applied sciences

MDPI

Article

Evaluating Uncertainty Quantification in Medical Image Segmentation: A Multi-Dataset, Multi-Algorithm Study

Nyaz Jalal [1,*], Małgorzata Śliwińska [1], Wadim Wojciechowski [2], Iwona Kucybała [3], Miłosz Rozynek [2], Kamil Krupa [2], Patrycja Matusik [2,3], Jarosław Jarczewski [3] and Zbisław Tabor [1]

[1] Department of Biocybernetics and Biomedical Engineering, AGH University, al. Mickiewicza 30, 30-059 Kraków, Poland; malgorzata.anna.sliwinska@gmail.com (M.Ś.); ztabor@agh.edu.pl (Z.T.)
[2] Department of Radiology, Jagiellonian University Medical College, ul. Botaniczna 3, 31-503 Kraków, Poland; wadim.wojciechowski@uj.edu.pl (W.W.); milosz.rozynek@doctoral.uj.edu.pl (M.R.); kamilkrupa955@gmail.com (K.K.); patrycja.s.matusik@gmail.com (P.M.)
[3] Department of Diagnostic Imaging, University Hospital, ul. Jakubowskiego 2, 30-688 Kraków, Poland; iwona.kucybala@uj.edu.pl (I.K.); jaroslaw.dawid.jarczewski@gmail.com (J.J.)
* Correspondence: nyaz@agh.edu.pl

Abstract: Deep learning is revolutionizing various scientific fields, with medical applications at the forefront. One key focus is automating image segmentation, a process crucial in many clinical services. However, medical images are often ambiguous and challenging even for experts. To address this, reliable models need to quantify their uncertainty, allowing physicians to understand the model's confidence in its segmentation. This paper explores how the performance and uncertainty of a model are influenced by the number of annotations per input sample. We examine the effects of both single and multiple manual annotations on various deep learning architectures. To tackle this question, we employ three widely recognized deep learning architectures and evaluate them across four publicly available datasets. Furthermore, we explore the effects of dropout rates on Monte Carlo models by examining uncertainty models with dropout rates of 20%, 40%, 60%, and 80%. Subsequently, we evaluate the models using various measurement metrics. The findings reveal that the influence of multiple annotations varies significantly depending on the datasets. Additionally, we observe that the dropout rate has minimal or no impact on the model's performance unless there is a substantial loss of training data, primarily evident in the 80% dropout rate scenario.

Keywords: segmentation; uncertainty; Monte Carlo dropout; regression; ensembles

check for updates

Citation: Jalal, N.; Śliwińska, M.; Wojciechowski, W.; Kucybała, I.; Rozynek, M.; Krupa, K.; Matusik, P.; Jarczewski, J.; Tabor, Z. Evaluating Uncertainty Quantification in Medical Image Segmentation: A Multi-Dataset, Multi-Algorithm Study. *Appl. Sci.* **2024**, *14*, 10020. https://doi.org/10.3390/app142110020

Academic Editor: Cosimo Nardi

Received: 29 September 2024
Revised: 25 October 2024
Accepted: 30 October 2024
Published: 2 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Artificial intelligence (AI) algorithms are revolutionizing various fields by tackling complex problems and exceeding the capabilities of traditional methods. AI, particularly deep learning (DL), has emerged as a powerful tool in medical diagnosis. AI excels at tasks such as visualization, quantification, segmentation, object tracking, and detection, significantly improving the analysis of medical data [1]. DL unlocks its true potential in analyzing vast quantities of medical data. This capability allows it to uncover previously unknown relationships and achieve superior performance in medical tasks. Consequently, DL has fueled groundbreaking advancements in drug discovery, robot-assisted surgery, personalized medicine, and early cancer detection [2]. One key application of DL is in medical image segmentation. This technique empowers clinicians by assisting in object detection, such as organs and tumors. Additionally, it helps determine size and type with high accuracy. Notably, DL segmentation achieves results comparable to expert human segmentation, while offering significant advantages. Unlike manual methods, DL is less labor-intensive, time-consuming, and susceptible to human bias. However, despite these advantages, widespread adoption of DL segmentation in clinical practice faces a hurdle: the lack of uncertainty quantification. Uncertainty is critical because without it, the model

treats all inputs, both certain and uncertain, as equally reliable. This limitation can pose challenges for informed clinical decision-making [3].

Uncertainty quantification is critical for understanding how confident a DL model is in its predictions. Several techniques can be employed to equip a DL with the ability to express uncertainty. These methods include Bayesian Neural Networks (BNNs), Ensembles, Monte Carlo (MC) Dropout, Quantile Regression (QR), and Evidential Deep Learning (EDL) [4]. While uncertainty plays a significant role, how the medical images are segmented when preparing ground truth training data is another key factor affecting a DL's output. This process, often performed by experts, can introduce bias. Even with identical architecture and data, models trained on differently segmented samples by different physicists can produce significantly different results [5]. To address this, previous research has explored using multiple segmentation approaches. This can involve collaborative segmentation, where experts reach an agreement or segment independently and then combine the results for training. Yet, previous research does not clearly indicate how and to what extent multiple annotations influence the model performance. A segmentation model configuration also plays a crucial role, particularly in the settings related to uncertainty estimation. While we understand that adjustments to model parameters will affect performance, the magnitude of this impact remains unclear. A specific example is the dropout rate in MC models [6]. How does changing this rate influence both the model's accuracy and its uncertainty estimates? Furthermore, is this effect consistent across different models and datasets? In other words, does the dropout rate that yields the best performance (or uncertainty estimates) for one model (or dataset) have any bearing on other models (or datasets)? Again, previous research utilizes different configurations, yet a universal impact of this parameter should be found. Lastly, previous research on uncertainty measurement typically relies on a single dataset focusing on a specific image type and body part. While this approach may lead to good performance on that dataset, it does not guarantee that the uncertainty estimates will generalize to other datasets. This limitation hinders their generalizability and clinical applicability.

This paper explores how multiple annotations for medical image segmentation tasks affect the uncertainty of DL models. We explore this by employing three well-established DL architectures and four datasets. These datasets, labeled by multiple experts, encompass various body parts and image types with different sample sizes and dimensionality. This diverse selection allows for unbiased experiments to assess the impact of multiple annotations on model performance. Additionally, we test MC dropout models with various dropout rates to analyze the influence of dropout on uncertainty and identify the optimal rate. Furthermore, we leverage multiple well-known metrics to comprehensively measure and evaluate the performance of models incorporating uncertainty. Based on this comprehensive analysis, we summarize our key contributions as follows:

- The influence of multiple annotations on segmentation models based on public image segmentation datasets is assessed.
- The influence of dropout rates on the MC models is studied.
- The quantification of uncertainty on different body parts and image types is analyzed.
- The performance of three well-known DL architectures is evaluated in an unbiased environment.
- Various metrics are used to quantify performance of uncertainty estimation methods, including a gray-level generalization of the Dice coefficient, introduced in the present study.
- A new medical images dataset with multiple annotations is made available for public for segmentation uncertainty assessment.

The subsequent sections of the paper are organized as follows: Section 2 reviews the existing literature, highlighting the strengths and weaknesses of previous studies compared to our approach. Section 3 presents an overview of the datasets, model architectures, and metrics employed in this study. Section 4 presents the results based on the gray-level Dice coefficient (GDC) metric (introduced in this paper), followed by a discussion and

conclusion in Section 5. Additionally, Appendices A and B showcase analogous results in Section 3 for the Normalized Mutual Information (NMI) and Normalized Root Mean Square Error (NRMSE) metrics, respectively.

2. Related Work

Many previous works focused on uncertainty measurement with DL for segmenting and classifying medical images. The following section reviews prior research related to DL model architecture, training and testing datasets, the dropout rate for measuring uncertainty with MC models, anatomic locations, image types, and ground truth uncertainty.

Wickstrøm et al. [7] estimate uncertainty for polyps in a colorectal cancer dataset. Three types of architecture named U-Net, SegNet, and FCN-8 were used to segment polyps in the 912 RGB images for 36 patients. The MC dropout was adopted to measure uncertainty with a fixed 50% dropout rate and 10 samples per prediction. Among the neural networks used, FCN-8 achieved the highest IoU and accuracy. Do et al. [8] used 1115 samples to predict segmentation uncertainty, and a 50% dropout rate was set to assess segmentation uncertainty. U-Net architecture was used with a 5 by 5 filter layer and batch normalization to segment myocardial blood flow in the 478 ASL (Myocardial arterial spin labeling) image dataset. This work also investigated the number of dropout samples on inference time and different hyperparameters selection for the Tversky loss function on the estimation accuracy. Hann et al. [6] also proposed U-Net architecture to estimate uncertainty with the ensemble technique. A total of twelve architectures were used; the first six U-Net architectures were created based on 7, 11, 15, 19, 23, and 27 convolutional layers. The remaining architectures were created using a pixel-wise voting technique of the output of the U-Net architectures. The ensemble models were trained on cardiac disorders datasets with a 50% dropout rate and 20 samplings per prediction. After 240 epochs of training, the ensemble model outperformed all individual architectures.

Works like Alonso-Caneiro et al. [9], Song et al. [10], Weldebirhan Arega et al. [11] used similar uncertainty measurements for other medical datasets. For example, Song et al. [10] introduced a U-Net model for segmenting lesions in the oral cancer dataset. The dataset originally contained 492 images, and six types of augmentation, including rotation, zoom, vertical-flipping, horizontal-flipping, brightness adjustment, and gamma correction, were applied to the dataset. Again, a fixed 50% dropout rate with 100 samples per prediction was applied, and the model with MC dropout outperformed the dropout-less U-Net counterpart. Alonso-Caneiro et al. [9] introduced an optical coherence tomography dataset for Stargardt disease. Likewise, a 50% dropout rate was used, yet 2, 5, 10, 20, and 30 samples per prediction were used differently to investigate the effect of sampling on the uncertainty segmentation. A multi-class variant of U-Net model was used to segment the eye's vitreous, retina, and choroid-scleral regions. Finally, Weldebirhan Arega et al. [11] proposed a U-Net architecture for segmenting cardiac regions for a magnetic resonance imaging (MRI) dataset. This work differs from previously mentioned papers since a 10% dropout rate was used instead of 50% and only five samples were used to estimate uncertainty of a single prediction. At the same time, the U-Net model trained for 1000 epochs (higher than others), achieved better results when compared with other models trained with the same dataset.

Roy et al. [12] proposed U-Net architecture for segmenting selected brain parts. Multiple datasets were used in this work: the IXI dataset and part of the MALC dataset were used to train the model, and the remaining samples of the MALC, the ADNI dataset, the CANDI dataset, and the IBSR dataset were used to test the performance of the Bayesian model. The 20% dropout rate and 3, 6, 9, 12, 15, and 18 samples per prediction were used to segment 33 brain structures. Hiasa et al. [13] proposed a semi-supervised segmentation with MC dropout to segment the structures of the human hip. The authors studied multiple DL architectures trained on both labeled and unlabeled datasets. Firstly, the authors trained a deterministic U-Net to separate body parts from other objects in the image. Then, the Bayesian U-Net was used for segmentation and finding uncertainty maps based on the labeled dataset. Finally, the unlabeled dataset was segmented based on the trained model;

if the segmentation uncertainty was high, then experts were asked to correct the automated segmentation; otherwise, the image was processed in the next iteration.

