUDING AND EXCLUDING UNCERTAINTY USING TWO-SIDED PAIRED SAMPLE T-TEST. **BOLD** VALUES INDICATE BEST PERFORMANCE FOR EACH METHOD, WHILE UNDERLINED VALUES INDICATE OVERALL BEST PERFORMANCE ACROSS DIFFERENT METHODS. TABLE WITH ACTUAL P-VALUES CAN BE FOUND IN SUPPLEMENTARY MATERIAL TABLE I

	Method	Input						ADAS-13		MMSE	
		T1 MR sequence	Mean seg.	Uncertainties			Segm. samples	RMSE (↓)	r (↑)	RMSE (↓)	r (↑)
				Var.	Entr.	MI					
1	**Baseline-1**	✓						7.87 ± 0.92	0.47 ± 0.09	2.28 ± 0.17	0.46 ± 0.11
2	**MC-Dropout**	✓	✓					7.77 ± 0.76	0.48 ± 0.06	2.28 ± 0.12	0.47 ± 0.08
3		✓	✓	✓				7.47 ± 0.76 *	0.54 ± 0.05 *	2.23 ± 0.10 *	0.51 ± 0.05 *
4		✓	✓		✓			7.71 ± 0.77	0.47 ± 0.06	2.28 ± 0.15	0.48 ± 0.08
5		✓	✓			✓		7.72 ± 0.78	0.46 ± 0.05	2.27 ± 0.14	0.47 ± 0.07
6		✓	✓	✓	✓	✓		**7.45** ± **0.72** *	**0.54** ± $\underline{\mathbf{0.04}}$ *	**2.22** ± **0.11** *	**0.51** ± $\underline{\mathbf{0.06}}$ *
7		✓					✓	7.51 ± 0.71 *	0.51 ± 0.06 *	2.24 ± 0.15 *	0.49 ± 0.07 *
8	**Dropout Ensemble**	✓	✓					7.67 ± 0.74	0.50 ± 0.04	2.26 ± 0.12	0.48 ± 0.09
9		✓	✓	✓				7.38 ± 0.71 *	0.57 ± 0.05 *	2.17 ± 0.13 *	0.51 ± 0.05 *
10		✓	✓		✓			7.59 ± 0.72	0.50 ± 0.04	2.26 ± 0.11	0.47 ± 0.08
11		✓	✓			✓		7.68 ± 0.68	0.50 ± 0.04	2.25 ± 0.13	0.48 ± 0.07
12		✓	✓	✓	✓	✓		**7.36** ± **0.73** *	**0.57** ± **0.06** *	**2.15** ± **0.16** *	**0.53** ± **0.06** *
13		✓					✓	7.37 ± 0.61 *	0.54 ± 0.04 *	2.19 ± 0.17 *	0.52 ± 0.06 *
14	**Deep Ensemble**	✓	✓					7.69 ± 0.74	0.50 ± 0.05	2.27 ± 0.11	0.47 ± 0.08
15		✓	✓	✓				7.38 ± 0.71 *	0.57 ± 0.05 *	2.19 ± 0.14 *	0.52 ± 0.06 *
16		✓	✓		✓			7.60 ± 0.70	0.51 ± 0.05	2.27 ± 0.14	0.47 ± 0.09
17		✓	✓			✓		7.69 ± 0.67	0.49 ± 0.05	2.25 ± 0.14	0.48 ± 0.06
18		✓	✓	✓	✓	✓		$\underline{\mathbf{7.34}}$ ± $\underline{\mathbf{0.73}}$ *	$\underline{\mathbf{0.57}}$ ± $\underline{\mathbf{0.04}}$ *	$\underline{\mathbf{2.15}}$ ± $\underline{\mathbf{0.17}}$ *	$\underline{\mathbf{0.54}}$ ± $\underline{\mathbf{0.07}}$ *
19		✓					✓	7.38 ± 0.63 *	0.55 ± 0.05 *	2.18 ± 0.18 *	0.52 ± 0.06 *

clinically relevant improvements. For example, Fig. 3 depicts qualitative results for three MS patient cases (top to bottom), where the propagation of uncertainties enabled the correction of both false positive (bottom case) and false negative (top two cases) lesions. The system learned how to interpret the uncertainties in the (incorrect) inferences made in those areas, and corrected the errors.

