Visualization, Quantification, And Analysis Of Inter-rater Variability To Enhance Deep Learning-based Medical Image Segmentation Of Paraspinal Muscles

Parinaz Roshanzamir

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Electrical and Computer Engineering) at

Concordia University

Montréal, Québec, Canada

December 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Parinaz Roshanzamir**

Entitled: **Visualization, Quantification, And Analysis Of Inter-rater Variability To Enhance Deep Learning-based Medical Image Segmentation Of Paraspinal Muscles**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Electrical and Computer Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Wei-Ping Zhu

_____ External Examiner
Dr. Tristan Glatard

_____ Examiner
Dr. Wei-Ping Zhu

_____ Co-supervisor
Dr. Hassan Rivaz

_____ Co-supervisor
Dr. Yiming Xiao

Approved by _____
Yousef R. Shayan, Chair
Department of Electrical and Computer Engineering

_____ 2023

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Visualization, Quantification, And Analysis Of Inter-rater Variability To Enhance Deep Learning-based Medical Image Segmentation Of Paraspinal Muscles

Parinaz Roshanzamir

Deep learning-based medical image segmentation has revolutionized healthcare diagnostics. While the accuracy offered by these models is important, ensuring their practical implementation requires a comprehensive reliability assessment. Among the sources of uncertainty, inter-rater variability, which reflects natural disagreements among annotators, has been historically overlooked. Studying this variability can be a key factor in improving the robustness and reliability of models.

Basing our experiments on paraspinal muscle segmentation, which has significant value in studies related to low back pain, in this dissertation, we first proposed a novel multi-task TransUNet model to accurately segment paraspinal muscles while predicting inter-rater labeling variability visualized using a variance map of raters' annotations. Benefiting from the transformer mechanism and convolution neural networks, our algorithm was shown to perform better or similar to the state-of-the-art methods while predicting and visualizing multi-rater annotation variance per muscle group in an intuitive manner. Subsequently, we studied the relationship between inter-rater variability and aleatoric/epistemic uncertainties, in the context of DL model architecture and label fusion methods. Specifically, we measured aleatoric and epistemic uncertainties using test-time augmentation, test-time dropout, and deep ensemble to explore their relationship with inter-rater variability. Furthermore, we compared UNet and TransUNet to study the impacts of Transformers on model uncertainty with two label fusion strategies. This thesis provides novel frameworks for visualizing, understanding and quantifying inter-rater variability to better inform relevant deployment and implementation of DL models for medical image segmentation.

Acknowledgments

First and foremost, I would like to express my deep gratitude to my supervisors, Dr. Hassan Rivaz and Dr. Yiming Xiao, for their guidance and support throughout my studies. I appreciate the innovative ideas, effective leadership, and thorough reviews they provided, offering me an excellent opportunity to learn and thrive in my research. I also want to extend my gratitude to my colleagues at the IMPACT and HEALTH-X labs, who have been very helpful and supportive.

Special thanks to my dear family and friends, without whose generous support none of this would have been possible. I am deeply grateful for all the love and support that I have received from them.

This thesis, along with all its findings, was highly dependent on the carefully collected dataset of paraspinal muscles. My gratitude goes out to all those individuals who played a role in collecting and annotating the data.

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and NVIDIA for donation of the GPU.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, advancements in deep learning (DL) have reshaped the landscape of various fields. From the developments in natural language processing [45] to revolutionizing healthcare with groundbreaking applications in medical imaging and disease diagnosis [27, 67], deep learning has emerged as a driving force behind cutting-edge technological innovations.

Among all of the medical applications of deep learning, medical image segmentation is considered to be a core step in many automatic image analyses. The segmentation process consists of creating precise boundaries within complex imaging data to identify distinct structures and anomalies. An example of an automatic medical image segmentation is demonstrated in Fig. 1.1, where brain tumors are segmented from MRIs. This process is crucial for accurate diagnosis and prognosis, treatment planning, and continuous monitoring of medical conditions. Segmenting certain anatomical features in medical images and showing their spatial boundaries can assist clinicians in detecting abnormalities and facilitate more comprehensive and precise monitoring of patients' health. Manual image segmentation is very time-consuming. Moreover, segmenting medical images requires the raters to have a certain level of expertise which makes it even more challenging to find eligible raters for the task. Automatic image segmentation, especially through deep learning-based models can mitigate these issues and produce accurate results by incorporating the opinions of multiple experts in a significantly shorter amount of time. These advancements play a vital role in the development of sophisticated computer-aided diagnostic tools, accelerating the speed and improving the accuracy of medical assessments.

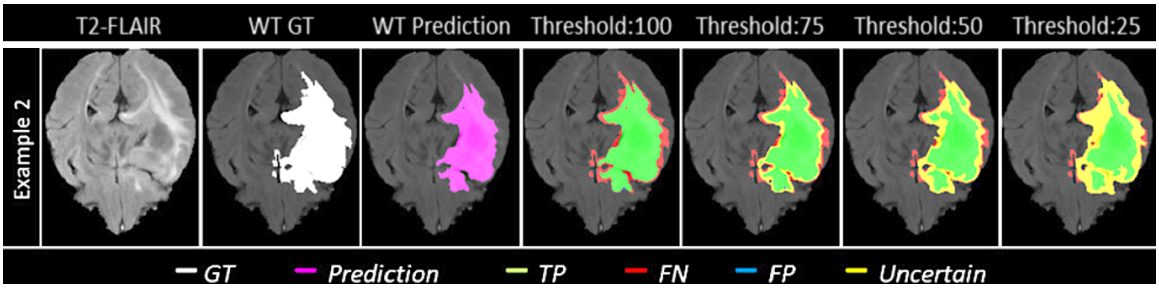

Figure 1.1: An example of a brain tumor segmentation task, showing model predictions along with the uncertainty maps thresholded at multiple levels [37]

Paraspinal muscles provide essential support to the spine, and abnormalities in their composition or function have been closely linked to various musculoskeletal disorders, including low back pain which is a leading cause of global disability. A more detailed description of low back pain and its relation to paraspinal muscles is provided in section 1.1. Due to the importance of automatic paraspinal muscle segmentation for studying the top musculoskeletal disorder in adults, we decided to base our experiments on this segmentation task in the presented thesis. We used a dataset of axial MRI scans of patients with low back pain, collected in a collaboration between Concordia and Western Universities. More details regarding the dataset have been provided in the following chapters.

As the DL-based models increasingly find applications in critical decision-making processes, understanding and quantifying uncertainties associated with their predictions becomes a necessity for their practical implementation. The importance of uncertainty assessment in medical applications cannot be overstated, as it plays a critical role in enhancing the reliability and interpretability of deep learning models. In the realm of healthcare, where decisions directly impact patients' well-being, understanding the confidence and limitations of DL model predictions is extremely important and can facilitate trust and transparency in the integration of these models into clinical workflows. Uncertainty estimation can serve as a valid assessment of the reliability of computer-facilitated diagnostic results and enable clinicians to make more informed decisions, eventually resulting in optimizing patient care and clinical decision-making.

1.1 Paraspinal Muscle Segmentation and Its Importance in Low Back Pain-related Studies

Low Back Pain (LBP) is the most common skeletomuscular disorder in adults and has a lifetime prevalence of more than 80% [43]. Understanding the underlying reason for LBP can facilitate its effective treatment and prevention. Many recent studies suggest an association between the morphological and composition (e.g., the fat vs. muscle ratio) changes of the muscles attaching to the spine (i.e., paraspinal muscles) and LBP [10]. Precise segmentation of these muscles plays a critical role in quantifying muscle morphology and composition and allows clinicians and researchers to detect changes, identify potential abnormalities, and tailor targeted interventions for individuals experiencing low back pain. Therefore, the study and refinement of paraspinal muscle segmentation pave the way for more effective diagnostics and personalized treatments for conditions associated with low back pain. The segmentation of paraspinal muscles is usually done using axial MRI scans of the lumbar spine at multiple disc levels (see Fig. 1.2(a) and (b)). The high variations in individual muscle morphology and composition, an example of which is shown in Fig. 1.2(c), is one of the challenges of paraspinal muscle segmentation, and it can also affect the uncertainty.

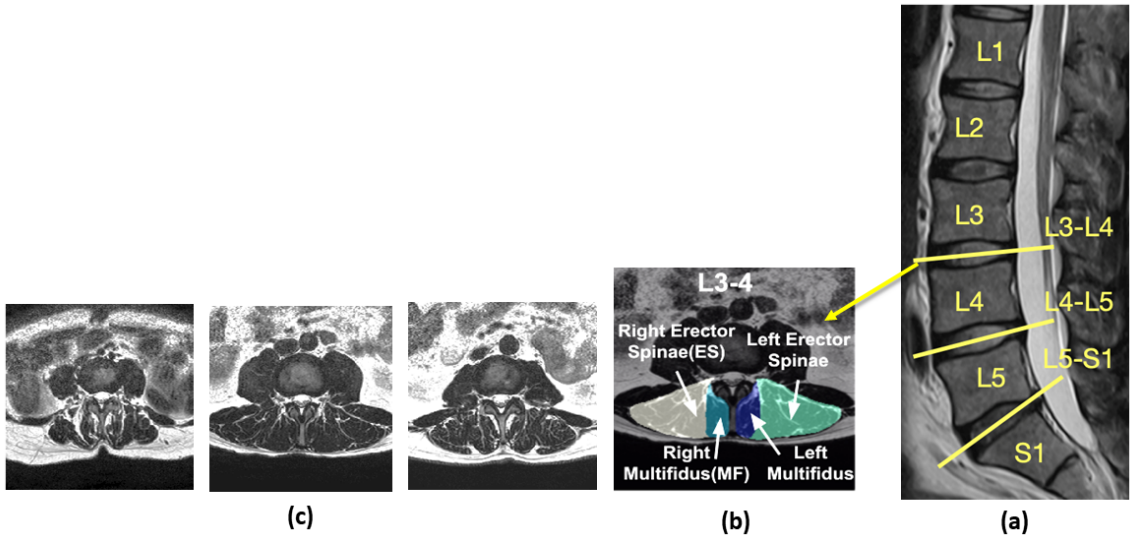

Figure 1.2: (a) An illustration of the location of disc levels L3-L4, L4-L5, and L5-S1; (b) is an example of Multifidus (MF) and Erector Spinae (ES) paraspinal muscles at level L3-L4 and (c) shows variations of paraspinal muscles in the same disc levels between individuals