Abdar et al. [4] introduced three uncertainty assessment approaches, including MC dropout, ensemble MC, and deep ensemble, to classify the skin cancer tissues. Each of those uncertainty measurements was tested with DenseNet201, MobileNetV2, ResNet152V2, and InceptionResNetV2 deep neural network architectures. Also, the authors proposed a model based on the Three-Way Decision (TWD) theory, which combines multiple uncertainty measurements and can be adjusted based on each dataset. The experiment used two binary skin cancer datasets, 20% and 50% dropout rates, and 50 samples per prediction. Gudhe et al. [14] developed a BNN architecture called BayesNuSeg. The model was trained on the histopathology image datasets called PanNuke to perform a multi-class segmentation task. The BayesNuSeg model contained one encoder and two decoders. The encoder contained five convolutional layers, and the decoders generated a seed map and pixel embedding. The proposed architecture was compared with multiple Bayesian and deterministic models, and with 5-fold cross-validation, 50% dropout rate, and 500 epochs of training, the BayesNuSeg model achieved the highest IoU and F1-score.

Unlike the above papers, the following studies investigated the influence of dropout rate for measuring uncertainty. Antico et al. [15] used ultrasonography (US) and MRI datasets for segmenting knee femoral cartilage. Bayesian CNN with MC dropout was used to measure uncertainty and compared with a similar CNN architecture without MC dropout. The model was trained with a 10% and 50% dropout rate to evaluate the dropout rate on the model. Fifty samples per prediction were used to quantify uncertainty, and after 31 epochs, the Bayesian CNN achieved a higher Dice score for both MRI and US images compared to CNN without uncertainty assessment. Camarasa et al. [16] introduced U-Net architecture for multi-class segmentation of the carotid artery MRI dataset. The goal of this work was to compare the performance and reliability of various uncertainty metrics. One of the experiments tests nine different dropout rates, including 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%. After 600 epochs of training, the model with a 30% dropout achieved 99%, 76%, and 85% Dice scores for background, vessel wall, and lumen, respectively. Largent et al. [17] also investigated the MC dropout rate by testing Bayesian U-Net architecture with 10%, 20%, 30%, 40%, and 50% dropout rates. In addition to the fixed dropout rate, they also implemented a so-called "concrete dropout", which treats the dropout rate as a hyperparameter that can be optimized during the training. The performance of the Bayesian U-Net was compared with DenseNet, U-Net without uncertainty, and the ensemble model by combining DenseNet and U-Net. The models trained based on MRI images of the infant's brain after 100 epochs U-Net with concrete dropout achieve the highest result.

Singh et al. [2] proposed a framework called SkiNet, which segments lesions in images of skin cancer and then classified the lesions into predefined types. The dataset was imbalanced and contained seven classes of skin lesions; sample augmentation was performed to balance the dataset. Again, MC dropout was adapted to add uncertainty assessment to the U-Net model. The dropout rate from the 40% to 70% range was applied (without specifying the exact number), and the best result was achieved with a 50% dropout. Ng et al. [3] proposed different techniques to measure uncertainty in two MRI cardiac datasets. The techniques were ensemble, Backprop, Stochastic Segmentation Networks, and MC dropout. In the case of MC dropout, both 10% and 50% with 50 samples per prediction were evaluated. Both datasets are multi-class segmentation datasets, and after 50 epochs of training, the ensemble model achieved superior segmentation accuracy. Yet, uncertainty with 10% MC achieved better results compared to 50% MC dropout. Lastly, Bórquez et al. [18] adapted the uncertainty concept for classifying breast cancer tissues using BNN. The dataset contained four classes based on the HER2 factor categorized as 0, 1+, 2+, and 3+. One of the experiments tested a dropout rate of 20%, 50%, and 80% during the testing process. Using 5-fold cross-validation after 200 epochs, the model with a 20% dropout rate achieved the highest result compared to other dropout rates.

Finally, Zhao et al. [1] proposed the Bayesian approach and integrated it with the original structure of the U-Net architecture. The authors claimed that the model makes various predictions with high uncertainty for hard inputs. Conversely, the model remains consistent when input is easy during a training process. The model was trained and evaluated based on two public MRI cardiac datasets. The proposed model slightly outperforms MC dropout and achieves higher efficiency compared to the ensemble model. Baumgartner et al. [5] segmented two medical datasets using U-Net architecture. To measure uncertainty, the hierarchical probabilistic method creates the conditional distribution for input samples. The model generates low-resolution segmentation samples and then improves the distribution in higher resolutions using multiple latent variables. The model uses LIDC-IDRI 3D CT images and 2D MRI for the prostate, both annotated by four radiologists. The model trained with single and multiple annotations separately achieved superior results compared to deterministic and probabilistic U-Net. Kohl et al. [19] combined U-Net architecture with a conditional variational autoencoder (CVAE) to assess uncertainty of segmentation. The CVAE part creates low-dimensional latent space, and then a random sample from this distribution was provided to the U-Net part to predict probabilistic segmentation. A medical and non-medical dataset were used to train and evaluate the model performance. The medical dataset is the LIDC-IDRI, but only 2D sections were used. We tried to reimplement the last two mentioned models based on their public GitHub repository and compare their result with our findings; however, we failed since the code was not maintained for a long time and there was a package inconsistency preventing us from running the code.

The previous works mainly focused on specific elements of uncertainty measurement with DL by applying uncertainty architecture for particular datasets or using specific hyperparameters. We believe measuring uncertainty with DL should be covered in more detail in the following aspects:

- Segmenting (or classifying) images or lesions differ from expert to expert; individual assertions are usually biased and insufficient for robust labeling. Therefore, adding uncertainty to ground truth images is essential.
- A robust model needs to be selected based on experiments with different image modalities, different anatomic locations, and multiple task types (multi-class segmentation, binary segmentation).
- The impact of dropout rate on uncertainty estimation in MC dropout architectures remains unclear. Some studies achieved the best results with low dropout rates, while others found success with high dropout rates. Therefore, it is essential to assess the impact of dropout rate in a fair and unbiased setting across various image types and with a sufficient training dataset.

To fulfill the mentioned gaps, we separately trained the ensemble, the regression, and MC models with 20%, 40%, 60%, and 80% dropout rates. Each model was separately trained with four public datasets labeled by multiple experts. Finally, multiple metrics were used to measure and evaluate the performance of uncertainty models. Table 1 shows the comparison between our and existing literature.

Table 1. Summarize the reviewed papers by examining various aspects of uncertainty.

Study	Feature 1 [1]	Feature 2 [2]	Feature 3 [3]	Feature 4 [4]	Feature 5 [5]
[14]	NO	YES	NO	NO	NO
[2]	NO	NO	NO	YES	NO
[17]	NO	NO	NO	YES	YES
[18]	NO	NO	NO	YES	NO
[4]	YES	NO	NO	YES	YES
[13]	NO	YES	NO	NO	NO
[11]	NO	NO	NO	NO	NO
[15]	YES	NO	NO	YES	NO

Table 1. *Cont.*

Study	Feature 1 [1]	Feature 2 [2]	Feature 3 [3]	Feature 4 [4]	Feature 5 [5]
[9]	NO	NO	NO	NO	NO
[7]	NO	NO	NO	NO	NO
[10]	NO	NO	NO	NO	NO
[8]	NO	NO	NO	NO	NO
[12]	NO	YES	NO	NO	NO
[3]	NO	NO	NO	YES	YES
[16]	NO	NO	NO	YES	NO
[6]	NO	NO	NO	NO	NO
[1]	YES	NO	NO	NO	YES
[5]	YES	YES	YES	NO	NO
[19]	NO	NO	YES	NO	YES
Our Work	YES	YES	YES	YES	YES

[1] Multiple medical image types, [2] multiple anatomic locations, [3] multiple ground truths, [4] dropout rate influence, [5] model comparison.

3. Materials and Methods

3.1. Materials

To assess the model's performance, four datasets with varying sizes and characteristics were used to enhance its generalization. As shown in Table 2, these datasets differ in several aspects, including image type, image size, body part, image dimensions, and dataset complexity. This diversity ensures the model's robustness and allows for a fair comparison between models.

Table 2. Summarize the datasets property utilized in this paper.

Properties	KNEE	SIJ	LUNG	HEART
Total Num.	47	345	1338	114
Training Num.	31	267	1029	65
Testing Num.	16	78	309	49
Image Size	448 by 448	Vary	Vary	512 by 512
Image Type	MRI	MRI	CT	X-ray
Num. Objects [1]	1	1	1	2
Segmented objects	Synovitis	Bone Oedema	Lung Nodules	Lung and Heart
Num. Radiologists	3	3	3 or 4	3
Body Part	Knee	Sacroiliac Joint	Chest	Chest
Image Dimension	2D	2D	3D	2D

[1] In addition to background.

3.1.1. Knee Dataset

This dataset was introduced by Gut et al. [20] for rheumatoid arthritis patients. The dataset targets synovitis areas in the human knee joint, containing 3D T1-weight contrasted MRI images. The dataset contains 205 cross-section 2D images annotated by three radiologists, yet only 48 of them were annotated due to a clear view of epiphyses of the tibia and femur. The MRI images were obtained by a 3.0 Tesla Achieva MRI scanner from Philips Healthcare. The MRI images were acquired in a coronal plane using an eight-channel knee coil. As one of these 48 images was used to establish manual annotation procedure between different annotators, the dataset's final version consists of 47 images (with 141 segmented objects) of which 31 images were used for training, and the remaining 16 images were used for testing. Figure 1 shows a sample from a Knee dataset with segmentation by three experts.

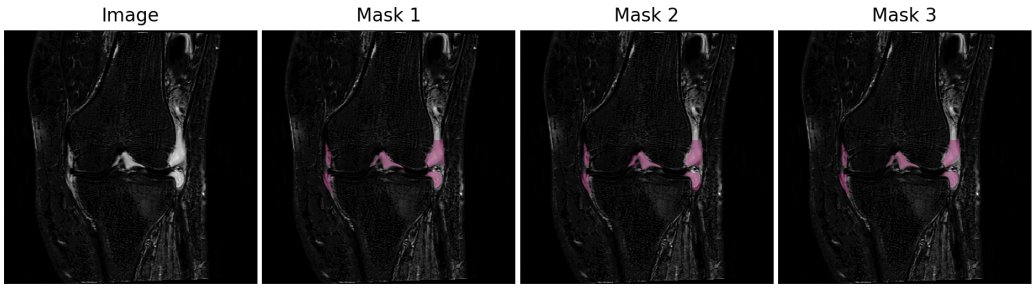


Figure 1. The original image and three annotations by different radiologists from the Knee dataset.