Fig. 4 shows example cases for three patients (top to bottom), where the downstream brain tumour segmentation network makes use of synthesized MRI sequences (here T1ce and FLAIR). The first example (top row) shows that propagating the synthesized T1ce image to the downstream tumour segmentation network results in confusion between enhancing tumour and core tumour, as the enhancing portion is not well synthesized in the generated T1ce. This result is not unsurprising as T1ce is the post-contrast injection T1 MRI, and accurate synthesis of enhanced tumour without injection remains an open problem. Importantly, the system produces an uncertainty map that indicates that the synthesis uncertainty is higher in this region, and conveys the uncertainty

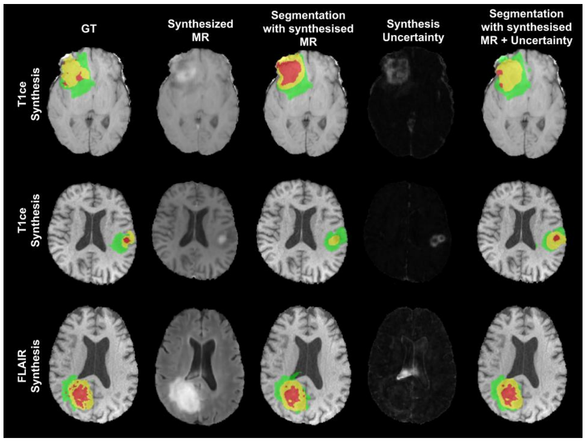

Fig. 4. Examples of three patient cases (top to bottom) demonstrating the 3D U-Net performance on the multi-class brain tumour segmentation task [42] based on synthesized MRI sequences. From Left to Right: Expert manual segmentation, synthesized MR sequence, segmentation using real MRI (3 sequences) + synthesized MRI, synthesis uncertainty, segmentation using real MRI (3 sequences) + synthesized MRI + synthesis uncertainty. First two rows: T1ce synthesis. Last row: FLAIR synthesis. Labels: edema (green), non-enhancing or necrotic tumour core (red), enhancing tumour (yellow).

information to the segmentation network. This enables the segmentation network to learn to correct these errors and leads to an improvement in the results. This can also be seen in the example in second row, where the uncertainty allows the network to fix errors and correctly identify enhancing and non-enhancing core. The third example shows the results of FLAIR synthesis, where an erroneous bright spot appears within the ventricle. This leads to the segmentation network erroneously predicting edema within the ventricle (which is clinically impossible) when the uncertainty is not propagated. However, the uncertainty maps indicate that the network is not confident in its synthesis prediction in this region. As such, cascading the uncertainty maps permits the network to learn to correct its error.

The results for all 3 clinical pipelines demonstrate that in multi-step medical image processing pipelines, that would otherwise accumulate errors can benefit from including the network uncertainty for each task as input to subsequent tasks.

B. MC-Dropout vs. Deep Ensemble vs. Dropout Ensemble

The next set of experiments compare the performance of uncertainty propagation using different methods for estimating sample variance uncertainties: MC-Dropout [14], Deep Ensemble [22], and Dropout Ensemble [16]. Tables I, II, and III, Row-2 and Row-3, Row-8 and Row-9, and Row-14 and Row-15 report results for MC-Dropout, Dropout Ensemble, and Deep Ensemble, respectively. These results indicate that ensemble methods, Deep Ensemble and Dropout Ensemble, achieve 1-5% higher performance over MC-Dropout when only mean predictions are propagated across tasks. The performance gains improve by a further 1-4% when the sample variance uncertainties are additionally propagated to the downstream task of interest. A marginal performance gain of Dropout Ensemble over Deep Ensemble can be seen, both with and without uncertainty propagation.

C. Effect of Different Uncertainty Measures

Experiments were devised in order to compare the effects of propagating each of the different uncertainty measures: sample variance, entropy and MI (Sec: III-B), as well as the effectiveness of cascading all three measures at once for all three uncertainty estimation strategies. Experiments were performed for the clinical pipelines of MS lesion segmentation/detection and AD clinical score prediction, but not for the brain tumour segmentation pipeline as estimating entropy or MI in the context of image regression (synthesis) in this context is an open research problem. (Sec.III-B.1) [64]. Tables I and III show that the sample variance gives better performance gains over entropy and MI for both the MS T2 lesion detection task and AD clinical score prediction task. However, passing all three uncertainty measures simultaneously shows the best improvement in the performance of downstream tasks (Row-7, Row-13, and Row-19), indicating that each provides different yet relevant summary statistics [5].

D. Statistics vs Samples

Finally, the effectiveness of passing summary statistics calculated across samples are examined against propagating the samples themselves for all three uncertainty estimation strategies. Multiple samples are generated (Sec:IV-B) from the Task-1 network for these uncertainty estimation strategies. During Task-2 network training, one random sample from the available Task-1 output samples is provided as input. During inference, all Task-1 samples are independently passed to the Task-2 network. The output samples from the Task-2 network are then used to estimate the sample mean, which serves as the final Task-2 output. Table I, Table II, and III indicate that passing samples instead of statistics across samples results in similar performance in the contexts explored in this paper.