1.2 A Deeper Look Into Uncertainty

As discussed earlier, understanding uncertainty has a vital role in demonstrating the limitations and reliability of DL-based model performance. Uncertainty in this context is often classified into two main classes: aleatoric and epistemic uncertainties. Epistemic uncertainty in deep learning models pertains to the lack of knowledge or uncertainty arising from limitations in the model itself, and it can be reduced by changing the model architecture, more comprehensive training, or acquiring more diverse and relevant training data. It reflects the model's ignorance about certain less-represented cases of the data and can be measured through techniques, such as model ensembling and Bayesian approaches. Mitigating epistemic uncertainty is essential for enhancing model reliability and confidence, as it involves refining the model's understanding of the relevant features of the data and minimizing uncertainties that arise from inadequate or incomplete knowledge during the training process.

On the other hand, aleatoric uncertainty in DL models refers to the unpredictability arising from the inherent randomness or noise in the data itself. Unlike epistemic uncertainty, which can be reduced with more training data or better model architecture, aleatoric uncertainty is irreducible. Understanding and quantifying aleatoric uncertainty is crucial for creating robust models that can acknowledge and appropriately respond to the variability present in real-world datasets.

While both types of uncertainties are reflected collectively in the results of the trained DL models at inference time, better understanding and quantifying these distinct forms of uncertainty is crucial to refine the interpretability and robustness of deep learning models. Specifically, the exploration into uncertainty not only provides a deep theoretical understanding of model behavior, but also establishes the groundwork for more informed decision-making, particularly in critical domains like healthcare, where the consequences of uncertainties bear significant weight.

Within the realm of DL models and their associated uncertainties, aleatoric and epistemic uncertainties stand as primary considerations. However, additional factors can also contribute to uncertainty, depending on the nature of the task at hand. One very important factor that significantly influences uncertainty is inter-rater variability, which emerges from defining the ground truths from multiple experts when differences in visual perception, domain expertise, and personal preferences

exist across raters [3]. Addressing the diversity of opinions among raters becomes particularly crucial in domains like medical tasks, where model bias towards a single expert is often undesirable, and obtaining consensus among multiple raters can help ensure the quality of ground truth labels to better guide the model training. Nevertheless, since the precise origin of inter-rater variability is still not fully understood and is difficult to quantify precisely, further research is required to unravel its precise relationship with other forms of uncertainties and distinctions. To gain insights regarding inter-rater variability and take good advantage of it, it is instrumental to determine the most effective way to present it and maximize its positive impact on enhancing the interpretability of the DL models. However, despite the importance and complexity of inter-rater variability, it has often been overlooked in the majority of studies within this field, and therefore, further exploration and investigation are indeed necessary.

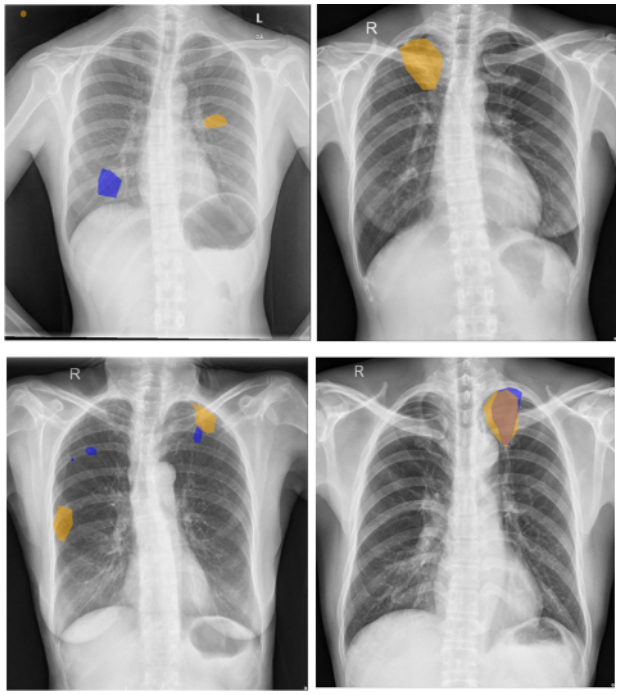

Figure 1.3: Segmentation of abnormalities in chest X-rays at different levels of agreement. Blue and yellow areas correspond to abnormalities annotated by Annotator 1 and Annotator 2, respectively. No agreement is observed in (a) and (b) while certain levels of agreement between the annotators are observed in (c) and (d). [66]

1.3 Inter-rater Variability: A Potential Cause of Uncertainty That Is Overlooked

In the context of medical image segmentation, inter-rater variability becomes evident when experts draw boundaries around structures in different ways, an example of which is shown in Fig. 1.3 for identification of abnormalities in lung X-rays. These differences in rater opinions regarding the shape of the boundaries around the structures and their spatial location within the image can significantly impact deep learning models and introduce a layer of uncertainty into the model predictions. The variation in annotations can be due to various factors, such as differences in the level of expertise among the raters or in the general annotation style that they tend to follow. In medical imaging tasks, these disagreements can also be caused by anatomical variations and image noise, which can potentially affect other types of uncertainty and result in a correlation between them. An overview of these relations is illustrated in Fig. 1.4. Similar to other types of uncertainty, recognizing and addressing inter-rater variability is essential for enhancing the robustness and reliability of deep learning models, particularly in domains where precision is crucial, such as medical imaging.

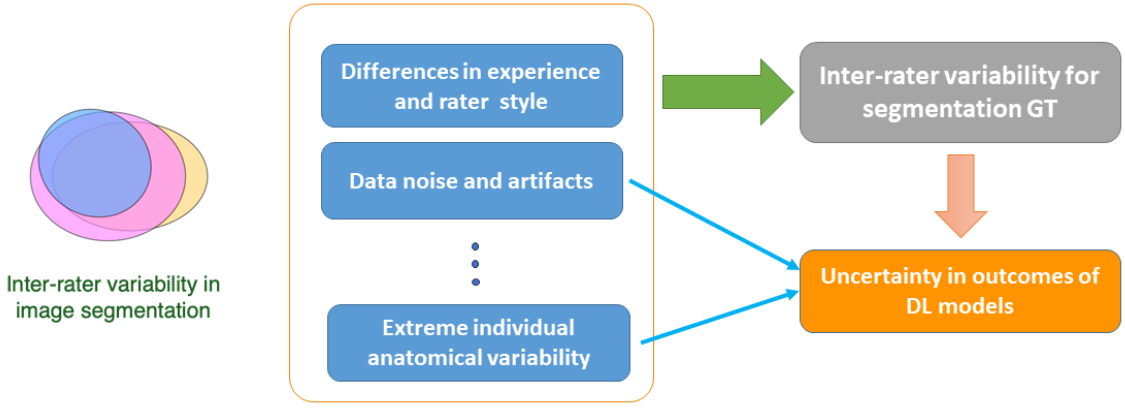

Figure 1.4: An overview of some of the factors effective in inter-rater variability and how they relate to uncertainty in model predictions.

1.4 Thesis Contribution

In summary, in this thesis, we first propose a novel multi-task TransUNet model to perform paraspinal muscle segmentation and predict the inter-rater variability, visually presented as the variance of rater annotations. Subsequently, our study extends to exploring the relationship between inter-rater variability and aleatoric/epistemic uncertainties, and the effect of model architecture and label fusion methods on them. The major contributions of our first work which is presented in Chapter 3 are as follows:

- We use the TransUNet model for paraspinal muscle segmentation for the first time.
- The proposed algorithm was trained and validated using a large dataset of paraspinal muscle MRIs with multi-rater annotations, and it has shown excellent segmentation performance in comparison to the state-of-the-art techniques and a newly proposed multi-task U-Net model.
- We proposed to use variance maps of soft labels from multiple raters to offer intuitive and easily comprehensible measures and visualization of inter-rater variability.
- Lastly and most importantly, to the best of our knowledge, we are the first to propose a multi-task DL model for joint paraspinal muscle segmentation and prediction of inter-rater variability, with the framework easily adaptable to other anatomical segmentation tasks.

Our second work, presented in Chapter 4 has the following main contributions:

- First, we are the first to compare Transformers and CNNs for their impacts on model uncertainties and the encoding of inter-rater variability, especially in a multi-class segmentation setting.
- Second, we explore the effect of label fusion methods during network training on DL model uncertainty (aleatoric and epistemic) for the first time.
- Lastly and most importantly, we explore the relationship between inter-rater variability and aleatoric/epistemic uncertainties, which has not been done so far.