3.1.2. SIJ Dataset

This dataset was originally introduced in Rzecki et al. [21], which investigated axial spondyloarthritis lesions (areas of damage) at the sacroiliac joints. The MR images of this dataset were acquired in the coronal oblique plane. The sacroiliac joints on both sides were examined from the front to the back parallel with the long axis of the sacrum. In total, 3D images of nine patients with sacroiliitis were studied. The dataset consists of 345 images (with 1035 segmented objects) annotated by three radiologists. In total, 267 images were used for training, and the remaining 78 images were used for testing. Figure 2 shows a sample from a SIJ dataset with segmentation by three experts.

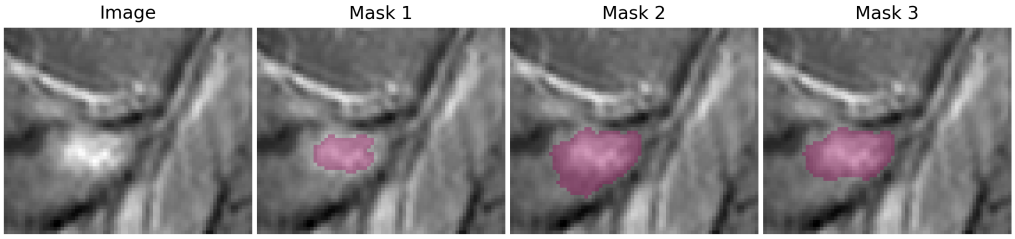


Figure 2. The original image and three annotations by different radiologists from the SIJ dataset.

3.1.3. Lung Dataset

The LIDC-IDRI is a public Lung dataset. It contains 1018 3D thoracic CT images of a chest from 1010 patients. Either three or four radiologists (out of 12) segmented abnormal lesions (lung nodules). This study implemented the following steps to create training and testing samples: first, for each CT scan, all spatial regions containing lung nodules were found based on manual annotations provided in the LIDC-IDRI database. Then, padding was added around each nodule, 20 pixels in the axial plane of a CT image and 2 pixels in the CT axial direction. Finally, the lung nodule regions with the padding were extracted from the CT scans and together with manual annotations were used as a dataset. The dataset's final version consists of 1338 3D patches, each patch containing a lung nodule. In total, 1029 patches were used for training, and 309 patches were used as testing samples. Figure 3 shows a sample from a Lung dataset with segmentation by three experts.

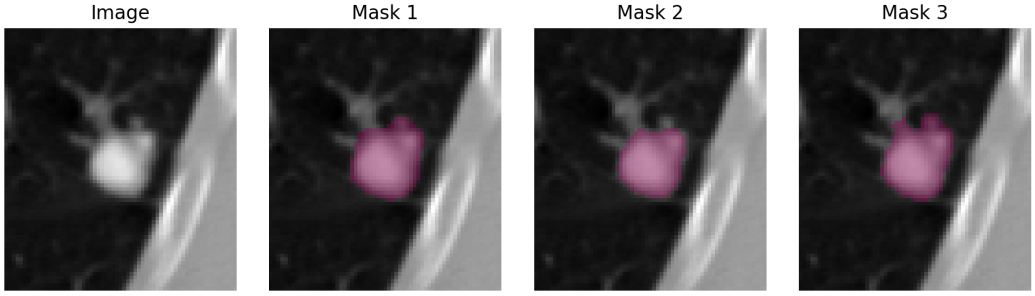


Figure 3. The original image and three annotations by different radiologists from the Lung dataset.

3.1.4. Heart Dataset

This dataset contains X-ray imaging of the chest. X-rays were prescribed mostly for routine check-ups or due to various cardiovascular or pulmonary disorders. Chest X-rays were taken in the posterior-anterior projection in the upright position with the chest in the mid-inspiratory condition using Siemens Siemens Ysio Max or Axiom Luminos dRF. The X-ray images were resized to a unified size of 512 by 512, and then three radiologists manually segmented the lungs and the heart. Radiologists separately annotated a single X-ray image and then analyzed it together to unify the process of segmentation of the lungs and heart. The dataset's final version consists of 114 2D images (342 segmentation); 65 images were used for training, and the remaining 49 images were used for testing. Figure 4 shows a sample from a Heart dataset with segmentation by three experts. We published this dataset in the following repository.

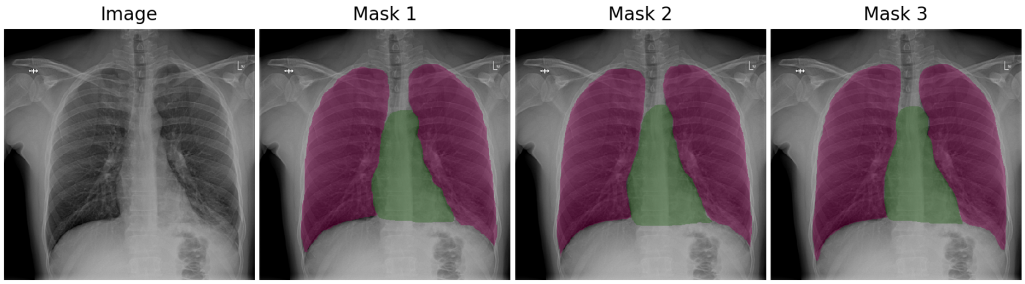


Figure 4. The original image and three annotations by different radiologists from the Heart dataset.

3.1.5. Dataset Preparation

The datasets were prepared for training using the following procedure:

- Image multiplication: For each image in each dataset, there are at least three manual annotations of the targeted regions. In the image datasets based on multiplication, each image was multiplied as many times as there were manual annotations corresponding to it. Thus, for example, for the Heart dataset, for which there are exactly three manual annotations for each X-ray image, the training set consisted of 65×3 X-ray images, of which only 65 were unique. The ground truth (GT) set consisted of unique 195 manual annotations.
- No multiplication of images: in the image datasets with no multiplication for each image, a single ground truth manual annotation is randomly selected from the available manual annotations and used for training. Thus, for example, for the Heart dataset, the training set consisted of 65 X-ray images, and the ground truth set also consisted of 65 manual annotations.
- Certainty maps: A certainty map is a function that assigns a probability to each image pixel or voxel that a structure of interest exists in this pixel or voxel. In this work, all available manual segmentations for each image were averaged, resulting in gray-level segmentation certainty maps. These certainty maps were used as ground truth (GT) for training. For example, in the Heart dataset, the training set included 65 X-ray images, and the ground truth consisted of 65 gray-level averages, each based on three manual annotations (since each X-ray image had three corresponding manual annotations).

The GT certainty maps, described above, were calculated for all data sets for all training and testing images. The GT certainty maps, calculated for testing images, were used to evaluate the trained models in the task of assessment of segmentation (un)certainty.

3.2. Architecture Description

The nnU-Net framework [22] is a standardized and flexible tool for medical image segmentation, proven to outperform other DL approaches in various tasks. To use nnU-Net for any specific segmentation task, the only requirement is to format the raw data according to specific guidelines, including file naming conventions, folder structure, and data con-

figuration files. Once prepared, nnU-Net ensures consistent preprocessing for all models, such as resampling, normalization, and cropping.

By default, nnU-Net employs five-fold cross-validation and trains five models. Since data generation in nnU-Net is separated from training, this ensures that, after finalizing the data generators (and thus fixing the data split into cross-validation folds), all models are trained on the exact same data, with identical augmentations. This guarantees a fair comparison of model performance.

As shown in Figure 5, we extended nnU-Net framework to train the following deep models for predicting segmentation (un)certainty:

- Ensemble models,
- MC dropout models,
- Regression models.

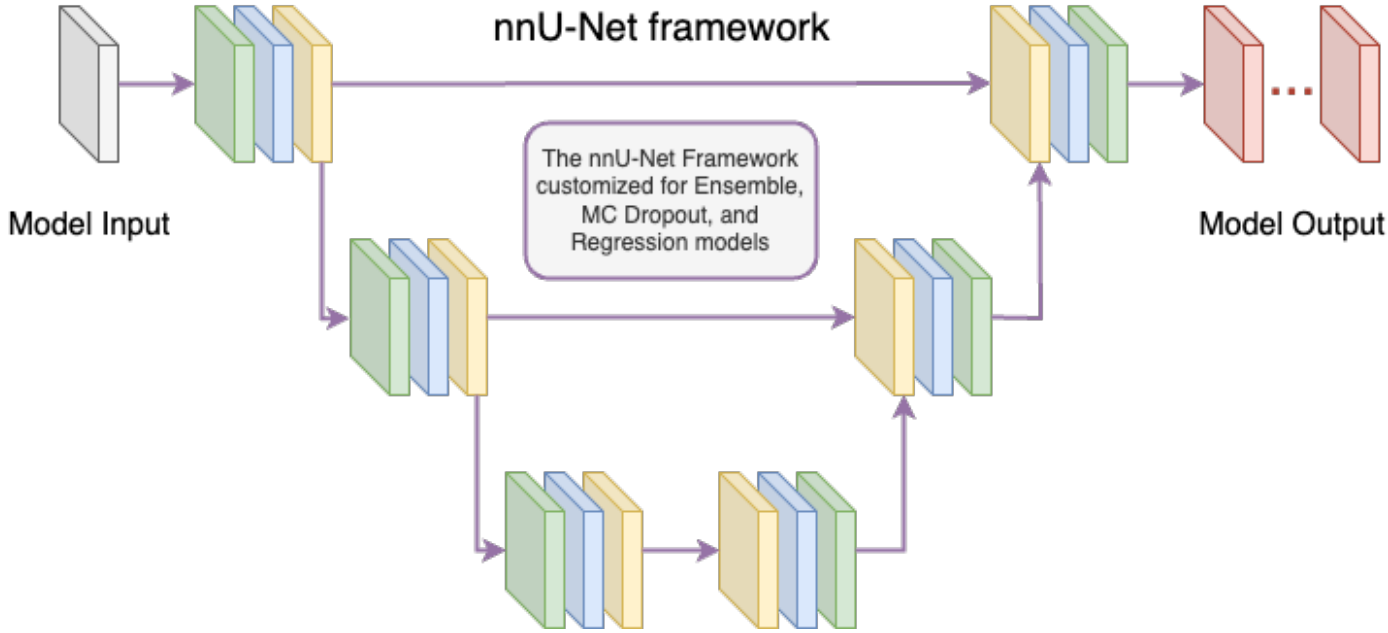


Figure 5. Show the model architecture based on the nnU-Net framework. The model has one input channel, and the number of output channels equals the number of segmented classes.

The ensembles used in this study consisted of the five models trained by default within the nnU-Net framework. For each input image, these five models generated five separate predictions, and the average of these predictions was used to estimate segmentation (un)certainty, which was then compared with the ground truth (GT) certainty maps, as described in the previous section. The ensembles were trained using two versions of the training sets: with and without multiplication. These experiments aimed to determine whether multiple annotations for a single image are necessary for reliable segmentation (un)certainty estimation.

To train the MC dropout models, dropout rates for U-Net were set to nonzero values of 20%, 40%, 60%, and 80%, with the network remaining in training mode during inference. For each dropout rate, 30 MC samples were drawn, and the average of these 30 predictions was used to estimate segmentation accuracy as a function of dropout rate. Specifically, the average of the MC dropout samples was compared with the GT certainty maps to assess segmentation (un)certainty.