VI. Conclusion

This work proposes a general deep learning framework for propagating uncertainties across a sequence of inference tasks within medical image analysis pipelines. It demonstrates that cascading uncertainties (e.g., based on MC dropout, Deep Ensemble) along with the outputs from the previous inference module can lead to improvements in performance of the downstream task. The framework was applied to three different contexts. First, we showed that by propagating voxel-based lesion segmentation uncertainties to a second segmentation network, lesion-level detection performance could be improved by reducing both FPs and FNs. Experiments were performed on a large-scale, multi-site MS patient brain MRI dataset acquired during different clinical trials. Next, using the publicly available BraTS dataset, we demonstrated that by propagating regression uncertainties from an MRI synthesis network, the performance of a downstream multi-class tumour segmentation task can be improved. In the last context, we demonstrated that uncertainty propagation from a voxel-level hippocampus segmentation network to a scan-level clinical score regression task in the context of images acquired from AD patients leads to improved predictions. These results

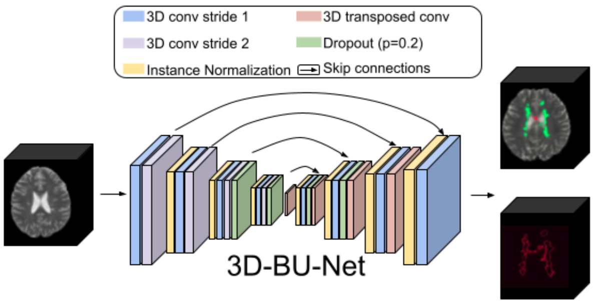

Fig. 5. Network architecture diagram for the BU-Net [5]. BU-Net provides the segmentation outputs and permits the estimation of the uncertainties associated with the outputs. BU-Net was used for both Task-1 and Task-2 in the MS lesion segmentation/detection pipeline depicted here and as a Task-1 network for hippocampus segmentation in the Alzheimer's Disease clinical score prediction pipeline.

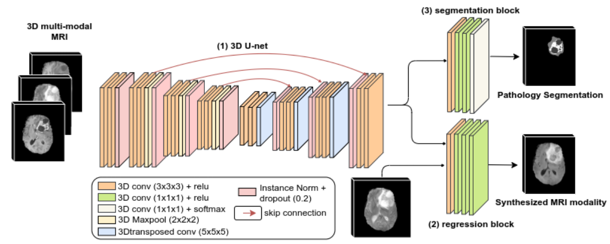

Fig. 6. Network architecture diagram of RS-Net [2]. We use RS-Net for the synthesis of the missing MRI sequence synthesis (Task-1) in the brain tumour segmentation pipeline. Note that T1, T2, and T1ce are used as inputs to the network when synthesizing FLAIR, while T1, T2, and FLAIR are used as inputs when synthesizing T1ce.

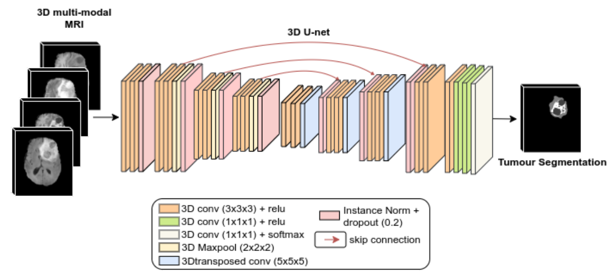

Fig. 7. Network architecture diagram of the modified 3D-U-Net [47], used for the multi-class brain tumour segmentation (Task-2) in the brain tumour segmentation pipeline. The inputs to this network vary depending on the experiment. For example, when assessing the effectiveness of uncertainty propagation, we also pass the uncertainties associated with the synthesized MR sequence as input to the network.

are encouraging and suggest that uncertainties can be propagated to a downstream task of interest to improve performance in cascaded medical image processing pipelines where the upstream task is related to the downstream task of interest.[4] The expectation is that the results are generalizable to other clinical pipelines. Results can be further improved with better calibrated uncertainties [39], [68]. Improvements on the performance of downstream tasks based propagation of sample free uncertainty estimations [69] or learned sample-based models [70] should also provide benefits, with an added decrease in inference time. Our experiments also showed that by propagating Task-1 samples to the Task-2 network as a proxy to the uncertainty associated with the Task-1 output, we could achieve similar performance. This is important as samples could better represent the Task-1 output distribution when it is multi-modal, compared to a single statistic like sample variance. It should be noted that the performance improvements resulting from uncertainty propagation are dependent on the number of samples taken to estimate the uncertainties (as we show in Appendix:A - Figure 9), as well as sample generation method. As a result, it would be important to tune these hyper-parameters for optimal performance in the particular application of interest.

An end-to-end system for uncertainty propagation in a medical image analysis pipeline requires access to ground-truth labels at all inference stages for the same training data. This data is generally not available in real clinical contexts (see, for example, the ADNI clinical score prediction pipeline). In general, the vast majority of medical image analysis tasks are developed independently and without consideration of downstream tasks of interest. Should it be possible, an end-to-end system where relevant uncertainty measures for a task are learned depending on the downstream task of interest may be an exciting and essential research direction to explore.

Finally, future work will explore the impact of uncertainty propagation on the uncertainties of the downstream task's outputs. One could expect better uncertainty quantification in a downstream task of interest with uncertainty propagation. It would also be interesting to propagate labeling uncertainties associated with different tasks [24], [25], if multiple annotations for each patient case are available.