1.5 Outline

The structure of this thesis is as follows. In the upcoming chapter, we offer an overview of deep segmentation models, with an emphasis on the incorporation of transformer and self-attention blocks into existing architectures. Additionally, we delve deeper into the uncertainty categories and the methodologies for their quantification. Finally, we investigate optimal approaches for label fusion in multi-rater datasets and the effective presentation of inter-rater variability. Chapter 3 introduces a novel multi-task model that is designed for simultaneous paraspinal muscle segmentation and inter-rater variability prediction. Moving on to Chapter 4, we explore the sources of inter-rater variability and how they affect the potential interplay of inter-rater variability with other uncertainty types. This section investigates how model architecture and label fusion methods influence these correlations and the preservation of inter-rater variability. The concluding chapter summarizes the thesis and considers potential avenues for future enhancements to the developed systems.

Chapter 2

Materials and Methodology

In this chapter, we start with an introduction to deep segmentation models and the advantages of the addition of transformers and attention blocks to them. Thereafter, our primary focus is on delivering a brief literature review on the uncertainty measurement techniques and explaining the methods deployed throughout our study. Subsequently, we provide insights into label fusion techniques and the formation of a consensus in multi-rater datasets.

2.1 Deep Learning-based Segmentation Models

Deep learning-based segmentation models have seen significant advancements and notable successes across diverse applications. A foundational work in this domain is the UNet architecture proposed by Ronneberger et al. [47]. UNet introduced a novel encoder-decoder structure with skip connections, facilitating the precise segmentation of medical images (see Fig. 2.1). Building upon this architecture, subsequent studies have delved into refining segmentation models for different applications [34, 44, 38].

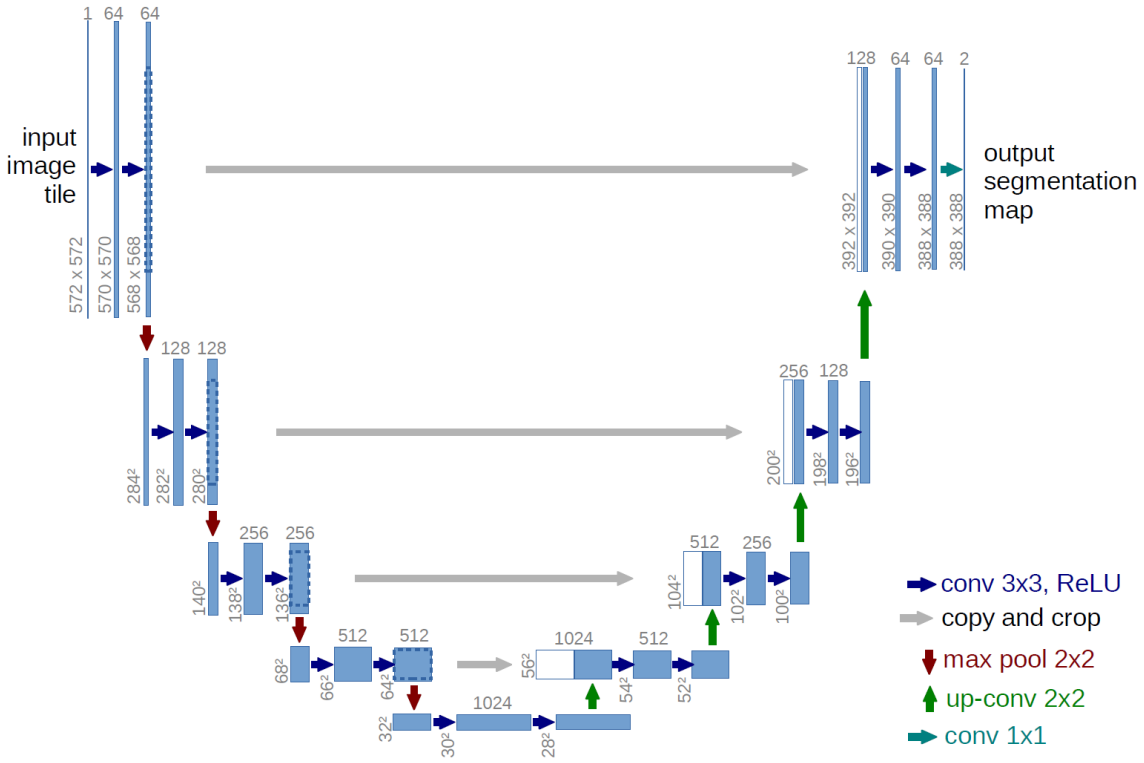

Figure 2.1: UNet model architecture [47]

Originally designed for natural language processing, transformers found their way into image classification through the introduction of the Vision Transformer (ViT) model [14]. In a ViT, each image is divided into multiple patches and then the sequence of the linear embeddings of these patches is fed to a transformer, as shown in Fig. 2.2. Compared to purely Convolutional Neural Networks (CNNs), ViTs are able to more effectively capture long-range dependencies and have outperformed CNNs in several classification tasks.

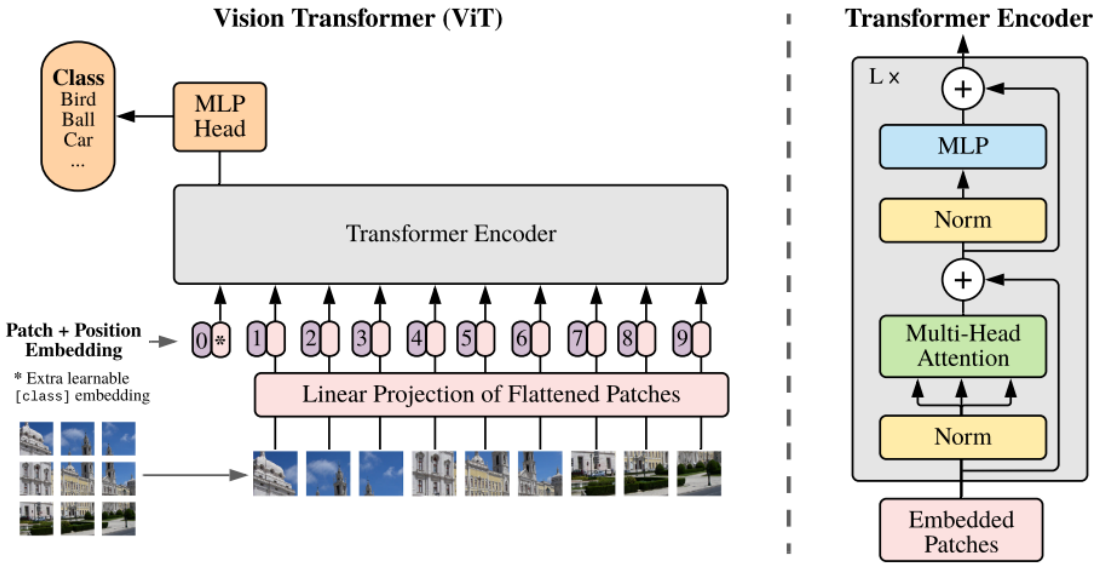

Figure 2.2: ViT model architecture [14]

After the ground-breaking introduction of UNet models to the segmentation field, the incorporation of transformer architectures has become a noteworthy advancement in segmentation models, showcasing improved performance and refined feature extraction capabilities [7, 65]. Among the proposed models that combine the encoder-decoder style of the UNet models with transformers, TransUNet [9] is a widely used one. The TransUNet model presents notable advantages over the traditional UNet architecture. Due to the integration of ViT in its encoder and also, benefiting from the encoder-decoder architecture of the UNet, TransUNet is able to capture long-range dependencies and contextual information effectively. The model architecture is shown in Fig. 2.3. This enhanced ability to consider global context contributes to improved segmentation accuracy, especially in scenarios where intricate spatial relationships and contextual understanding are crucial. The TransUNet model's attention to global context and adaptability to diverse scales enhances its robustness and generalizability, positioning it as a compelling advancement over the UNet architecture in tasks demanding subtle spatial understanding and complex feature relationships.

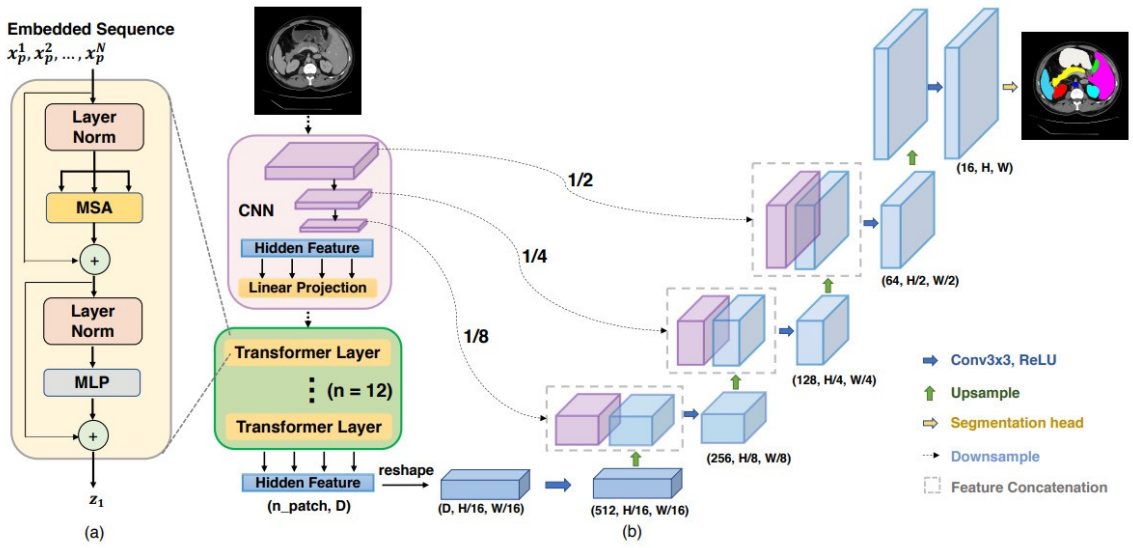

Figure 2.3: TransUNet model architecture [9]

Prior research has consistently compared models incorporating additional transformer blocks, such as TransUNet, exclusively in terms of accuracy, often focusing solely on Dice scores. However, an overlooked aspect in these studies is the exploration of the impact of these architectural enhancements on uncertainties. One of our goals in this study is to bridge this gap by investigating the influence of model architecture on uncertainty. Specifically, we conducted a comparative analysis between TransUNet and UNet to confirm whether the favorable outcomes observed in segmentation accuracy extend to a reduction in uncertainty. Additionally, we aim to determine if other factors contribute to the observed differences in the uncertainty as well.