Converting the nnU-Net segmentation framework into a regression framework requires some modifications. The segmentation task must be reformulated as a multi-channel input task, where the number of input channels equals one plus the number of segmentation classes for foreground objects. For all datasets except the Heart dataset, this results in two input channels, as we are solving a binary segmentation task. In the case of the Heart

dataset, which contains two foreground classes (lungs and heart), there are three input channels. The first channel always contains the original image, while the remaining channels hold the ground truth (GT) certainty maps for the foreground objects. Although manual segmentation masks are processed by nnU-Net data generators, they are not used during training. Instead, GT certainty maps are extracted from the input data channels and serve as reference data during training. The original nnU-Net loss function is replaced by the mean absolute error (MAE) loss to train the regression models. These models directly predict segmentation certainty maps, which are then compared with the GT certainty maps during testing. Our re-implementation of nnU-Net is available publicly on GitHub.

3.3. Metric Description

The following metrics were used to evaluate the model's performance:

- Gray-level Dice coefficient (GDC): This metric measures the overlap between two gray-level, positive-definite images (such as the certainty maps analyzed in this work) and is an extension of the Dice coefficient (DC), which is commonly used to assess the quality of segmentations. As shown in Equation (1), the GDC is an enhanced version of the DC that can be applied to both binary and gray-level images, evaluating segmentation accuracy within the same range. Similar to the standard DC, the GDC value ranges from 0 to 1, where 0 indicates no overlap between the images, and 1 indicates perfect overlap.

$$\mathrm{GDC} = \frac{2 \times \sum_{i=1}^{M} \sum_{j=1}^{N} \sqrt{Ref_{(i,j)} * Seg_{(i,j)}}}{\sum_{i=1}^{M} \sum_{j=1}^{N} Ref_{(i,j)} + \sum_{i=1}^{M} \sum_{j=1}^{N} Seg_{(i,j)}} \quad (1)$$

- Normalized Root Mean Square Error (NRMSE): This metric measures the similarity between two images. As shown in Equation (2), the NRMSE is estimated by calculating the RMSE (Root Mean Square Error) divided by the Euclidean normalization for the reference image. In addition to the Euclidean, other normalizations can be used with the NRMSE, such as min_max and mean normalization. The expected score is between 0 and 1, and a lower value of NRMSE indicates better matching between the images.

$$\mathrm{NRMSE} = \frac{\sqrt{\mathrm{MSE}}}{\| Ref \|} \quad (2)$$

- Normalized Mutual Information (NMI): This metric measures the similarity between two images. As shown in Equation (3), the NMI is estimated by calculating the entropy I between the two images divided by the mutual information H of the same images. The score value is between 0 and 1 (In the scikit-image python package, the range is between 1 and 2), and a higher value of NMI indicates better matching between the images.

$$\mathrm{NMI}\ (S, T) = \frac{I(S;T)}{\sqrt{H(S) \cdot H(T)}} \quad (3)$$

4. Results

This section presents the performance of each DL architecture on individual datasets. The results are visualized in three subsections using Box plots, Bland–Altman plots, and Correlation plots. All results are based on the GDC metric. For NMI and NRMSE metrics, please refer to Appendices A and B, respectively. It is important to note that the y-axis scale varies across the figures because model performance differs between datasets.

4.1. Box Plot

Figure 6 presents a summary of the model's performance using the GDC metric in a box plot. The performance of the models varies significantly depending on the dataset

used. For instance, all models exhibit strong performance with the Heart dataset, whereas the outcomes are less favorable with the Knee and SIJ datasets. A key observation from the box plot is that the dropout rate has minimal or no impact on models performance, as evidenced by similar results across different dropout rates for both the SIJ and Lung datasets. In the Heart dataset, only 80% dropout scores less (yet very accurate) results. However, the Knee dataset shows fluctuation in performance by reaching below 80% mean, and then the results improve for 40% and 60% dropouts and decline again for 80% dropouts. Another indication is the impact of multiple or single annotations on the segmentation performance. Again, the results highly depend on the datasets; for example, in the Lung dataset, both models perform equally, with a median of around 92% and an almost similar IQR distribution. The most obvious difference is for the SIJ dataset when the multiple annotations ensembles (MA) model performs a 20% higher median compared to the single annotation ensemble (SA) model. However, all the models perform below 80% median for this dataset. Furthermore, the MA model achieved a higher GDC median for the Heart dataset for segmenting the heart (HHS). Oppositely, the SA model scored higher results in the case of the Knee dataset. Moreover, when comparing the ensemble models with models with MC dropouts, the results show that at least one MC dropout version scores a higher median than the SA and MA models. Also, the IQR distribution for ensemble models is much wider in the Lung and SIJ datasets; for other datasets, a version of MC dropout models always achieves a tighter IQR distribution. Finally, the regression model performs very poorly for the Knee dataset but scores competitive results for other datasets. Nevertheless, all MC models with a dropout rate equal to or less than 60% always perform better than the regression model. However, when the regression and ensemble models are compared, the model's performance varies depending on the datasets, yet the regression model usually scores a wider distribution.

4.2. Bland–Altman Plot

Figure 7 shows the Bland–Altman plot for the ensemble model with multiple annotations compared to other models using the GDC metric; each row in the graph represents the implementation of a separate dataset. The impact of using multiple annotations varies across datasets. In the Knee dataset, for instance, the average performance with the MA model is 3% lower than using a single annotation, indicated by a negative mean difference (MD) of −3%. Conversely, for other datasets, the MA model leads to positive improvements. The Lung dataset sees a 1% increase, the HHS dataset a 5% increase, and the SIJ dataset a significant 12% increase in performance compared to using a single annotation. The Limits of Agreement (LoA) between the models provide another key insight. The datasets with high GDC scores show a very narrow LoA, indicating strong consensus among the models. Conversely, datasets with lower GDC scores, like SIJ and Knee, exhibit a much wider LoA, suggesting greater disagreement between the models' predictions. Moreover, the dataset dependence phenomenon is also true when comparing the MA model and regression model. The regression model achieves a −9% MD for the SIJ dataset, yet the MA model performs a higher score for the remaining datasets. In the case of the Knee dataset, the MA model scored +28% MD as the highest score, and +1% is the lowest positive MD for the Heart dataset for segmenting the lungs (HLS). Furthermore, when the results of the MA model are compared with different versions of MC dropout, MC with 40% and 60% dropout rates always achieved better (positive) MD; for MC with 20% and 80% dropout rates, there is a narrow MD (both positive/negative) between the models. In summary, the MA model shows weaker performance with the SIJ datasets compared to other datasets, achieving competitive but not optimal results. A similar pattern is observed with the LoA: narrower LoA values correlate with higher GDC scores in the datasets, while wider LoA values occur in scenarios where GDC scores are lower.

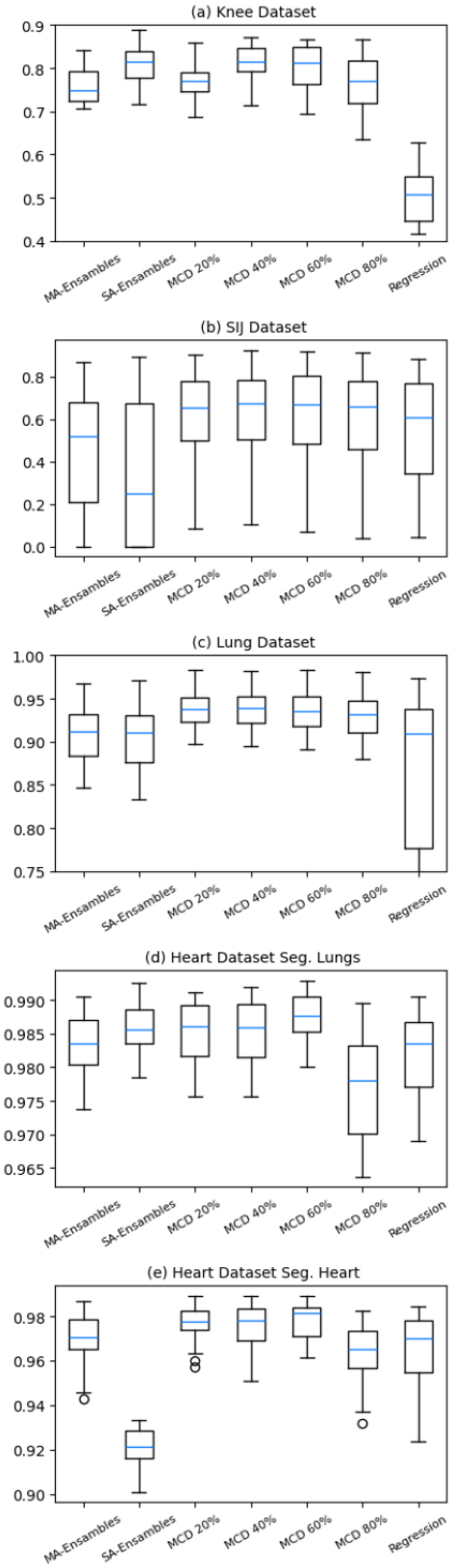


Figure 6. Shows box plots for the GDC metric.



Figure 7. Shows Bland-Altman plots for the GDC metric.

4.3. Correlation Plot

Figure 8 explains the correlation plot for the GDC metric, the ensemble model with multiple annotations on the x-axis, and the remaining models on the y-axis. The correlation plot shows the same dataset-related conclusion as the previous plots. For example, when the SA and MA models are compared, the correlation plot shows wide spreads of points for the Knee and SIJ datasets, which indicates the models' disagreement. Also, in the SIJ dataset, the graph shows two clusters. One at the bottom indicates that the MA model is scored higher for those samples, and the other at the top right indicates the agreement between the models. For the Lung and Heart datasets (for the Heart dataset, the y- and x-axis ranges are very small, which is why we consider them as a clear cluster), the plot shows a cluster of points at the top right very tightly around the lines, which suggests a strong correlation between the models. Moreover, when the MA model is compared with the MC models, we see the same distribution for all the graphs, which clearly indicates that changing the dropout rate has little impact on the model's performance. For example, for the Knee and SIJ datasets, the plot shows a wide distribution of the points as an indicator of disagreement between the models (also true for the regression model). One interpretation of Knee dataset results is that the dataset size is relatively small. Again, for the Lung and Heart datasets, the graph shows a cluster of points in the top right for all MC dropout versions, and the points are very tight to the line as proof of the agreement between the models. Yet, some outliers appear directly below the cluster, representing samples of MC models that score low GDC. Finally, comparing the MA and regression models shows a strong agreement for the Heart dataset. For the Lung dataset, the points are widely spread on the right side of the plot, which suggests the MA model outperforms the regression model. The IQR of the regression model in Figure 6 is much wider than that of other models, which draw the same conclusion as shown in a correlation plot.