[4]Propagating uncertainties from a skull stripping task to a hippocampus segmentation task might not lead to performance improvement, as the two tasks are not directly related.

Appendix

A. Implementation Details

In section, we provide details about the network architecture, implementation details and the training process for all three pipelines explored in the paper: Multiple Sclerosis lesion segmentation/detection (Sec:II-A), brain tumour segmentation (Sec:II-B), and Alzheimer's disease clinical score prediction (Sec:II-C). Note that all our experiments were implemented using PyTorch, and ran on a machine equipped with an NVIDIA Titan Xp GPU with 12 GBs of memory.

1) MS T2 Lesion Segmentation Detection Pipeline: The pipeline (Sec:II-A) consists of a cascade of two binary lesion segmentation tasks. We chose an off-the-shelf BU-Net [5] architecture[5] for both Task-1 and Task-2 networks, which can be seen in Figure 5. The only differences between the two networks were their inputs. For the Task-1 network, the inputs consisted of all the MR sequences. The Task-2 network takes as input the MR sequences, the Task-1 network output, and the uncertainties associated with the Task-1 network output (in the case of the proposed framework). These additional inputs marginally increase the total number of parameters for the Task-2 network. For exact architecture details, readers can refer to the BU-Net [5] paper.

Both the Task-1 and Task-2 networks were trained to minimize a weighted binary cross-entropy loss function for

[5]We reimplemented the model architecture in PyTorch following the code (link) provided by the authors.

TABLE IV

COMPARING OVERALL MS T2 LESION DETECTION PERFORMANCE USING AREA UNDER CURVE (AUC) OF ROC-LIKE CURVES, ILLUSTRATING TPR (TRUE POSITIVE RATE) VS. FDR (FALSE DETECTION RATE) ACROSS (A) LARGE LESIONS (51+ VOXELS), AND (B) MEDIUM LESIONS (10-50 VOXELS) WITH SEVERAL INPUT COMBINATIONS. THE INCLUSION OF THE ASSOCIATED UNCERTAINTIES WITH OUTPUTS FROM TASK-1, IN ADDITION TO TASK-1 OUTPUTS, AS INPUTS TO THE TASK-2 NETWORK RESULTS IN IMPROVED DETECTION PERFORMANCE. **BOLD** VALUES INDICATE THE BEST PERFORMANCE FOR EACH METHOD, WHILE UNDERLINED VALUES INDICATE OVERALL BEST PERFORMANCE ACROSS DIFFERENT METHODS

	Method	Input						AUC Med. lesions (↑)	AUC large lesions (↑)
		MR sequences (T1, T2, FLR, T1ce, PDw)	Mean Segm.	Uncertainties			Segm. Samples		
				Var.	Entr.	MI			
1	**Baseline-1**	✓						0.9768	0.9977
2	**MC-Dropout**	✓	✓					0.9775	0.9979
3		✓	✓	✓				0.9834	**0.9992**
4		✓	✓		✓			0.9768	0.9982
5		✓	✓			✓		0.9782	0.9979
6		✓	✓	✓	✓	✓		**0.9852**	0.9989
7		✓					✓	0.9831	0.9981
8	**Dropout Ensemble**	✓	✓					0.9801	0.9988
9		✓	✓	✓				0.9858	0.9992
10		✓	✓		✓			0.9807	0.9992
11		✓	✓			✓		0.9851	0.9992
12		✓	✓	✓	✓	✓		**0.9861**	**0.9993**
13		✓					✓	0.9844	0.9989
14	**Deep Ensemble**	✓	✓					0.9806	0.9988
15		✓	✓	✓				0.9855	0.9992
16		✓	✓		✓			0.9812	0.9991
17		✓	✓			✓		0.9856	0.9993
18		✓	✓	✓	✓	✓		**0.9867**	**0.9993**
19		✓					✓	0.9854	0.9990

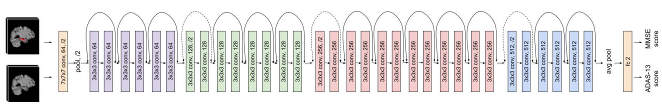

Fig. 8. Network architecture diagram of modified 3D-ResNet-34 [47] for the Alzheimer's Disease clinical regression pipeline for predicting both ADAS-13 and MMSE scores. In our framework, input to this network varies depending on the experiment. For example, when assessing the effectiveness of uncertainty propagation, uncertainties associated with the hippocampus segmentation is also provided as input to the network.

the lesion segmentation task. Here, class weights were taken as an inverse of the frequency of both lesion/non-lesion voxels within the brain mask. After every epoch, class weights were decayed with a factor of 0.95, which results in equally weighted binary cross-entropy after around 50 epochs. The networks were trained using an Adam optimizer with an initial learning rate of 0.0002 and a weight decay of 0.00001 for a total of 250 epochs. The learning rate was decayed with a factor of 0.995 after each epoch.