2.2 Measuring Segmentation Accuracy

In the context of segmentation, several metrics are used to measure the accuracy of the results. Provided below is a short explanation of some of the most commonly used methods.

- **Dice Score Coefficient (DSC):** This score quantifies the degree of overlap between the predicted segmentation and the ground truth. Ranging from 0 to 1, DSC is calculated as twice the intersection of the segmented and reference regions divided by the sum of their volumes, as summarized in Eq. 1.

$$DSC = \frac{2 \times Area\ of\ overlap}{Area\ of\ prediction + Area\ of\ ground\ truth} \tag{1}$$

- **Intersection over Union (IoU):** Similar to the Dice score, the Jaccard index measures the overlap between the predicted and ground truth regions, but with a slightly different method. This metric is calculated as the area of the overlap between the segmented and reference regions divided by the area of their union and is shown in Eq. 2.

$$IOU = \frac{Area\ of\ overlap}{Area\ of\ union} \tag{2}$$

- **Hausdorff Distance (HD):** This distance metric quantifies the maximum distance between two segmentation boundaries.

- **Precision:** This metric assesses the accuracy of positive predictions and is calculated as the number of true positives (TP) divided by the sum of true positives and false positives (FP), as shown in Eq. 3.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

- **Recall:** This metric quantifies the ability of a segmentation algorithm to correctly identify positive instances. It is calculated as the ratio of true positives to the sum of true positives and false negatives (FN), as indicated by Eq. 4.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

Among the above-mentioned metrics, the Dice score is more commonly used in medical image segmentation due to its effectiveness in capturing the spatial agreement between predicted and true regions.

2.3 Uncertainty Quantification Methods

2.3.1 Bayesian Deep Learning

In a traditional deep learning model, parameters are treated as fixed values which limits the ability to calculate the uncertainty arising from both the model and the data. Bayesian deep learning combines conventional deep learning techniques with Bayesian statistical methods, introducing a probabilistic framework to the existing models. In this approach, the model parameters are treated as distributions rather than fixed values, and we try to estimate these distributions through model training. This approach of having distributions instead of fixed values helps facilitate uncertainty estimations and increases the accuracy of the models.

The concept of Bayesian deep learning and uncertainty measurement has gained significant attention in the research community, with several studies contributing to its development and application. Blundell et al. [5] introduced the "Bayes by Backprop" algorithm, which efficiently learns a probability distribution on the weights of a neural network, enabling uncertainty measurement in deep learning models. This pioneering work laid the basis for integrating Bayesian principles into deep learning, allowing for the quantification of uncertainty in model predictions.

Today, Bayesian deep learning continues to be used by researchers, serving as a method to train probabilistic models for a diverse range of both medical and non-medical tasks [12, 4, 35, 55]. However, Bayesian methods can be challenging to scale for large neural networks due to the computational cost of sampling from complex probability distributions. Therefore, estimation methods have been proposed to perform Bayesian model averaging; two widely used methods are Monte Carlo (MC) dropout and Model Ensembles.

2.3.2 Monte Carlo Dropout

Monte Carlo (MC) dropout, also known as test-time dropout (TTD), has gained popularity as a method for Bayesian approximation and uncertainty estimation within deep learning models, addressing challenges associated with pure Bayesian deep learning. Following the introduction of "Bayes by Backprop" by Blundell et al. [5], Gal and Ghahramani [17] proposed leveraging dropout

during inference as an approximation method in Bayesian neural networks, providing a robust theoretical basis for uncertainty estimation. The utilization of dropout during inference introduces randomness, by deactivating random nodes with each iteration through the model. This results in a slightly different model in each iteration, which acts similarly to sampling from the weights' distribution. Consequently, iterations with the same input yield different outputs, allowing us to interpret these variations as a means to measure uncertainty. In a medical image multi-class segmentation task, the averaging process for every voxel can be summarized as follows:

$$E(y_c) = \frac{1}{T}\sum_{t=1}^{T} P(y = c|\theta_t) \tag{5}$$

where *y* represents an image voxel, *c* is one of the possible classes present in the image, y_c is the expected probability of voxel *y* belonging to class *c*, θ_t represents the model parameters in iteration *t*, and *T* is the total number of iterations through the model with the same input, using dropout during inference.

With the incorporation of MC dropout during inference, a pivotal consideration arises: how to interpret the resulting average and determine the type of uncertainty it encapsulates. Particularly in segmentation tasks, the most widely used approaches include computing metrics such as entropy, variance, or mutual information (MI) of the outputs and utilizing them as representations of epistemic uncertainty. In a comprehensive study conducted by Camarasa et al. [6], the accuracy of these widely used epistemic uncertainty metrics was examined. The findings indicated that, overall, entropy and variance outperformed other measures, yielding uncertainty maps that more closely align with the errors present in the model outputs. A simple illustration of the process of using test-time dropout with entropy for segmentation is shown in Fig. 2.4. Using the entropy of the resulting average, the epistemic uncertainty is calculated as:

$$H(y) = \sum_{c=1}^{C} P(y = c) log(P(y = c)) \tag{6}$$

where *y* is a voxel in the image, *C* is the total number of classes, and *P(y=c)* is the probability of the voxel output belonging to class *c*.

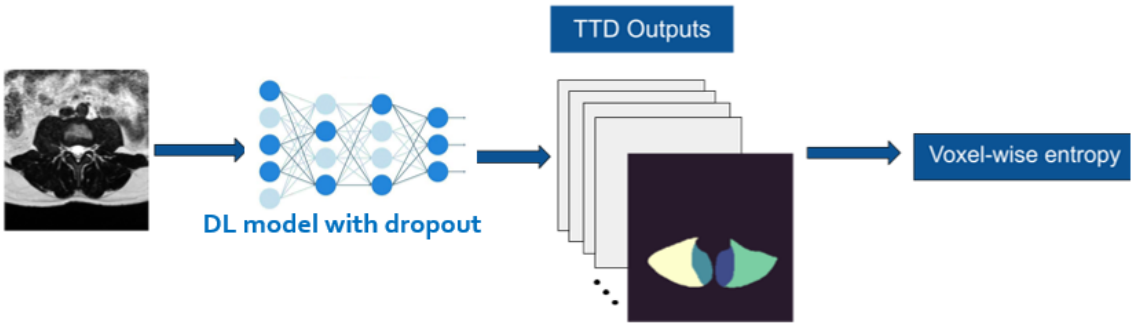

Figure 2.4: An overview of test-time dropout used for epistemic uncertainty estimation in a segmentation task

In summary, the utilization of MC dropout during test time has become a widely adopted technique among researchers, serving as an approximation method for Bayesian networks to estimate uncertainty [8, 1, 50, 2]. Nevertheless, the determination of the optimal and accurate metric for uncertainty measurements remains an open and ongoing research topic.

While certain studies have proposed methods to estimate both aleatoric and epistemic uncertainties from TTD samples [36], the majority of the studies estimate these uncertainties separately. One notable technique for aleatoric uncertainty estimation is test-time augmentation, which will be further discussed in the subsequent sections. In addition to MC dropout, another frequently employed approach for Bayesian approximation involves training an ensemble of independent models, a topic we will explain in the following section.

2.3.3 Model Ensembles

In another effort to perform Bayesian approximation and uncertainty estimation, Lakshminarayanan et al. [30] introduced a method to predict uncertainty through the use of an ensemble of independently trained models. In this method, an ensemble of neural networks is trained with shared architecture but different initialization, and at test time, by combining the predictions of multiple models, accurate uncertainty estimates can be obtained, similar to MC dropout. The method is considered scalable and straightforward, making it applicable to various deep learning architectures and tasks. Calculating uncertainty from a deep ensemble in a segmentation task is shown in Fig. 2.5. To measure the epistemic uncertainty with this averaging method, we can use Eq. 5 and 6,

similar to test time dropout.

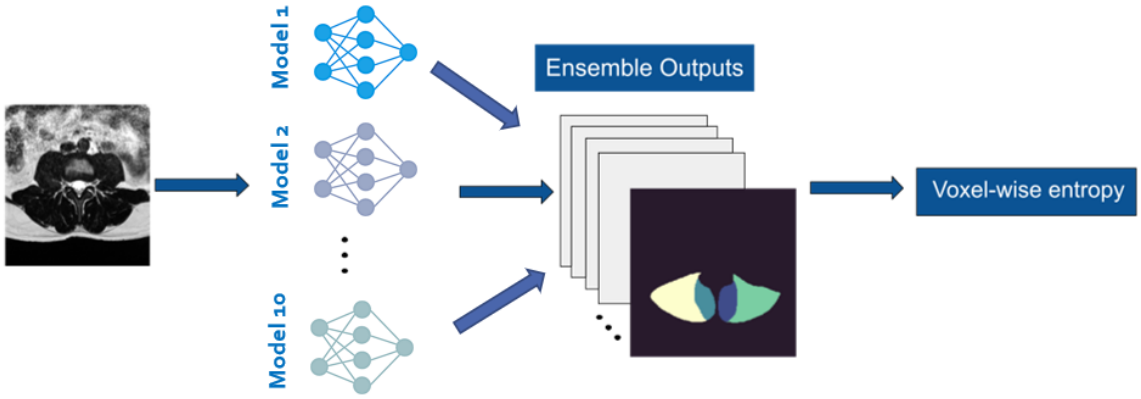

Figure 2.5: An overview of deep ensembles used for epistemic uncertainty estimation in a segmentation task

Model ensembles have become a widely adopted and effective technique for uncertainty measurement in deep learning, offering enhanced predictive performance and robust quantification of uncertainty [51, 1]. While, in general, ensembling tends to yield superior results compared to test-time dropout, it is not a universal rule and can vary based on factors such as the specific task, the model architecture, the training methods employed, and the characteristics of the dataset.