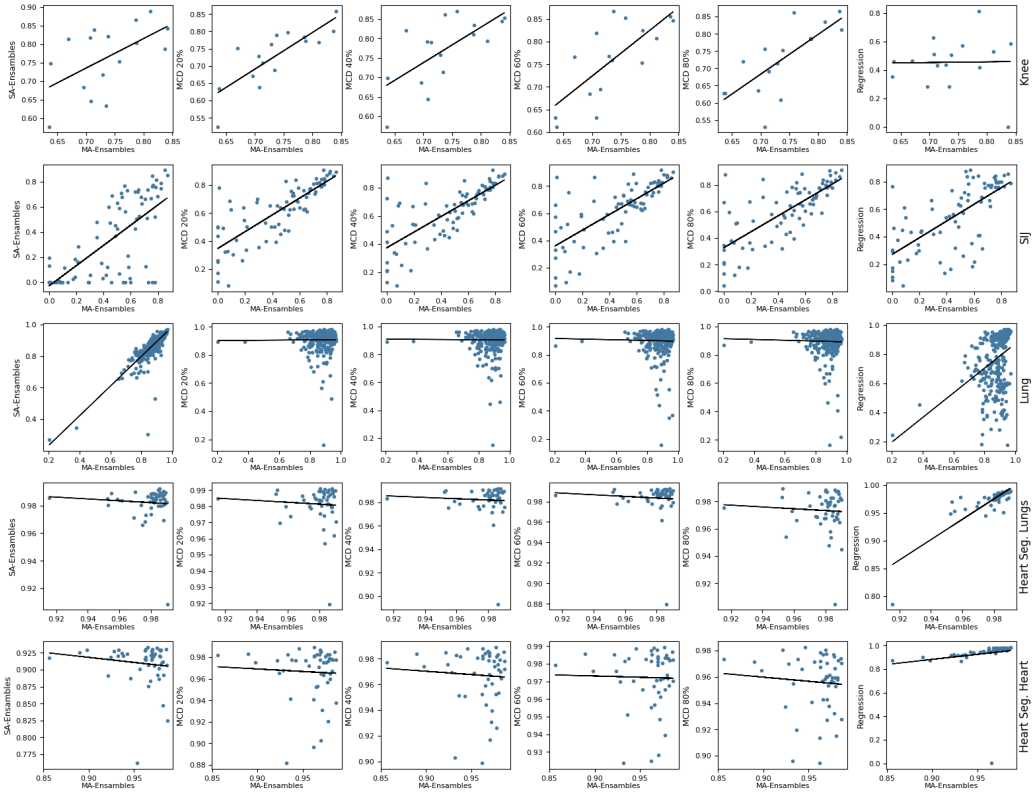


Figure 8. Shows correlation plots for the GDC metric.

5. Discussion

Prior studies often focused on specific aspects of the uncertainty segmentation, such as using a particular dataset or model architecture. However, these results can have limited generalizability, as a model might perform well on a specific dataset but poorly on unseen data or produce unrealistic certainty values, hindering its clinical application. This work aimed to shed light on multiple aspects of using deep learning with uncertainty quantification for medical image segmentation. Most prior studies relied on single-annotator datasets. A drawback of this approach is the inherent variability in expert opinions. To address this limitation, at least three experts segmented each dataset in this study. Subsequently, models were trained on both single-annotation and multiple-annotation versions to evaluate the impact of multiple annotations on uncertainty measurement. Another limitation in prior research was its narrow focus on specific datasets or image types, potentially hindering model generalization. In contrast, we trained our models on diverse datasets covering various body parts and imaging modalities, including CT scans, MRIs, and X-rays. Our study employed three established architectures—ensembles, regressions, and MC dropout—implemented in a customized nnU-Net framework. Each architecture was trained separately on all datasets to facilitate a fair comparison and to understand how performance varies with dataset complexity. Furthermore, we investigated the impact of MC dropout rates (20%, 40%, 60%, 80%) on uncertainty models.

The present study uses four datasets: Knee, SIJ, Lung, and Heart, each showing varied model performance, largely influenced by their distinct characteristics detailed in Table 2. Dataset size is a crucial factor: Knee, with the smallest size (47 images), results in modest performance across models (median GDC below 85%). Conversely, the Lung dataset, the largest (1338 images), consistently performs well across all models (median NRMSE below 0.1, except for MA at 0.25). The ample data in the Lung dataset likely enhances model learning. However, larger sample sizes do not always lead to better model performance. The Heart and SIJ datasets illustrate this point. Despite having only 65 images, the Heart dataset yields superior performance across all models. In contrast,

the SIJ dataset, with 345 images, shows significantly lower performance in both GDC and NMI metrics, with all models scoring median GDC below 80%. These contrasting results suggest that factors beyond sample size, such as annotation diversity and data complexity, significantly influence model efficiency.

We further explored how multiple annotations affect model performance. Intuitively, more annotations should improve performance due to the combined expertise of multiple observers. However, our findings show no consistent advantage for models trained with multiple annotations over those trained with single annotations. In some cases, single-annotation models even outperform multiple-annotation models. Notably, the MA model scored 5% higher GDC median for the HHS dataset and 0.3 less NRMSE median for the SIJ dataset. Conversely, the SA model scored 7% higher GDC median for the Knee dataset and 0.15 lower median NRMSE for the Lung dataset. Generally, both models demonstrate similar performance in other cases. Our results show that multiple annotations lead to a clear improvement in uncertainty estimation in only one case—heart segmentation in X-ray images, with a roughly 5% increase in GDC. However, we believe this improvement may be due to chance, as no significant enhancements were observed for other datasets when multiple annotations were used for training. We think the limited impact of multiple annotations on uncertainty prediction accuracy is related to the local properties of the images being segmented. Specifically, when the boundaries of the segmented objects are clear, annotators can delineate them with high certainty. Conversely, when borders are blurred, annotators tend to make more errors in delineation. A model can learn to estimate uncertainty based on these local relationships between boundary clarity and delineation errors. Multiple annotations are not necessary to capture this correspondence. Local examples of discrepancies between object boundaries and manual delineations are sufficient, and these can be gathered from a single annotator's work. Therefore, as our study suggests, averaging multiple annotations has only a marginal effect on improving uncertainty estimation.

When comparing ensemble models with MC dropout versions on the NMI metric, ensemble models generally outperform MC models except for the SIJ dataset, where performance is equal. For the GDC and NRMSE metrics, the performance difference between the two models is usually negligible. Ensemble models also consistently outperform regression models on all datasets for the NMI metric. However, GDC and NRMSE performance varies by dataset. For instance, the ensemble model outperforms the regression model on the Knee dataset but performs similarly on the Heart dataset for GDC, with the regression model showing a wider IQR. The regression model slightly outperforms the ensemble model on the Heart dataset for NRMSE. The study also examines the impact of dropout rates on MC models. Performance differences across various dropout rates are minimal, except for the 80% dropout rate, which shows slightly lower performance. Occasionally, specific dropout rates yield better results for certain datasets. For example, the 60% dropout model achieves the best NMI and GDC scores on the Heart dataset, while the 40% dropout model produces the highest GDC score on the Knee dataset. The limited impact of the dropout rate may suggest that the model's weights are organized in such a way that critical information is distributed across multiple neurons or connections. This redundancy allows the model to preserve key information even when some neurons are randomly 'dropped out' during training.

Furthermore, this study used three metrics to evaluate model performance: GDC, NMI, and NRMSE. The conclusions drawn from these metrics vary by dataset and model. Sometimes, the metrics agree, such as with the Knee dataset, where the regression model performs the worst across all three metrics, while the SA model and MC with 40% dropout achieve the highest scores. In contrast, the Lung dataset shows significant disagreement between the metrics. All models score a median GDC above 90%, but all score below 1.4 on the NMI metric. The NRMSE metric aligns more with GDC, though ensemble models show performance decline and numerous outliers. Another key finding is that GDC scores are consistently higher than NMI and NRMSE. For example, in the Heart dataset, all models

achieve GDC scores near 97%, but none surpass an NMI score of 1.7 or an NRMSE score lower than 0.1. Additionally, the analysis revealed potential biases within the metrics. The SIJ dataset shows outstanding NRMSE scores, yet poor performance on GDC and NMI metrics. Similarly, the NMI metric tends to favor the MA model in most cases, suggesting a bias towards this model.

Computational complexity is an important consideration for any model, especially those with high resource demands, such as ensemble models or MC dropout. In our case, this complexity is minimized by using a well-known framework like nnU-Net, which automatically optimizes the model's hyperparameters and adapts to the data. Additionally, the experiments were run on a high-performance system with ample RAM and GPU resources. As a result, the model's inference time is very fast. For example, the dropout model takes approximately 0.2 and 0.16 s to segment a single sample from the Lung and SIJ datasets, respectively. Similar results are observed with other models, such as the ensemble model, which requires around 0.13 and 0.25 s to segment a single sample from the Knee and Heart datasets, respectively. Also, it is important to note that there were instances where the model failed to segment images or produced low-quality results. These failures can occur at the dataset level, where all models struggled to segment certain datasets, as seen with the SIJ dataset. Additionally, failures can be specific to a particular model; for example, the regression model produced significantly lower-quality segmentation for the Knee dataset compared to other models. These issues could be attributed to factors such as dataset complexity, the number of samples, and the models' capabilities.

In conclusion, this study explored the effects of multiple annotations and dropout rates on uncertainty segmentation using three deep learning architectures across four datasets. We assessed the influence of multiple annotations on segmentation models using public datasets and analyzed the impact of dropout rates on MC models. Additionally, we examined uncertainty quantification across different body parts and image types. The study evaluated the performance of three widely used deep learning architectures in an unbiased setting, utilizing various metrics to quantify the performance of uncertainty estimation methods, including a gray-level generalization of the Dice coefficient introduced in this work. The main conclusion of this paper is that the choice of uncertainty estimation method, whether MC dropout or ensembles, and the specific MC dropout rate, does not significantly impact performance. Similarly, the study design—whether using multiple or single annotations—appears to have little effect on the accuracy of uncertainty segmentation. Based on our findings, we recommend that simpler experimental designs, such as using a single annotation per image and either MC dropout or ensembles, can be used interchangeably for assessing segmentation uncertainty.

Author Contributions: Conceptualization, N.J. and Z.T.; methodology, N.J.; software, N.J., M.Ś. and Z.T.; validation, N.J. and Z.T.; formal analysis, N.J.; investigation, N.J.; data curation, W.W., I.K., M.R., K.K., P.M. and J.J.; writing—original draft preparation, N.J.; writing—review and editing, M.Ś., W.W., I.K., M.R., K.K., P.M., J.J. and Z.T.; visualization, N.J.; supervision, Z.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Science Center, Poland, grant number is ARTIQ/0004/2021.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: Data is contained within the article.