2) Brain Tumour Segmentation Pipeline: The pipeline (Sec:II-B) consists of two different sequential inference tasks. The first network (Task-1) is designed for a 3-to-1 synthesis of a missing MRI sequence in the presence of a brain tumour. RS-Net[6] [2] was chosen for this task and can be seen in Fig. 6. RS-Net is a multi-task network designed to jointly perform the synthesis of the missing image while performing the segmentation of the tumour, with the premise that this would improve the synthesis in the tumor area. RS-Net was trained for a total of 400 epochs using an Adam optimizer with a learning rate of 0.0002 and a weight decay of 0.00001. The learning rate was decayed with a factor of 0.995 after each epoch. The network was trained to minimize a combined weighted mean squared error and weighted cross-entropy loss [2].

[6]Readers are requested to refer RS-Net paper for the exact network architecture details.

A modified 3D U-Net [47], depicted in Fig. 7, was developed for multi-class brain tumour segmentation (Task-2 Network). Similar to the original 3D U-Net, the network consists of encoder and decoder paths that embed convolution, pooling, and deconvolution operations. High-resolution features from the encoder path were combined with the up-sampled output of the decoder to preserve high-resolution features. Each convolution was followed by rectified linear unit activation (ReLU). Instead of using the batch-normalization layer used in the original U-Net, we used instance normalization [48]. Instance normalization typically improves performance for small batch sizes. The network was trained using Adam optimizer with a learning rate of 0.0002 and weight decay of 0.00001 for a total of 240 epochs to minimize weighted cross-entropy loss. Here, the weights are defined such that the weight increases whenever there are fewer voxels in a particular class. After every epoch, class weights were decayed with a factor of 0.95, which results in equally weighted binary cross-entropy after around 50 epochs. Inputs to the network varies depending on the experiment. For example, in the proposed framework 3D U-Net takes as input the real MR sequences, the RS-Net synthesized MR sequence, and the uncertainties associated with the synthesized MR sequence. These additional inputs result in a marginal increase in the total number of parameters.

3) Alzheimer's Disease Clinical Score Prediction Pipeline: The pipeline described in Sec:II-C consists of two cascaded inference tasks. The BU-Net [5] was used for the binary hippocampus segmentation task (Fig. 5). The T1 weighted MRI was provided as an input to the BU-Net. The network was trained to reduce the weighted binary cross-entropy loss using an Adam optimizer with a learning rate of 0.0002 and a weight decay of 0.00001 for a total of 250 epochs. Here, class weights were taken as an inverse of the frequency of both hippocampus/background voxels within the brain mask. The learning rate was decayed with a factor of 0.995 after each

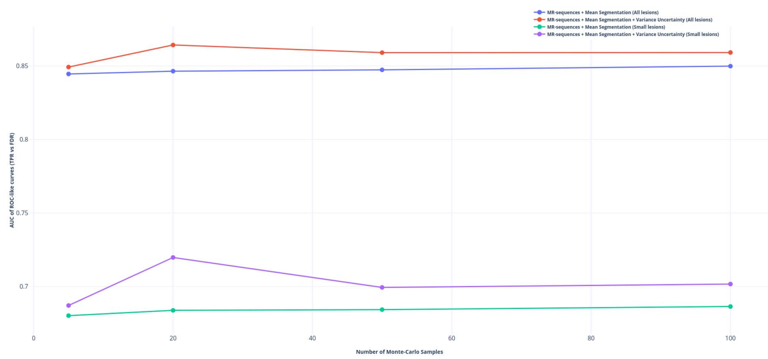

Fig. 9. Comparing overall MS T2 lesion detection performance using Area Under Curve (AUC) of ROC-like curves, illustrating TPR (true positive rate) vs. FDR (false detection rate) across all lesions, and small lesions (3-10 voxels). Here we evaluate the impact of number of samples used to estimate uncertainty (variance) measure for MC-Dropout uncertainty estimation method. From the plot we can see that for all lesion detection and small lesion detection, highest performance is achieved when 20 samples are used to estimate uncertainty. With increase in number of samples, performance saturates.

epoch. After every epoch, the class weights were decayed with a factor of 0.95, which results in equally weighted binary cross-entropy after around 50 epochs.

A 3D ResNet34 [59] architecture was designed for the task of clinical score prediction (Task-2).[7] The network (Fig. 8) was modified to be a multi-task network, such that it predicts both ADAS-13 and MMSE scores simultaneously. The network was trained to reduce the combined mean squared error losses for both ADAS-13 and MMSE. An Adam optimizer with a learning rate of 0.0002 and a weight decay of 0.00001 was used to train the network for a total of 200 epochs. The learning rate was decayed with a factor of 0.995 after each epoch.

Acknowledgment

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how to apply/ ADNI Acknowledgement List.pdf.

References

[1] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris, "Multimodal MR synthesis via modality-invariant latent representation," *IEEE Trans. Med. Imag.*, vol. 37, no. 3, pp. 803–814, Mar. 2017.