2.3.4 Test-time Augmentation

As previously discussed, aleatoric uncertainty arises from the inherent factors of the data and is affected by the variability of the dataset. Therefore, by generating multiple augmented versions of the inputs, we can simulate the variability of the data, and measure the aleatoric uncertainty as the variations of the model output when given the augmented versions of the same input data. This method is called test-time augmentation (TTA), and was used by Wang et al [59] to estimate the aleatoric uncertainty of a model trained for a medical image segmentation task. In employing test-time augmentation (TTA), augmentation techniques like geometric transformations are systematically applied to input data during the inference phase. By generating multiple augmented versions of the same input, TTA provides the model with diverse perspectives, allowing it to capture the

inherent variability in the data. Ideally, the model's output should remain consistent across all augmented versions for the same input. Consequently, any observed variability in the model's outputs can be interpreted as the uncertainty arising from the inherent variability in the data, i.e., aleatoric uncertainty. An illustration of the application procedure of TTA in a medical image segmentation task is illustrated in Fig. 2.6.

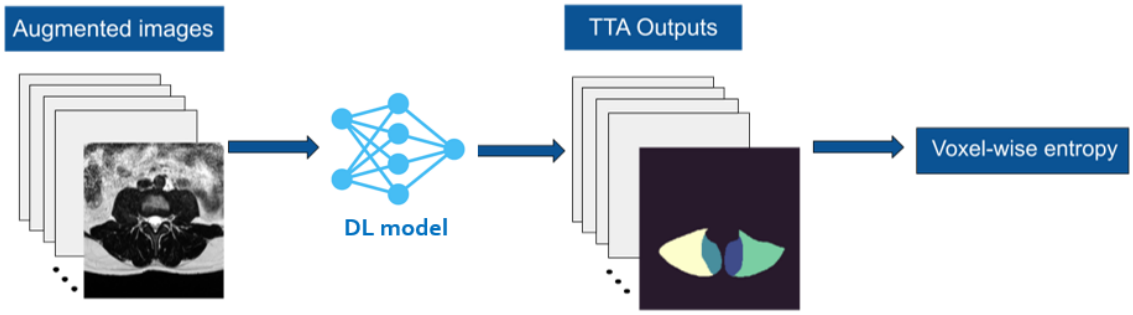

Figure 2.6: An overview of test-time augmentation used for aleatoric uncertainty estimation in a segmentation task

The model outputs for the augmented versions of the same input can be formulated in a similar fashion to the outputs of TTD or model ensembles. Therefore, the aleatoric uncertainty can be calculated using Eq. 5 and 6 as well.

2.3.5 Inter-rater Variability

In a multi-rater segmentation dataset, revealing and quantifying disagreement among raters is often achieved through the entropy of their opinions [32]. The potential dissimilarity in the ground truth annotations may signal challenging-to-segment areas within an image, indicating the need for additional attention to these areas in the model's output as well. To effectively identify and highlight regions with higher rater disagreement, the model needs to reflect such disagreement in its predictions, ideally aligning the entropy of the ground truth with that of the predictions. Consequently, the entropy of the predictions becomes a valuable representation of inter-rater variability in the model outputs. By comparing the entropies of the ground truth and the prediction, we can estimate how well a model preserves inter-rater variability through its predictions.

While employing entropy proves sufficient for quantifying inter-rater variability, alternative

methods may offer improved visual representations of these disagreements, enhancing user comprehension of variability maps. In Chapter 4, we opted for utilizing the variance of the signed distance maps, followed by a sigmoid function, to portray inter-rater variability for each image, as illustrated in Fig. 3.3. This approach yields smoother maps, enhancing visual interpretability, particularly in datasets with a limited number of annotators, which is common for pixel/voxel-wise medical image annotations.

2.4 Interpreting Inter-rater Variability: How to Find A Consensus

When dealing with multi-rater segmentation datasets, the establishment of a dependable ground truth is pivotal for training and assessing deep learning models. Various methodologies have been devised to aggregate annotations from multiple raters and derive a consensus, each presenting distinct characteristics and applications. One straightforward method involves employing the majority vote, where the label most commonly agreed upon among the raters is chosen as the ground truth. This simplistic yet robust method is effective, especially when a clear majority consensus exists. It can also efficiently be implemented for multi-class segmentation datasets. Alternatively, the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm includes refining the estimation of true segmentation by iteratively considering observed segmentations and the estimated performance levels of each rater [61]. This algorithm is particularly valuable when dealing with diverse levels of expertise among raters.

While majority vote and STAPLE have been widely used in classic and DL-based image segmentation methods, for DL models, random sampling involves randomly selecting one of the raters' annotations as the ground truth during each epoch [24, 32] is another technique to allow the integration of multi-rater opinions. Furthermore, this approach is shown to generate better-calibrated outputs and uncertainty maps.

Each of these approaches possesses unique merits and should be selected based on the specific characteristics of the dataset, the nature of the annotations, and the desired properties of the derived ground truth for subsequent model training and evaluation in deep learning tasks.

2.5 Quantification of Inter-Rater Variability

As we mentioned earlier (section 2.3.5), inter-rater variability can be quantified and visualized through the entropy or the variance of the rater annotations. In an ideal model, the entropy of the model predictions for each voxel should be representative of the underlying inter-rater variability in the ground truths. For example, in a 3-class segmentation task with 3 different raters, if the average ground truth for a specific voxel is [0.33, 0.33, 0.33] (i.e., each rater has placed this voxel in a different class), then ideally we want to see model predictions that are close to this ground truth. Therefore, one way to measure the inter-rater variability preservation is to compare the average ground truths and the probabilistic outputs. One of the metrics used for this comparison is the Brier score. Inter-rater variability preservation and the Brier score are explained in detail in Chapter 4.

Chapter 3

Joint Paraspinal Muscle Segmentation and Inter-rater Labeling Variability Prediction with Multi-task TransUNet

A version of this chapter was presented at the 25th International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2022 during the Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE) workshop.

- **Roshanzamir, P.**, Rivaz, H., Ahn, J., Mirza, H., Naghdi, N., Anstruther, M., Battié, M.C., Fortin, M. and Xiao, Y., 2022, September. Joint paraspinal muscle segmentation and inter-rater labeling variability prediction with multi-task transunet. In International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (pp. 125-134). Cham: Springer Nature Switzerland. [48]

3.1 Introduction

Paraspinal muscles are critical to stabilizing the spinal column [60] and many recent studies [16, 64] have suggested a link between the morphological and composition (e.g., fatty infiltration) changes of these muscles and painful spinal disorders that lead to low back pain (LBP), such as disc herniation and lumbar spinal stenosis (LSS), with a high prevalence in the adult population worldwide. With superior soft tissue contrast, magnetic resonance imaging (MRI) has become a staple for the relevant investigations, with the new trend to leverage recent computational and imaging techniques to explore image-based biomarkers for more accurate diagnosis and prognosis. To achieve this, good MRI segmentation of paraspinal muscles, especially at different spinal levels plays a critical role. While manual image segmentation is known to be time- and labor-consuming, the high variations in muscle morphology and composition across individuals and spinal levels create additional challenges for segmentation in comparison to other anatomical structures such as the brain. While automatic segmentation algorithms, especially deep learning (DL) methods [44, 52] have been shown to mitigate these issues, only a few were proposed for paraspinal muscle segmentation. Furthermore, most DL methods only offer deterministic outcomes, making it impossible to assess the reliability of their results, which is important in practice. Different from epistemic and aleatoric uncertainties that are used to evaluate the reliability of automatic segmentation [15, 41], methods that predict and visualize inter-rater disagreement in DL-based segmentation, which also offer important insights regarding the reliability of the results, are unfortunately often overlooked. Thus, new techniques are necessary to address this gap.

Automatic paraspinal muscle segmentation has its unique challenges, and only a few techniques [23, 33, 63] were proposed, focusing primarily on deep learning techniques using a single rater's annotation. Earlier, Li *et al.* [33] proposed a U-Net based model with additional residual blocks at each layer and a feature pyramid attention module added to the bottleneck. They obtained a Dice score coefficient (DSC) of 94.9% in multifidus segmentation at three spinal levels. Xia *et al.* [63] combined conditional random fields as recurrent neural networks with a U-Net model to add spatial constraints for labels and achieved a DSC of 95% on the segmentation of the multifidus and erector spinae. Most recently, Huang *et al.* [23] proposed a two-stage coarse-to-fine segmentation

framework with attention gates and achieved the best DSC of 94.4±3.5% on multifidus identification. However, none of the relevant existing works explored the prediction and visualization of inter-rater segmentation variabilities, and all used convolutional neural networks (CNNs). While advantageous in extracting low-level image features, CNNs often fail to capture long-range dependencies in the input data, which can be leveraged to improve segmentation performance, especially in the context of high anatomical variabilities. To address this concern, attention gates have been employed to improve U-Net [44] Most recently, exploiting self-attention and long-range dependencies, Transformers, which were first proposed in machine translation [57] also lend their powers in vision tasks [14] To benefit from both CNNs and Vision Transformers (ViT), TransUNet [9] is a recent encoder-decoder-style model that combines a U-Net with a Transformer. This DL model can capture long-range dependencies while extracting the high-resolution features captured by a CNN. To the best of our knowledge, it has not been used for paraspinal muscle segmentation.