Acknowledgments: We gratefully acknowledge the funding support by the program "Excellence Initiative—Research University" for the AGH University in Krakow and the ARTIQ project: UMO-2021/01/2/ST6/00004 and ARTIQ/0004/2021. The funding source was not involved in planning, conducting the research, or preparing the article. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2023/016441.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. NMI

Appendix A.1. Box Plot

Figure A1 summarizes the model's performance in the box plot using the NMI metric (the range value is between 1, which indicates perfectly uncorrelated images, and 2, which indicates perfectly correlated images either in a positive or negative direction). Again, like the GDC metric, there are no clear winners; the models perform differently based on the datasets. However, the ensemble models perform better than the MC dropout and regression models except for the SIJ dataset, in which the MC dropout versions perform the same as the MA model and are better than the SA model. Yet, none of the models achieved perfect correlation; the Heart dataset scored around 1.7 median correlation as the best result. Also, there is a significant discrepancy between the GDC and NMI metrics on some datasets. For instance, on the Heart dataset, the SA model achieved a very low GDC score, with a median of only 92%. However, the same model achieves a much higher NMI score, with a median around 1.7. Interestingly, even within the MC dropout versions themselves, there is no significant difference in performance. There is a fluctuation among all the datasets. For instance, the Heart dataset achieved the best results with a 60% dropout rate, while the Lung dataset performed best with a 20% dropout rate. This trend continued with slight variations across other versions as well. Moreover, there is no clear superiority between the MA or SA models in segmentation performance. Similarly, the results highly depend on the datasets; for example, in the Knee dataset, The SA model achieved a higher median than the MA model but with a much wider IQR distribution. Conversely, the MA model performed better for the SIJ dataset. For the remaining datasets, the median and IQR distributions are almost the same. Finally, the regression model scored very low NMI for the Knee and Lung datasets but scored competitive results for other datasets. Nonetheless, when the regression and ensemble models were compared, the ensemble models consistently scored higher NMI median. Similar to the GDC metric, all MC models with a dropout rate equal to or less than 60% always score better NMI than the regression model.

Appendix A.2. Bland–Altman Plot

Figure A2 shows the Bland–Altman plot for the ensemble model with multiple annotations compared to other models using the NMI metric; each row in the graph represents the implementation of a separate dataset. The SA and MA NMI are scored almost equally for all the datasets. The SA model performs slightly better for the Knee and Heart datasets, and the MA model performs slightly better for the SIJ and Lung datasets. However, the Lung dataset scored the narrowest LoA, and the Heart dataset scored a wide LoA (poorer agreement). Previously, the Heart dataset achieved a very tight LoA (strong agreement) for GDC metrics. The correlation plot for both metrics confirms this conflict in which the points are very tight to the line for the GDC, and for the NMI, the points are spread all around. Furthermore, when the results of the MA model were compared with different versions of MC dropout, MA consistently scored higher NMI regardless of the dropout rates. However, the different range depends on the datasets. For example, for the SIJ dataset, the MD is +2%, meaning the MA model performs better on average by 2%. For the Lung and Knee dataset, MD is approximately +10%. For the Heart, the MD is above +15%, and the number increases when the dropout rate increases; for instance, for Lung segmentation with 80% dropout, the MD reaches +30%. This result is completely opposite of the result of the GDC metric in which MC models with 40% and 60% dropout rates always achieve better (positive) MD (The result of the RMMSE metric is in between NMI and GDC, discussed in the next appendix). Also, the LoA is very wide for all comparison models. Similar results appeared when the MA model was compared with the regression model. The MA model scores +21%, +5%, +26%, and +24%, and +22% MD for the Knee, SIJ, Lung, HHS, HLS, respectively. The conclusion is that the NMI metric always shows superiority toward the MA model compared to MC dropout and regression models.

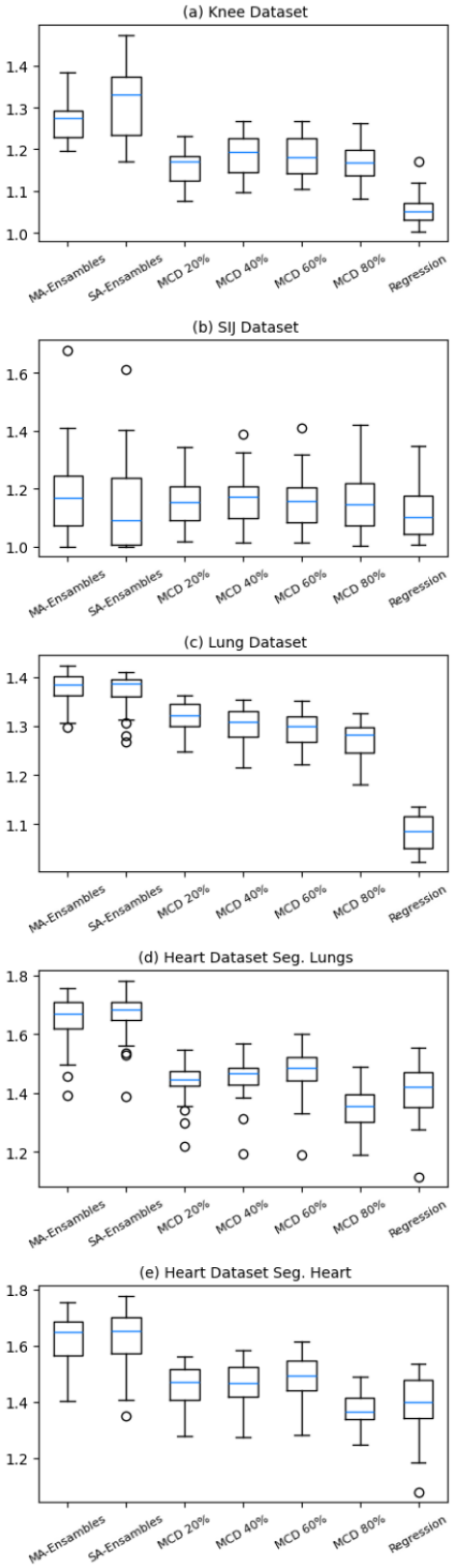


Figure A1. Shows box plots for the NMI metric.

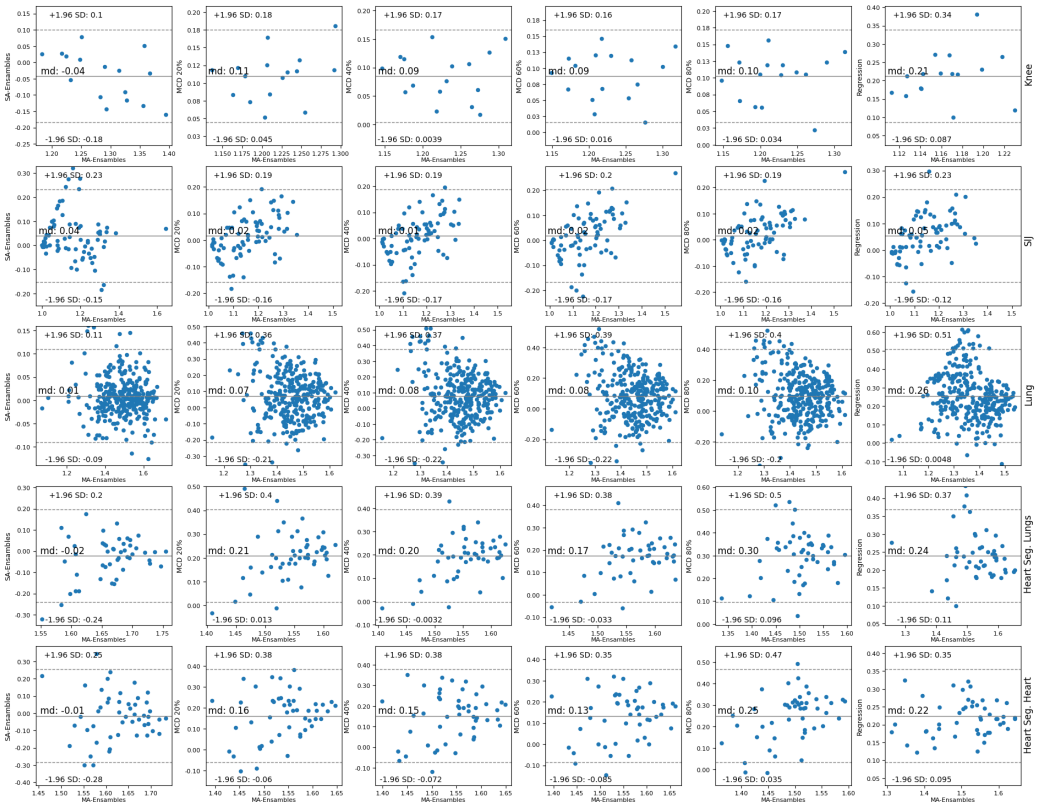


Figure A2. Shows Bland–Altman plot for the NMI metric.

Appendix A.3. Correlation Plot

Figure A3 explains the correlation plot for the NMI metric, the ensemble model with multiple annotations on the x-axis, and the remaining models on the y-axis. When comparing the MA and SA models, the points in the Knee dataset are scattered around the line (similar to the GDC plot). The SIJ dataset shows two patterns: first, the horizontal points in the bottom left, meaning the MA model outperforms the SA model; second, the remaining points are close to the line, meaning agreement between the models. The Bland–Altman plot shows the same conclusion when the MA scored +12% average bias compared to the SA. For the Lung and Heart datasets, the plot shows points that are very close to the scatter line, which indicates strong agreement between the models. Also, in the Lung dataset, the plot shows a robust positive linear correlation, meaning that both models make the same predictions across different ranges of data points. Furthermore, when the MA model was compared with MC versions, opposite to the GDC metric in the Lung dataset, there were no cluster points in the top right; instead, the points were widely scattered on the right side of the plot. This phenomenon is correct for all MC models, which indicates that the MA model scored higher NMI compared to the MC models (Bottom right is the area where MC prediction is declining). The box plot confirms this result in which the MA model's average and IQR were higher than the MC versions. For the SIJ dataset, both the MA and MC models score low NMI, yet there is a high consistency among the models. For the Heart dataset, the MA model scored higher NMI compared to MC versions; also, points are spread more widely compared to the GDC metric, especially for higher dropout rates in which there is conflict between the models for a subset of samples. Finally, comparing the MA and regression models for the Heart dataset revealed that the MA model performed better, yet there is a positive correlation indicating consistency between the models. For the Lung dataset, the points are widely spread on the right side of the plot, which suggests the MA model scored high NMI for multiple samples in which regression scores the low NMI. Again, similar to MC models, both models score low yet consistent results for the SIJ dataset.