[2] R. Mehta and T. Arbel, "RS-Net: Regression-segmentation 3D CNN for synthesis of full resolution missing brain MRI in the presence of tumours," in *Simulation and Synthesis in Medical Imaging* (Lecture Notes in Computer Science), vol. 11037, A. Gooya, O. Goksel, I. Oguz, and N. Burgos, Eds. Springer, Sep. 2018, pp. 119–129. [Online]. Available: https://dblp.org/rec/conf/miccai/MehtaA18.bib, doi: 10.1007/978-3-030-00536-8_13.

[3] A. V. Dalca, G. Balakrishnan, J. V. Guttag, and M. R. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *Medical Image Computing and Computer Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 11070, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Springer, Sep. 2018, pp. 729–738. [Online]. Available: https://dblp.org/rec/conf/miccai/DalcaBGS18.bib, doi: 10.1007/978-3-030-00928-1_82.

[4] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No new-net," in *Proc. Int. MICCAI Brainlesion Workshop*. Springer, 2018, pp. 234–244.

[7] https://github.com/kenshohara/3D-ResNets-PyTorch/blob/master/models/resnet.py

[5] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Med. Image Anal.*, vol. 59, Jan. 2020, Art. no. 101557.

[6] A. Tousignant, P. Lemaître, D. Precup, D. L. Arnold, and T. Arbel, "Prediction of disease progression in multiple sclerosis patients using deep learning analysis of MRI data," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, 2019, pp. 483–492.

[7] A. M. Dale, B. Fischl, and M. I. Sereno, "Cortical surface-based analysis: I. Segmentation and surface reconstruction," *NeuroImage*, vol. 9, no. 2, pp. 179–194, Feb. 1999.

[8] B. B. Avants, N. Tustison, and G. Song, "Advanced normalization tools (ANTs)," *Insight J.*, vol. 2, no. 365, pp. 1–35, Jun. 2009.

[9] J. Fan, X. Cao, Q. Wang, P.-T. Yap, and D. Shen, "Adversarial learning for mono-or multi-modal registration," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101545.

[10] J. Kleesiek *et al.*, "Deep MRI brain extraction: A 3D convolutional neural network for skull stripping," *NeuroImage*, vol. 129, pp. 460–469, Apr. 2016.

[11] A. Simkó, T. Löfstedt, A. Garpebring, T. Nyholm, and J. Jonsson, "A generalized network for MRI intensity normalization," 2019, *arXiv:1909.05484*. [Online]. Available: http://arxiv.org/abs/1909.05484

[12] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "HeMIS: Hetero-modal image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 9901, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. B. Ünal, and W. Wells, Eds. Springer, Oct. 2016, pp. 469–477. [Online]. Available: https://dblp.org/rec/conf/miccai/HavaeiGCB16.bib, doi: 10.1007/978-3-319-46723-8_54.

[13] G. van Tulder and M. de Bruijne, "Why does synthesized data improve multi-sequence classification?" in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 9349, N. Navab, J. Hornegger, W. M. Wells, III, and A. F. Frangi, Eds. Springer, Oct. 2015, pp. 531–538. [Online]. Available: https://dblp.org/rec/conf/miccai/TulderB15.bib, doi: 10.1007/978-3-319-24553-9_65.

[14] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[16] L. Smith and Y. Gal, "Understanding measures of uncertainty for adversarial example detection," 2018, *arXiv:1803.08533*. [Online]. Available: http://arxiv.org/abs/1803.08533

[17] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1613–1622.

[18] H. Ritter, A. Botev, and D. Barber, "A scalable Laplace approximation for neural networks," in *Proc. 6th Int. Conf. Learn. Represent. Conf. Track (ICLR)*, vol. 6, 2018, pp. 1–15.

[19] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, and A. G. Wilson, "A simple baseline for Bayesian uncertainty in deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13153–13164.

[20] Y. Wen, D. Tran, and B. Jimmy, "BatchEnsemble: An alternative approach to efficient ensemble and lifelong learning," in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–20.

[21] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot ensembles: Train 1, get M for free," 2017, *arXiv:1704.00109*. [Online]. Available: http://arxiv.org/abs/1704.00109

[22] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. neural Inf. Process. Syst.*, 2017, pp. 6402–6413.

[23] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" 2017, *arXiv:1703.04977*. [Online]. Available: http://arxiv.org/abs/1703.04977

[24] S. Kohl *et al.*, "A probabilistic U-Net for segmentation of ambiguous images," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 6965–6975.

[25] C. F. Baumgartner *et al.*, "PHiSeg: Capturing uncertainty in medical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 11765, D. Shen *et al.*, Eds. Springer, Oct. 2019, pp. 119–127. [Online]. Available: https://dblp.org/rec/conf/miccai/BaumgartnerTCHM19.bib, doi: 10.1007/978-3-030-32245-8_14.