To enhance the quality of segmentation ground truths, it is often desirable to employ multiple raters, and a number of techniques have been proposed to explore inter-rater variability. Mirikharaji *et al.* [39] used an ensemble of DL models for lesion segmentation, with each trained on a single rater's annotation, and averaged the results to produce the final output. Ji *et al.* [25] proposed a CNN with embedded modules for encoding raters' expertise levels, generating one prediction per rater and then fusing the predictions based on their uncertainties. Finally, Lemay *et al*. [32] explored the advantages of soft segmentation for training segmentation models to reflect inter-rater variability. They concluded that relying solely on discrete labels as outputs will generate overconfident results and, in these cases, soft segmentation performs better. These prior works often train an ensemble of DL models or use a large model with many learnable parameters, which could be costly. Furthermore, they employed soft or probabilistic segmentation through averaging to represent the inter-rater annotation variability, which may not be visually intuitive or informative.

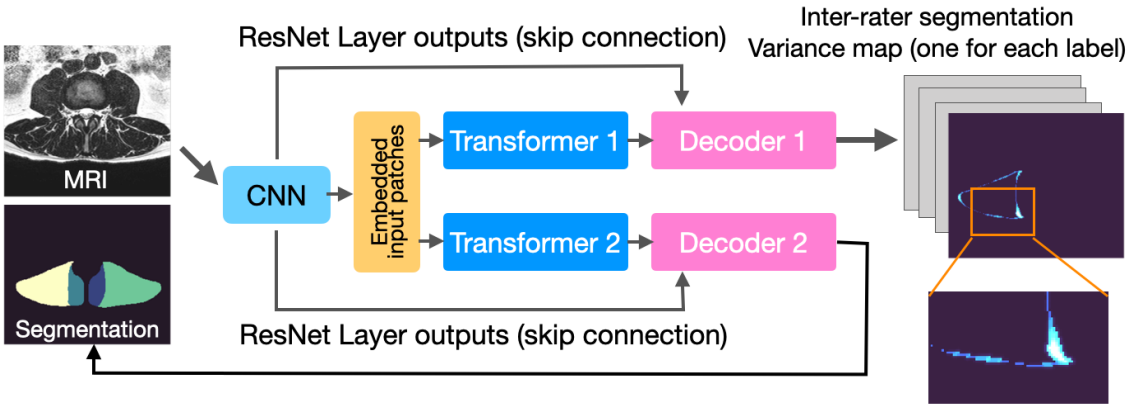

Figure 3.1: An overview of the proposed multi-task TransUNet

In this paper, we proposed a novel multi-task deep learning technique (see Fig. 3.1) to address the aforementioned issues and allow accurate segmentation of paraspinal muscles (i.e., multifidus muscles and erector spinae, left and right separated as shown in Fig. 3.2) and prediction and visualization of inter-rater segmentation inconsistency. As a design choice, we require the representation of variability across multiple raters to allow efficient training through deep learning models and be easily interpreted. Therefore, instead of averaging multiple segmentations to reflect the inconsistency, we decided to estimate the variance across multi-rater segmentations. Thus, our final algorithm consists of a multi-task TransUNet with shared convolutional layers to produce the patch embeddings from the input images, which are then fed into two task-specific transformers to produce the desired outputs. In Task 1, the features are decoded and up-sampled to produce the segmentations, and a similar decoding process is repeated for Task 2 to produce the pixel-wise variance map of multi-rater annotations.

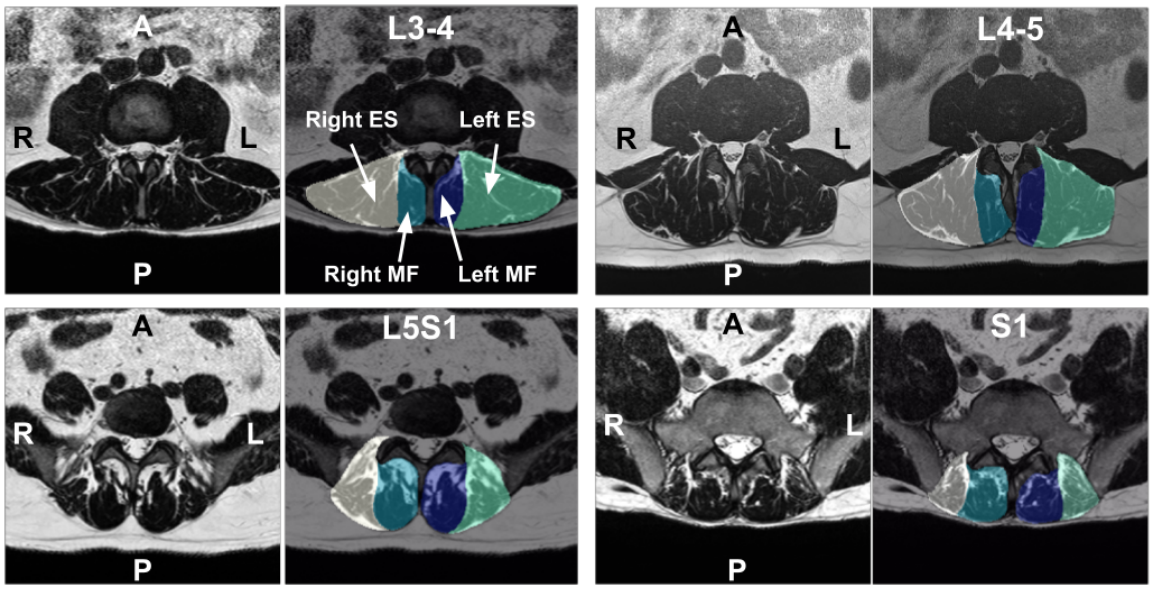

Figure 3.2: Axial MRIs of paraspinal muscles at four spinal levels (L3-L4, L4-L5, L5-S1, and S1). The names of the four muscles, the left and right multifidus (MF) and erector spinae (ES) muscles, along with their color-coded manual segmentations are shown in the top left image.

Our work has four major contributions: First, we use the TransUNet model for paraspinal muscle segmentation for the first time. Second, the proposed algorithm was trained and validated using a large dataset of paraspinal muscle MRIs with multi-rater annotations, and it has shown excellent segmentation performance in comparison to the state-of-the-art techniques and a newly proposed multi-task U-Net model. Third, we proposed to use variance maps of soft labels from multiple raters to offer intuitive and easily comprehensible measures and visualization of inter-rater variability. Lastly and most importantly, to the best of our knowledge, we are the first to propose a multi-task DL model for joint paraspinal muscle segmentation and prediction of inter-rater variability, with the framework easily adaptable to other anatomical segmentation tasks.

3.2 Materials and Methodology

3.2.1 Image Pre-processing and Multi-rater Annotation

From the European research consortium project, Genodisc, on commonly diagnosed lumbar pathologies (physiol.ox.ac.uk/genodisc), lumbosacral T2-weighted (T2w) MR images of 118 patients (59 male, age=30$\sim$59y) were selected, with the factors of sex and age roughly equally distributed among the subjects. Axial MRI slices of the L3-L4, L4-L5, L5-S1, and S1 spinal levels that are often affected by painful spinal disorders were acquired for analysis. In total, we have 444 MRI slices, including 105 scans at L3-L4, 117 scans at L4-L5, 118 scans at L5-S1, and finally 104 scans at S1. Note that due to imaging artifacts and cropping, not all patients have usable axial slices at all spinal levels. All axial MR images were first processed with N4 inhomogeneity correction [56] to remove field non-uniformity in the image. Then, the multifidus (MF) and erector spinae (ES) muscles were manually segmented for all patients (using the software ITK-SNAP (itksnap.org)) independently by three different raters, who have good knowledge in musculoskeletal anatomy and ITK-SNAP. As all raters have similar levels of expertise and experience in paraspinal muscle segmentation, we decided to use a majority voting scheme to fuse multi-rater annotations, and the final results were used for training and testing the proposed algorithm in terms of discrete anatomical segmentation. To produce the variance maps of multi-rater segmentations for inter-rater variability, instead of directly using discrete labels, we decided to first generate soft label maps by following the steps in [46] . In short, for a multi-class segmentation (4 classes in our case), each class was first binarized, and then transformed into a signed distance map. Then, a sigmoid function was applied to convert the signed distance map into a soft probabilistic label, and the variance across three raters was computed in a pixel-by-pixel manner. Thus, for each image, there are four variance maps, one for each muscle group. The steps for generating the variance map are shown in Fig. 3.3.