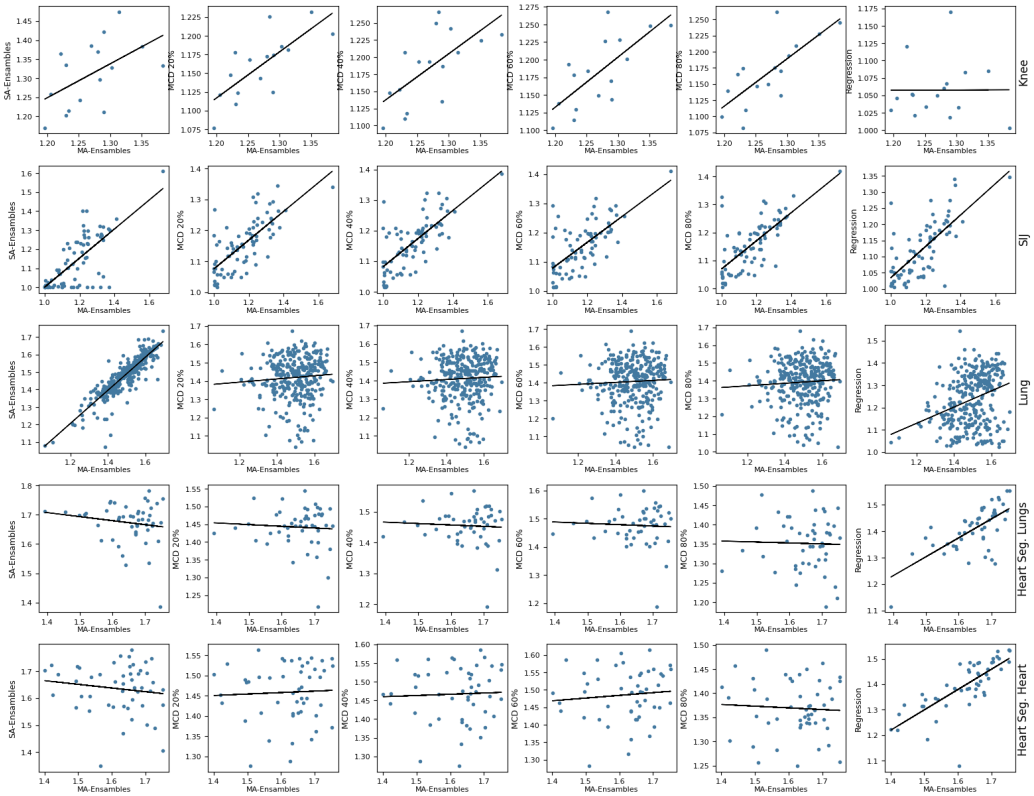


Figure A3. Shows correlation plot for the NMI metric.

Appendix B. NRMSE

Appendix B.1. Box Plot

Figure A4 summarizes the model's performance in the box plot using the NRMSE metric (the range value is between 0, which indicates minimum error rate, and 1 indicating maximum error rate). Again, like previous metrics, no model outperforms others; the models score differently based on the datasets. When MA and SA models were compared for the SIJ dataset, the MA model overtook the SA model; this conclusion is also correct for the GDC and NMI metrics. Oppositely, the SA model scored better for the Lung dataset with a lower median and less diverse IQR. Yet, the plot shows many outliers in both datasets. For the remaining datasets, both MA and SA models almost perform equally with the higher IQR for the MA in the Knee and HHS datasets. Furthermore, comparing the metrics shows some disagreement between the GDC and NRMSE metrics. For example, for the HHS dataset, the MA model achieved around 5% higher GDC median, yet the SA model scored a lower NRMSE median and narrower IQR for the same dataset. Similarly, when comparing the MC dropout models again, there is no obvious indication of disparity. There is a fluctuation among all the datasets. For example, the Knee dataset with 20% and 40% dropout rates scores equally, yet for the 60% and 80% dropout rates, both median and IQR scores worsen. However, for the SIJ and HLS datasets, the MC models perform equally except for an 80% dropout rate, in which the performance declines. For the Lung dataset, all MC models perform equally, and the MC model with a 20% dropout rate achieves the lowest median for the HHS dataset. Moreover, MC models perform better (yet with many outliers) than non-MC models only in the Lung dataset; in the remaining datasets, the box plot does not show any clear conclusion and shows high instability between the results. Finally, the regression model scores a high NRMSE median for the Knee and a very big IQR for Lung datasets but scores competitive results for other datasets. For example, the HHS dataset scored the minimum median and smallest IQR. Moreover, the GDC metric supported this result in which the regression model scored a high GDC median for the Heart dataset and scored a very big IQR for the Lung dataset (yet 90% GDC median).

Nonetheless, the regression model scored modest results for the NMI metric for the Heart and Lung datasets, demonstrating variance explanation between the measurement metrics.

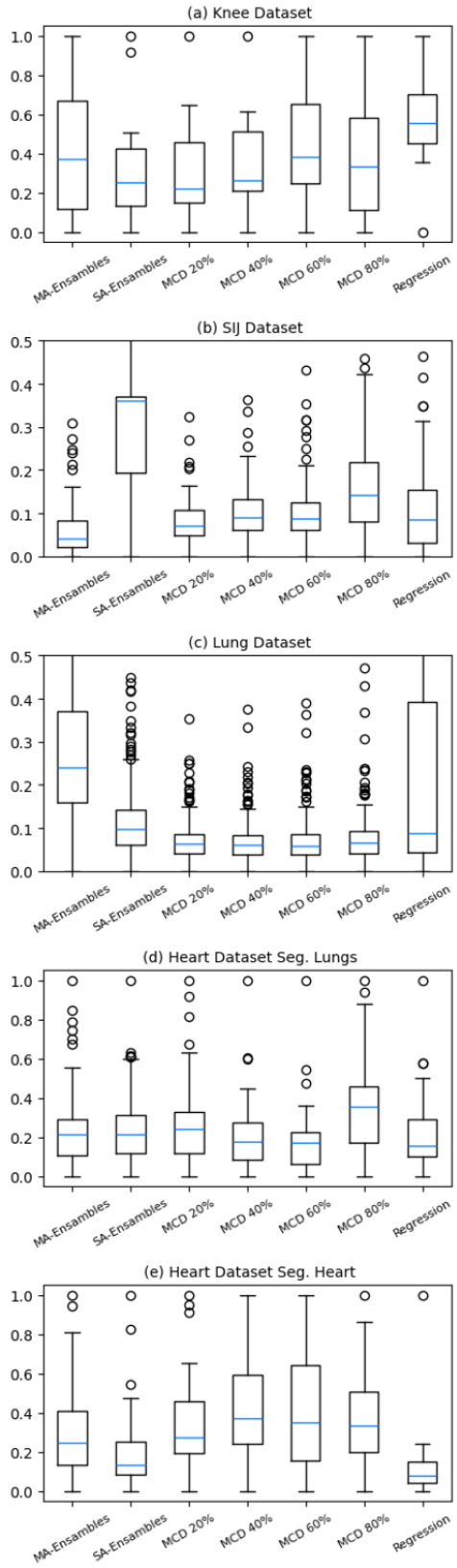


Figure A4. Shows box plots for the NRMSE metric.

Appendix B.2. Bland–Altman Plot

Figure A5 shows the Bland–Altman plot for the ensemble model with multiple annotations compared to other models using the NRMSE metric; each row in the graph represents the implementation of a separate dataset. Previously, the NMI metric suggested that ensemble models outperform other models; conversely, the GDC metric showed the superiority of MC dropout over ensemble models, especially with 40% and 60% dropout rates. Similar to the GDC, the NRMSE metric shows dataset dependence results. For example, for the Lung dataset, the MA model scored a positive bias (For NRMSE metric, positive MD means the MA model performs worse) compared to other models, suggesting the MA model performs poorly. For the SIJ dataset, the MA model outperforms the SA model by scoring −21% MD; it also outperforms the MC model with an 80% dropout rate by −8% MD and scored almost similar MD with the regression and remaining MC models. For the HHS dataset, the regression model scored the best NRMSE, yet the MA model scored higher MD compared to all MC versions but still less than the SA model. For the HLS dataset, similar to the GDC metric, MC models with 40% and 60% outperform the MA model by +5% and +8%, respectively. Yet the MA model scored higher bias −2% and −10% MD compared to 20% and 80% MC versions. However, both the regression and SA models outperform the MA model by +1% and +6% MD. Finally, for the Knee dataset, the MA model only outperforms the regression and MC with a 60% dropout, but the other models outperform the MA model by less than 10% or less MD. Considering LoA, the plot shows wide LoA in many cases, suggesting strong disagreement between the models, which is also indicated by the box and correlation plots.

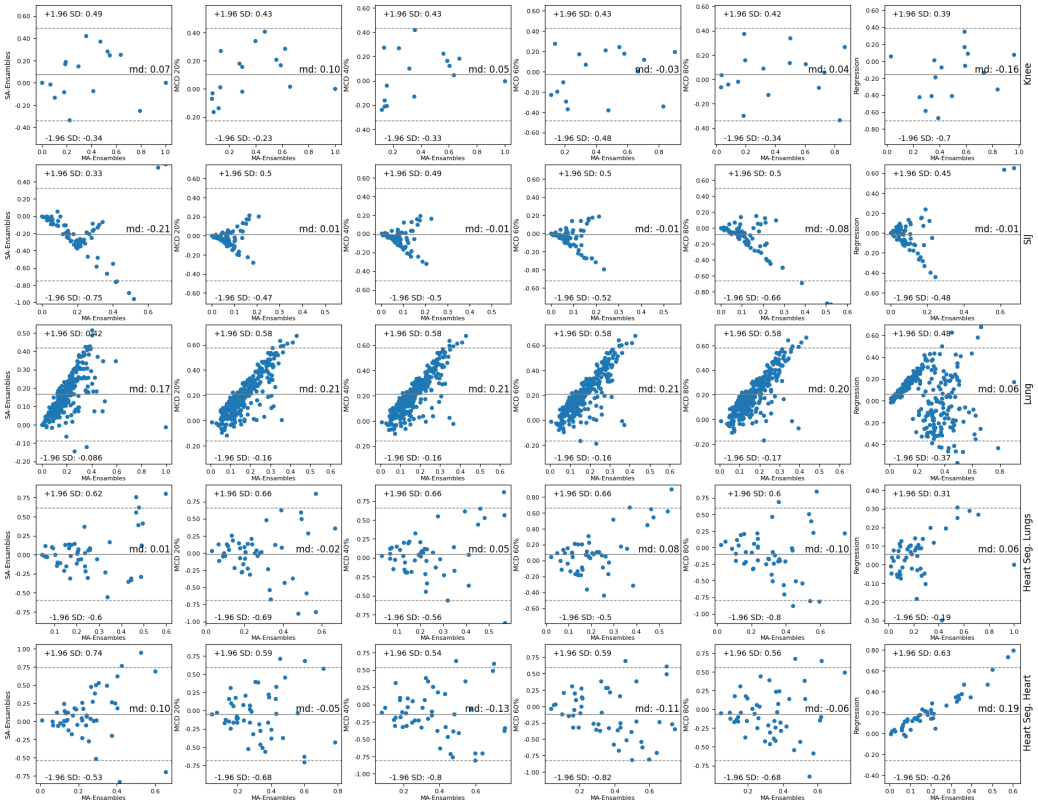


Figure A5. Shows Bland–Altman plot for the NRMSE metric.