[26] M. Monteiro *et al.*, "Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty," 2020, *arXiv:2006.06015*. [Online]. Available: http://arxiv.org/abs/2006.06015

[27] O. Ozdemir, B. Woodward, and A. A. Berlin, "Propagating uncertainty in multi-stage Bayesian convolutional neural networks with application to pulmonary nodule detection," 2017, *arXiv:1712.00497*. [Online]. Available: http://arxiv.org/abs/1712.00497

[28] A. G. Roy, S. Conjeti, N. Navab, and C. Wachinger, "Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control," *NeuroImage*, vol. 195, pp. 11–22, Jul. 2019.

[29] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Sci. Rep.*, vol. 7, no. 1, pp. 1–14, Dec. 2017.

[30] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 11765, D. Shen *et al.*, Eds. Springer, Oct. 2019, pp. 605–613. [Online]. Available: https://dblp.org/rec/conf/miccai/YuWLFH19.bib, doi: 10.1007/978-3-030-32245-8_67.

[31] S. Sedai *et al.*, "Uncertainty guided semi-supervised segmentation of retinal layers in OCT images," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 11764, D. Shen *et al.*, Eds. Springer, Oct. 2019, pp. 282–290. [Online]. Available: https://dblp.org/rec/conf/miccai/SedaiARJ0SWG19.bib, doi: 10.1007/978-3-030-32239-7_32.

[32] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 3868–3878, Dec. 2020.

[33] A. Jungo and M. Reyes, "Assessing reliability and challenges of uncertainty estimations for medical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 11765, D. Shen *et al.*, Eds. Springer, Oct. 2019, pp. 48–56. [Online]. Available: https://dblp.org/rec/conf/miccai/Jungo019.bib, doi: 10.1007/978-3-030-32245-8_6.

[34] R. Mehta, A. Filos, Y. Gal, and T. Arbel, "Uncertainty evaluation metric for brain tumour segmentation," 2020, *arXiv:2005.14262*. [Online]. Available: http://arxiv.org/abs/2005.14262

[35] D. Sharma, Z. Shanis, C. K. Reddy, S. Gerber, and A. Enquobahrie, "Active learning technique for multimodal brain tumor segmentation using limited labeled images," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 11795, Q. Wang *et al.*, Eds. Springer, Oct. 2019, pp. 148–156. [Online]. Available: https://dblp.org/rec/conf/miccai/SharmaSRGE19.bib, doi: 10.1007/978-3-030-33391-1_17.

[36] D. Zotova, A. Lisowska, O. Anderson, V. Dilys, and A. O'Neil, "Comparison of active learning strategies applied to lung nodule segmentation in CT scans," in *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention* (Lecture Notes in Computer Science), vol. 11851, L. Zhou *et al.*, Eds. Springer, Oct. 2019, pp. 3–12. [Online]. Available: https://dblp.org/rec/conf/miccai/ZotovaLADO19.bib, doi: 10.1007/978-3-030-33642-4_1.

[37] L. Venturini, A. T. Papageorghiou, J. A. Noble, and A. I. L. Namburete, "Uncertainty estimates as data selection criteria to boost omni-supervised learning," in *Medical Image Computing and Computer Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 12261, A. L. Martel *et al.*, Eds. Springer, Oct. 2020, pp. 689–698. [Online]. Available: https://dblp.org/rec/conf/miccai/VenturiniPNN20.bib, doi: 10.1007/978-3-030-59710-8_67.

[38] L. Herzog, E. Murina, O. Dürr, S. Wegener, and B. Sick, "Integrating uncertainty in deep neural networks for MRI based stroke analysis," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101790.

[39] A. Ashukha, A. Lyzhov, D. Molchanov, and D. Vetrov, "Pitfalls of in-domain uncertainty estimation and ensembling in deep learning," 2020, *arXiv:2002.06470*. [Online]. Available: http://arxiv.org/abs/2002.06470

[40] S. Fort, H. Hu, and B. Lakshminarayanan, "Deep ensembles: A loss landscape perspective," 2019, *arXiv:1912.02757*. [Online]. Available: http://arxiv.org/abs/1912.02757

[41] R. Mehta, T. Christinck, T. Nair, P. Lemaître, D. L. Arnold, and T. Arbel, "Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures* (Lecture Notes in Computer Science), vol. 11840, H. Greenspan *et al.*, Eds. Springer, Oct. 2019, pp. 23–32. [Online]. Available: https://dblp.org/rec/conf/miccai/MehtaCNLAA19.bib, doi: 10.1007/978-3-030-32689-0_3.

[42] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018, *arXiv:1811.02629*. [Online]. Available: http://arxiv.org/abs/1811.02629

[43] J. L. Prince, A. Carass, C. Zhao, B. E. Dewey, S. Roy, and D. L. Pham, "Image synthesis and superresolution in medical imaging," in *Handbook of Medical Image Computing and Computer Assisted Intervention*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 1–24.

[44] A. Jog, A. Carass, S. Roy, D. L. Pham, and J. L. Prince, "Random forest regression for magnetic resonance image synthesis," *Med. Image Anal.*, vol. 35, pp. 475–488, Jan. 2017.