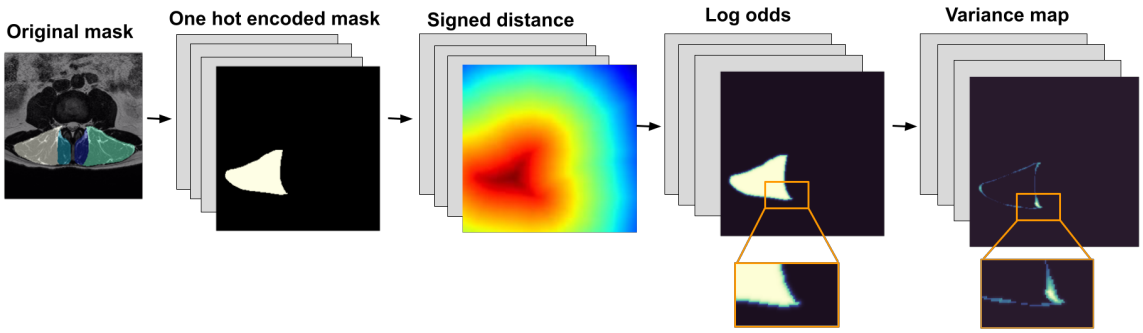

Figure 3.3: The steps of calculating the variance maps for assessing inter-rater variability.

3.2.2 Model Architecture

In this section, we provide a detailed description of the proposed TransUNet model which consists of 3 main parts: (1) The convolutional network (CNN) (2) Transformer and (3) Decoder. The goal of the CNN is to encode important image features and produce the most relevant flattened patch embeddings that can be fed to the task-specific transformers. In our model, the CNN has a ResNet34 backbone with 3 skip connections that feed the ResNet layer outputs to the decoders. In summary, the CNN produces N flattened $P \times P$ patches for each image of dimensions $C \times H \times W$ where $N = (HW)/P^2$ and C is the number of channels. As the encoded information is usable in both segmentation and variance map prediction, the two tasks in our model share the CNN layers. Next, the patches are simultaneously fed into two transformers. Similar to the encoder-decoder architecture of U-Net, the output of each transformer then goes through an up-sampler, where the task-specific information is decoded with the help of the feature maps provided by the skip connections from the ResNet. Finally, the discrete muscle segmentation is generated, as well as a 4-channel output, with each channel corresponding to the variance of the rater annotations of one muscle (left and right multifidus & erector spinae). The schematic of the proposed model is shown in Fig. 3.1.

As previously mentioned, we trained our model in a multi-task manner. The cost function for model training is a weighted sum of the losses of the individual tasks. For segmentation, we use a combination of Dice and cross-entropy loss and for variance estimation, we use the mean squared error (MSE). In summary, the total model loss is:

$$L_{model} = 0.4L_{DiceCE} + 0.6L_{MSE} \tag{7}$$

where L_{MSE} is the mean squared loss of Task 2 and L_{DiceCE} is the segmentation loss which is a weighted sum of Dice loss and cross-entropy:

$$L_{DiceCE} = 0.4L_{Dice} + 0.6L_{CE} \tag{8}$$

As TransUNet performs better in terms of generalization and capturing the global context, it can be more resistant to overfitting compared to other models like U-Net. This is especially helpful for both muscle segmentation and variance map estimation.

3.3 Experiments and Results

3.3.1 Experimental Set-up and Implementation Details

To better demonstrate the benefits of combining CNN and transformer architectures, besides the proposed multi-task TransUNet, we also devised a multi-task U-Net to perform the same joint tasks. The accuracy of image segmentation and variance prediction were assessed with DSC and MSE, respectively. Two-sided paired-sample t-tests were performed to compare the performance of the two proposed techniques, which were trained and tested in the same manner. More specifically, from the acquired 444 MRI slices, 20% of them are randomly sampled as the test set to report the algorithms' performance. For the remaining data, 80% and 20% served as training and validation sets, respectively. To improve the robustness of the networks, data augmentation was performed for the training data, where we applied random rotation, image mirroring (label IDs were also swapped accordingly), Gaussian noise, and Gaussian blurring, resulting in 1420 MRI slices in total. Furthermore, in each of the train, validation, and testing sets, there are approximately the same proportion of images from each spinal level. Finally, all MRI scans, discrete manual segmentation, and variance maps are resized to 256 × 256 pixels for network training. We trained the proposed multi-task DL models on an Alienware Aurora PC with Intel(R) Core(TM) i7-8700 CPU and 12 GB NVIDIA TITAN V GPU for 150 epochs, with a batch size of 2 and stochastic gradient descent

(SGD) optimization. The initial learning rate was 0.00125, and it was decreased gradually after each iteration. As mentioned in the previous section, a loss function (Eq. 7 & 8) that integrates cross-entropy and Dice losses, as well as MSE was used.

3.3.2 Quantitative and Qualitative Results

The quantitative assessments of the two proposed techniques (multi-task TransUNet and U-Net), as well as two recent paraspinal segmentation methods [23, 63] are listed in Table 3.1 for both the paraspinal muscle segmentation and the associated multi-rater variance map prediction. As shown in Table 3.1, the Dice scores of the proposed multi-task TransUNet are higher than the U-Net counterpart for 3 of the 4 muscles ($p<0.01$). Compared with previous reports [23, 33, 63], the proposed TransUNet has better or nearly similar accuracy in automatic identification of paraspinal muscles on average. For inter-rater segmentation variance map estimation, the multi-task TransUNet outperforms the U-Net for the left multifidus ($p<0.01$) and has lower mean errors for the left and right erector spinae muscles. To further demonstrate the results qualitatively, the outcomes of segmentation and variance prediction for one subject are shown in Fig 3.4 and 3.5, respectively. For the segmentation, we can see that TransUNet provides smoother tissue boundaries without "island labels" (see L5-S1 level segmentation of Fig. 3.4) produced in U-Net segmentations. Furthermore, U-Net produces variance maps with overall higher errors within the target muscle and background than the multi-task TransUNet.

Table 3.1: Quantitative evaluation of automatic segmentation (in DSC) and inter-rater labeling variance prediction (in MSE) for the proposed techniques. Here, superior performance of Multi-task TransUnet than the Multi-task U-Net is indicated by $^*(p < 0.01)$. L represents Left and R is Right.

Metric	Task 1 – segmentation (DSC (%))				Task 2 – variance prediction (MSE×10^5)			
	L_{MF}	R_{MF}	L_{ES}	R_{ES}	L_{MF}	R_{MF}	L_{ES}	R_{ES}
Multi-task U-Net	93.3±4.5	93.6±3.7	92.4±3.9	91.8±4.5	7.37±10.60	5.37±6.67	11.20±16.54	9.81±18.02
Multi-task TransUNet	*94.3±4.1	*94.5±3.4	92.8±4.5	*92.8±4.1	*6.41±9.62	5.57±8.62	10.83±16.03	9.66±17.87
Xia et al. (2019) [63]	95.0±2.3	94.5±1.9	90.6±5.8	91.3 ±4.6	N/A	N/A	N/A	N/A
Huang et al. (2020) [23]	94.4±3.5	94.9±3.0	94.4±3.4	94.4±3.4	N/A	N/A	N/A	N/A

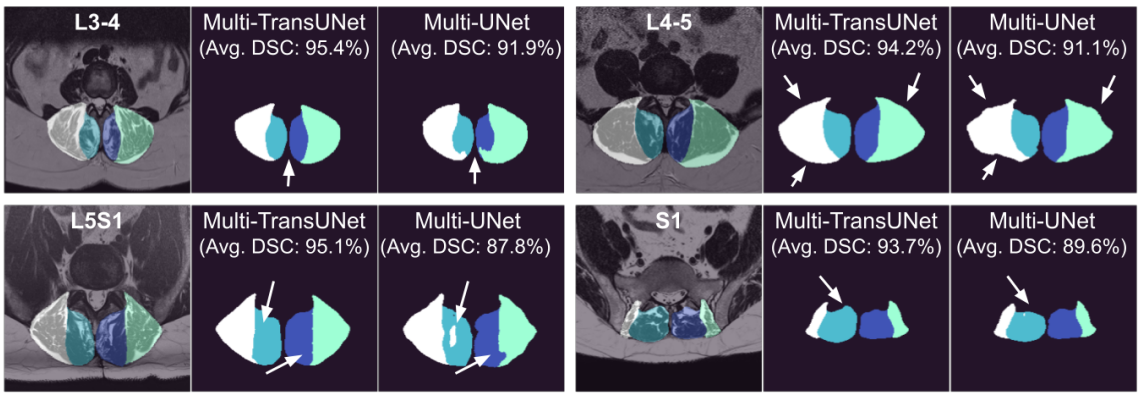

Figure 3.4: Automatic paraspinal muscle segmentation results at different spinal levels of a LBP patient, with arrows indicating the differences between results from the TransUNet and U-Net.

3.4 Discussion

With both quantitative and qualitative assessments, we can see that it is advantageous to combine CNN and Transformer architectures in comparison to using CNNs alone. More specifically, TransUNet offers smoother tissue boundaries in segmentation while nicely handles individual anatomical variabilities (see S1 level segmentation in Fig. 3.4). On the other hand, the U-Net model can introduce island labels and rough tissue borders. Previous approaches leverage additional conditional random fields as a post-processing step [63] or devise more complex hierarchical processing pipelines [23]. The capacity to encode long-range spatial information while extracting local features allows the proposed technique to achieve similar or better outcomes for tissue labeling with a simpler setup. This benefit also extends to the estimation of inter-rater annotation variability. Compared with the U-Net, the proposed method has significantly lower prediction error for the left multifidus and lower mean errors for the erector spinae muscles. In general, higher rater disagreements usually occur at the borders of the muscles, especially at the borders of muscles vs. bones, the posterior borders of erector spinae, and the posterior division between erector spinae and multifidus. This is because these regions have higher individual variability. For both proposed multi-task TransUNet and U-Net models, the mean errors in the predicted variance maps are higher for the erector spinae than the multifidus, likely because the anatomical variabilities are greater. For the same reason, the automatic segmentation for the erector spinae also has lower Dice scores than that of the multifidus.