Appendix B.3. Correlation Plot

Figure A6 explains the correlation plot for the NRMSE metric, the ensemble model with multiple annotations on the x-axis, and the remaining models on the y-axis. When comparing the MA model to other models for the Knee dataset, the plot shows a positive correlation, indicating strong agreement. However, this outcome is not robust due to the

number of samples in the Knee dataset. For the remaining dataset, when the MA and SA models were compared, the plot for the SIJ dataset shows a vertical line on the plot's left side, which clearly reveals that the MA model outperforms the SA model. The Bland–Altman plot shows the same conclusion when the MA model scored −21% average bias compared to the SA model. For the Lung dataset, the SA model scored better results, yet the points are evenly spread around the line, showing a positive correlation between the models. Also, the plot shows multiple points in the [0.5–1] x-axis range, revealing the MA low scoring. Both models perform well for the Heart dataset, yet for HHS, there is a broader spread of points interpreted in the box plot as a bigger IQR for the MA model. Furthermore, when the MA model was compared with MC versions, like the GDC metric in the Lung dataset, a cluster of points in the bottom left indicates that all MC models outperform the MA model. The Bland–Altman plot draws the same conclusion, in which all MC models score above +20% average bias, which also suggests that the MC dropout rate has a limited effect on the model's performance. For the HHS dataset, the MA scored better NRMSE; however, MC, with 40% and 60% dropout rates, scored better for segmenting the Lungs. Moreover, the points are more spread out when the dropout rate increases, suggesting correlation decreases when the dropout rate increases. For the SIJ dataset, the MA model and MC models perform equally, with strong agreement between the models. Finally, comparing the MA and regression models for the Heart dataset revealed that the regression model performed better (opposite of NMI), yet there is a positive correlation indicating consistency between the models (similar to NMI). For the Lung dataset, the points are spread all around the plot, indicating very low consistency between the models. Also, the box plot shows a big IQR for both models, presenting the same conclusion. Finally, for the SIJ dataset, a cluster point appears on the bottom left, suggesting strong consistency between the models with multiple outlier samples in both directions.

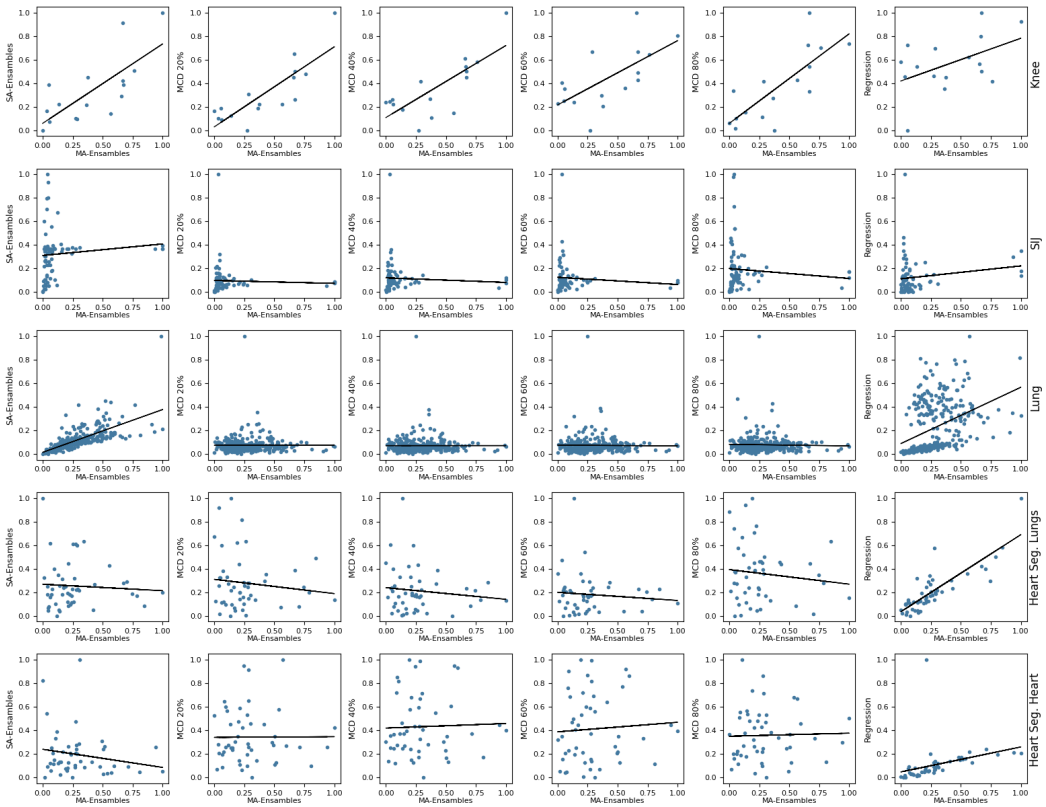


Figure A6. Shows correlation plot for the NRMSE metric.

References

1. Zhao, Y.; Yang, C.; Schweidtmann, A.; Tao, Q. Efficient Bayesian Uncertainty Estimation for nnU-Net. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022. MICCAI 2022*; Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022. [CrossRef]
2. Singh, R.K.; Gorantla, R.; Allada, S.G.R.; Narra, P. SkiNet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability. *PLoS ONE* **2022**, *17*, e0276836. [CrossRef] [PubMed]
3. Ng, M.; Guo, F.; Biswas, L.; Petersen, S.E.; Piechnik, S.K.; Neubauer, S.; Wright, G. Estimating Uncertainty in Neural Networks for Cardiac MRI Segmentation: A Benchmark Study. *IEEE Trans. Biomed. Eng.* **2023**, *70*, 1955–1966. [CrossRef] [PubMed]
4. Abdar, M.; Samami, M.; Dehghani Mahmoodabad, S.; Doan, T.; Mazoure, B.; Hashemifesharaki, R.; Liu, L.; Khosravi, A.; Acharya, U.R.; Makarenkov, V.; et al. Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning. *Comput. Biol. Med.* **2021**, *135*, 104418. [CrossRef] [PubMed]
5. Baumgartner, C.F.; Tezcan, K.C.; Chaitanya, K.; Hötker, A.M.; Muehlematter, U.J.; Schawkat, K.; Becker, A.S.; Donati, O.; Konukoglu, E. PHiSeg: Capturing Uncertainty in Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019, Proceedings of the 22nd International Conference, Shenzhen, China, 13–17 October 2019*; Proceedings, Part II 22; Springer International Publishing: Cham, Switzerland, 2019.
6. Hann, E.; Gonzales, R.A.; Popescu, I.A.; Zhang, Q.; Ferreira, V.M.; Piechnik, S.K. Ensemble of Deep Convolutional Neural Networks with Monte Carlo Dropout Sampling for Automated Image Segmentation Quality Control and Robust Deep Learning Using Small Datasets. In *Annual Conference on Medical Image Understanding and Analysis*; Springer International Publishing: Cham, Switzerland, 2021; pp. 280–293.
7. Wickstrøm, K.; Kampffmeyer, M.; Jenssen, R. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Med. Image Anal.* **2020**, *60*, 101619. [CrossRef] [PubMed]
8. Do, H.P.; Guo, Y.; Yoon, A.J.; Nayak, K.S. Accuracy, uncertainty, and adaptability of automatic myocardial ASL segmentation using deep CNN. *Magn. Reson. Med.* **2020**, *83*, 1863–1874. [CrossRef] [PubMed]
9. Alonso-Caneiro, D.; Kugelman, J.; Tong, J.; Kalloniatis, M.; Chen, F.K.; Read, S.A.; Collins, M.J. Use of uncertainty quantification as a surrogate for layer segmentation error in Stargardt disease retinal OCT images. In Proceedings of the DICTA 2021—2021 International Conference on Digital Image Computing: Techniques and Applications, Gold Coast, Australia, 29 November–1 December 2021. [CrossRef]
10. Song, B.; Li, S.; Sunny, S.; Gurushanth, K.; Mendonca, P.; Mukhia, N.; Patrick, S.; Peterson, T.; Gurudath, S.; Raghavan, S.; et al. Exploring uncertainty measures in convolutional neural network for semantic segmentation of oral cancer images. *J. Biomed. Opt.* **2022**, *27*, 115001. [CrossRef] [PubMed]
11. Weldebirhan Arega, T.; Bricq, S.; Meriaudeau, F. Leveraging Uncertainty Estimates to Improve Segmentation Performance in Cardiac MR. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis, Proceedings of the 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, 1 October 2021*; Proceedings 3; Springer International Publishing: Cham, Switzerland, 2021; pp. 24–33.
12. Roy, A.G.; Conjeti, S.; Navab, N.; Wachinger, C. Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage* **2019**, *195*, 11–22. [CrossRef] [PubMed]
13. Hiasa, Y.; Otake, Y.; Takao, M.; Ogawa, T.; Sugano, N.; Sato, Y. Automated Muscle Segmentation from Clinical CT using Bayesian U-Net for Personalized Musculoskeletal Modeling. *IEEE Trans. Med. Imaging* **2019**, *39*, 1030–1040. [CrossRef] [PubMed]
14. Gudhe, N.R.; Kosma, V.M.; Behravan, H.; Mannermaa, A. Nuclei instance segmentation from histopathology images using Bayesian dropout based deep learning. *BMC Med. Imaging* **2023**, *23*, 162. [CrossRef] [PubMed]
15. Antico, M.; Sasazawa, F.; Takeda, Y.; Jaiprakash, A.T.; Wille, M.L.; Pandey, A.K.; Crawford, R.; Carneiro, G.; Fontanarosa, D. Bayesian CNN for Segmentation Uncertainty Inference on 4D Ultrasound Images of the Femoral Cartilage for Guidance in Robotic Knee Arthroscopy. *IEEE Access* **2020**, *8*, 223961–223975. [CrossRef]
16. Camarasa, R.; Bos, D.; Hendrikse, J.; Nederkoorn, P.; Kooi, E.; van der Lugt, A.; de Bruijne, M. Quantitative Comparison of Monte-Carlo Dropout Uncertainty Measures for Multi-class Segmentation. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis, Proceedings of the Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, 8 October 2020*; Proceedings 2; Springer International Publishing: Cham, Switzerland, 2020; Volume 12443, pp. 32–41. [CrossRef]
17. Largent, A.; De Asis-Cruz, J.; Kapse, K.; Barnett, S.D.; Murnick, J.; Basu, S.; Andersen, N.; Norman, S.; Andescavage, N.; Limperopoulos, C. Automatic brain segmentation in preterm infants with post-hemorrhagic hydrocephalus using 3D Bayesian U-Net. *Hum. Brain Mapp.* **2022**, *43*, 1895–1916. [CrossRef] [PubMed]
18. Bórquez, S.; Pezoa, R.; Salinas, L.; Torres, C.E. Uncertainty estimation in the classification of histopathological images with HER2 overexpression using Monte Carlo Dropout. *Biomed. Signal Process. Control* **2023**, *85*, 104864. [CrossRef]
19. Kohl, S.A.A.; Romera-Paredes, B.; Meyer, C.; De Fauw, J.; Ledsam, J.R.; Maier-Hein, K.H.; Eslami, S.M.A.; Jimenez Rezende, D.; Ronneberger, O. A Probabilistic U-Net for Segmentation of Ambiguous Images. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; Volume 31.

20. Gut, D.; Trombini, M.; Kucybała, I.; Krupa, K.; Rozynek, M.; Dellepiane, S.; Tabor, Z.; Wojciechowski, W. Use of superpixels for improvement of inter-rater and intra-rater reliability during annotation of medical images. *Med. Image Anal.* **2024**, *94*, 103141. [CrossRef] [PubMed]
21. Rzecki, K.; Kucybała, I.; Gut, D.; Jarosz, A.; Nabagło, T.; Tabor, Z.; Wojciechowski, W. Fully automated algorithm for the detection of bone marrow oedema lesions in patients with axial spondyloarthritis–Feasibility study. *Biocybern. Biomed. Eng.* **2021**, *41*, 833–853. [CrossRef]
22. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.