[45] J. E. Iglesias, E. Konukoglu, D. Zikic, B. Glocker, K. Van Leemput, and B. Fischl, "Is synthesizing MRI contrast useful for inter-modality analysis?" in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 8149, K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, Eds. Springer, Sep. 2013, pp. 631–638. [Online]. Available: https://dblp.org/rec/conf/miccai/IglesiasKZGLF13.bib, doi: 10.1007/978-3-642-40811-3_79.

[46] J. C. Reinhold *et al.*, "Validating uncertainty in medical image translation," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 95–98.

[47] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 9901, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. B. Ünal, and W. Wells, Eds. Springer, Oct. 2016, pp. 424–432. [Online]. Available: https://dblp.org/rec/conf/miccai/CicekALBR16.bib, doi: 10.1007/978-3-319-46723-8_49.

[48] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022

[49] M. Goedert and M. G. Spillantini, "A century of Alzheimer's disease," *Science*, vol. 314, no. 5800, pp. 777–781, 2006.

[50] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.

[51] K. Gopinath, C. Desrosiers, and H. Lombaert, "Learnable pooling in graph convolution networks for brain surface analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 2, 2020, doi: 10.1109/TPAMI.2020.3028391.

[52] C. M. Stonnington, C. Chu, S. Klöppel, C. R. Jack, Jr., J. Ashburner, and R. S. Frackowiak, "Predicting clinical scores from magnetic resonance scans in Alzheimer's disease," *NeuroImage*, vol. 51, no. 4, pp. 1405–1413, 2010.

[53] L.-O. Wahlund *et al.*, "A new rating scale for age-related white matter changes applicable to MRI and CT," *Stroke*, vol. 32, no. 6, pp. 1318–1322, 2001.

[54] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "'Mini-mental state': A practical method for grading the cognitive state of patients for the clinician," *J. Psychiatric Res.*, vol. 12, no. 3, pp. 189–198, 1975.

[55] N. Bhagwat *et al.*, "An artificial neural network model for clinical score prediction in Alzheimer disease using structural neuroimaging measures," *J. Psychiatry Neurosci.*, vol. 44, no. 4, p. 246, 2019.

[56] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in Alzheimer disease," *Nature Rev. Neurol.*, vol. 6, no. 2, pp. 67–77, 2010.
[57] M. Chupin *et al.*, "Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI," *Hippocampus*, vol. 19, no. 6, pp. 579–587, 2009.
[58] M. Liu *et al.*, "A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease," *NeuroImage*, vol. 208, Mar. 2020, Art. no. 116459.
[59] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.
[60] D. J. C. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural Comput.*, vol. 4, no. 3, pp. 448–472, May 1992.
[61] R. M. Neal, *Bayesian Learning for Neural Networks*, vol. 118. Springer, 2012. [Online]. Available: https://link.springer.com/book/10.1007/978-1-4612-0745-0
[62] J. Lee *et al.*, "Wide neural networks of any depth evolve as linear models under gradient descent," 2019, *arXiv:1902.06720*. [Online]. Available: http://arxiv.org/abs/1902.06720
[63] Y. Gal, "Uncertainty in deep learning," M.S. thesis, Univ. Cambridge, Cambridge, U.K., 2016. [Online]. Available: https://mlg.eng.cam.ac.uk/yarin/thesis/thesis.pdf
[64] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," 2017, *arXiv:1703.02910*. [Online]. Available: http://arxiv.org/abs/1703.02910
[65] G. B. Frisoni *et al.*, "The EADC-ADNI harmonized protocol for manual hippocampal segmentation on magnetic resonance: Evidence of validity," *Alzheimer's Dementia*, vol. 11, no. 2, pp. 111–125, 2015.
[66] C. R. Jack *et al.*, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *J. Magn. Reson. Imag.*, vol. 27, no. 4, pp. 685–691, 2008.
[67] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," 2015, *arXiv:1511.02680*. [Online]. Available: http://arxiv.org/abs/1511.02680
[68] M.-H. Laves, S. Ihler, J. F. Fast, L. A. Kahrs, and T. Ortmaier, "Well-calibrated regression uncertainty in medical imaging with deep learning," in *Proc. Int. Conf. Med. Imag. Deep Learn. (MIDL)*, in Proceedings of Machine Learning Research, vol. 121, T. Arbel, I. B. Ayed, M. de Bruijne, M. Descoteaux, H. Lombaert, and C. Pal, Eds. Montréal, QC, Canada: PMLR, Jul. 2020, pp. 393–412. [Online]. Available: http://proceedings.mlr.press/v121/laves20a.html and https://dblp.org/rec/conf/midl/LavesIFKO20.bib
[69] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9690–9700.
[70] A. Malinin, B. Mlodozeniec, and M. Gales, "Ensemble distribution distillation," 2019, *arXiv:1905.00076*. [Online]. Available: http://arxiv.org/abs/1905.00076