For this study, multi-task learning allows us to estimate inter-rater disagreement without the need for model ensembles or other complex additional modules. Previous studies [29] also showed that multi-task learning is beneficial to enhance the training and performance of the DL tasks involved. The inter-rater variance maps have low pixel values that lead to small MSEs in the loss function, making training challenging for Task 2. We solved this by scaling the metric to a greater range (multiplying the values by 70) in the training process and received good results. The outputs are re-scaled to their original range at test time.

There are still several aspects of the proposed framework that can be improved in the future. First, more data will be incorporated to further enhance the performance against anatomical variabilities due to dividual differences and diseases. Second, we will further investigate the composition of different loss functions and the weights involved to help improve the performance. Lastly, we will experiment with different set-ups for the main architecture to verify their impacts on the accuracy of the desired tasks.

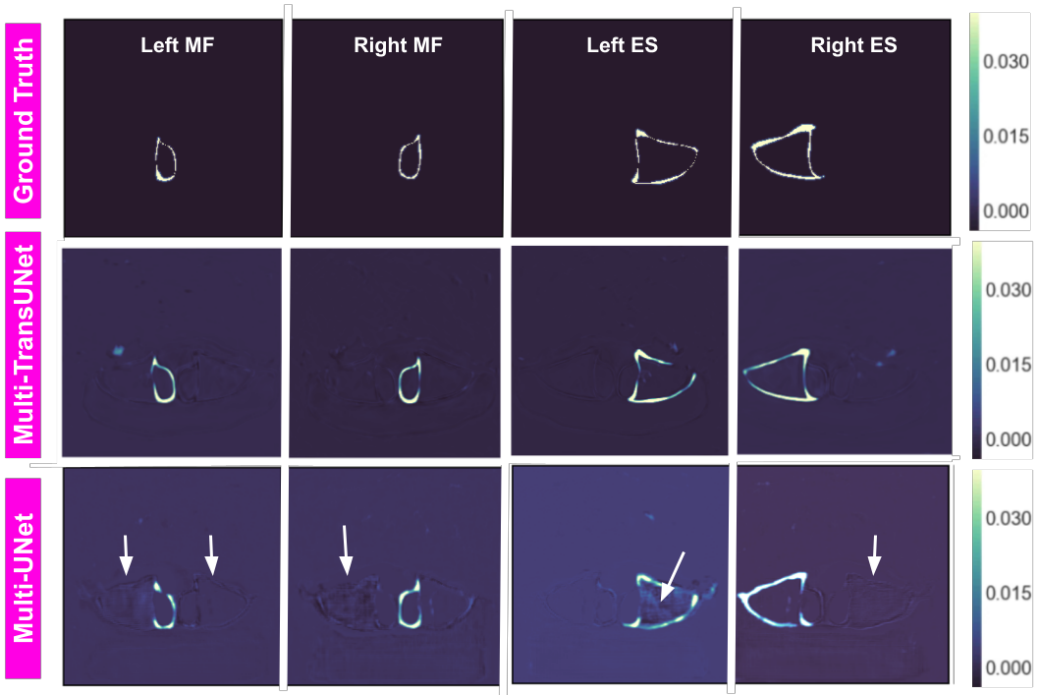

Figure 3.5: Results of the variance map estimation of one image at the L4-L5 level with different models. The arrows point to the areas with errors in inter-rater variance map prediction. The overall errors from multi-task U-Net are higher than those from multi-task TransUNet.

3.5 Conclusion

In this paper, we proposed a novel multi-task TransUNet for simultaneous paraspinal muscle segmentation at multiple spinal levels and prediction of the variance from multiple raters' annotations. While demonstrating the benefit of combining CNN and transformer architectures for the target tasks against the popular U-Net, the proposed technique offers similar or better segmentation accuracy than previous works. The resulting framework offers user-friendly and complementary information in addition to conventional uncertainty estimation, and can be easily extended to other segmentation tasks.

Chapter 4

How Inter-rater Variability Relates to Aleatoric and Epistemic Uncertainty: A Case Study with Deep Learning-based Paraspinal Muscle Segmentation

A version of this chapter was presented at the 26th International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 during the Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE) workshop.

- **Roshanzamir, P.**, Rivaz, H., Ahn, J., Mirza, H., Naghdi, N., Anstruther, M., Battié, M.C., Fortin, M. and Xiao, Y., 2023, October. How inter-rater variability relates to aleatoric and epistemic uncertainty: a case study with deep learning-based paraspinal muscle segmentation. In International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (pp. 74-83). Cham: Springer Nature Switzerland. [49]

4.1 Introduction

In recent years, deep learning (DL) techniques have shown remarkable success in various fields, including medical domains. However, using DL models in safety-critical applications, such as medical diagnosis and treatment, requires not only high accuracy but also a proper understanding of the model's uncertainty, which is crucial for the safety and adaptability of medical DL algorithms. In general, a DL model's uncertainty can be classified into two main categories: epistemic and aleatoric [13]. Epistemic uncertainty is related to the model's lack of knowledge about the data and can be reduced by collecting more data or optimizing the model's architecture and training process. On the other hand, aleatoric uncertainty is related to inherent factors of the data (e.g., noise) and cannot be reduced. Additionally, in image segmentation tasks, where supervised learning is widely used, another important factor that can affect uncertainty is inter-rater variability. Supervised segmentation algorithms require training data with well-annotated ground truth (GT) masks. While GT masks obtained by fusing annotations of multiple experts are commonly recommended, it is still costly and the best practice to combine different annotations is still being explored. Various factors can affect inter-rater variability in manual medical image segmentation, including differences in expertise, rater style [58], image noise, extreme anatomical variations among individuals, and so on. In turn, inter-rater variability propagates the influences of these factors to the resulting DL models through training as uncertainties of the algorithms. For example, image noise and measurement errors (e.g., due to partial volume effects) can result in aleatoric uncertainty while extreme individual anatomical variations, which may not be sufficiently represented in the data, can contribute to epistemic uncertainty. Therefore, knowledge regarding the relationship of inter-rater variability with aleatoric and epistemic uncertainties in DL models can help better understand their performance and reliability and inform the dataset design and learning strategies to improve them.

To date, various methods have been proposed to measure aleatoric and epistemic uncertainties of DL models. In terms of epistemic uncertainty, Bayesian DL estimates a distribution for each weight in the network and uses these distributions to measure epistemic uncertainty [28, 62]. As Bayesian networks can bear high computational costs, more efficient approaches have been reported for uncertainty estimation. Gal and Ghahramani [18] proposed dropout at test-time to approximate

Bayesian neural networks for uncertainty estimation while deep ensemble trained multiple versions of the same DL model to derive epistemic uncertainty [30]. Finally, variational inference [26] has also been used but limits the types of DL models in application. Different approaches have been reported ever since to improve the quality of the uncertainties obtained from these methods [31]. Another important aspect in accurate uncertainty quantification is the metric. Camarasa et al. [6] performed an extensive study on different metrics for measuring epistemic uncertainty using test-time dropout (TTD). They concluded that the measure of entropy produces uncertainty maps that are in correspondence with the misclassification in the model. Utilizing a sampling strategy similar to TTD, Wang et al. [59] proposed test-time augmentation (TTA) for aleatoric uncertainty assessment, which samples from the data distribution by using input data augmentation at inference time.

In training data, inter-rater variability can be measured as the entropy of the rater annotations. Lemay et al. [32] showed the superiority of random sampling and STAPLE in inter-rater variability preservation and image segmentation accuracy with a UNet. Jensen et al. [24] discovered that random sampling leads to higher classification accuracy and better-calibrated results. Vincent et al. [58] characterized rater style in terms of bias and variance of raters' annotations and explored the relationship between rater bias and data uncertainty. Nichyporuk et al. [42] proposed a segmentation model that can learn the bias in the annotations for better results. However, previous studies haven't explored the impact of label fusion methods on aleatoric and epistemic uncertainties, or the potential relationship between inter-rater variability and these uncertainties. In addition, most of them only used a conventional UNet as the base model. In recent years, Transformers that better model long-range dependencies via self-attention have gained popularity in vision tasks [14], and the hybrid Transformer-convolutional neural network (CNN) models that complement their merits, such as TransUNet [9], have shown better segmentation accuracy [48]. With a different mechanism, the addition of Transformers can have potential effects on the uncertainty of a segmentation model, which has not been investigated, but can be highly valuable.

To address the aforementioned knowledge gaps, In this study, we investigate inter-rater variability in relation to DL model uncertainties using commonly employed TTA, TTD, and deep ensemble techniques for MRI-based paraspinal muscle segmentation, with a comparison of UNet and TransUNet. Our work has three main novel contributions: First, we are the first to compare

Transformers and CNNs for their impacts on model uncertainties and the encoding of inter-rater variability, especially in a multi-class segmentation setting. Second, we explore the effect of label fusion methods during network training on DL model uncertainty (aleatoric and epistemic) for the first time. Lastly and most importantly, we explore the relationship between inter-rater variability and aleatoric/epistemic uncertainties, which has not been done so far. We hope the article will offer instrumental insights to facilitate the design and selection of DL datasets, training strategies, and model architectures.

4.2 Materials and Methodology

4.2.1 Inter-rater Variability

To measure inter-rater variability, we use the entropy of the average GT annotations for each image [32]. The pixel-wise entropy is calculated as:

$$H(y_i) = -\sum_{c=1}^{C} P(y_i = c) \log(P(y_i)) \tag{9}$$

where C is the number of classes, and y_i is the GT annotation at voxel *i,* obtained by simply averaging the one-hot GT masks from all raters. Similarly, the entropy of the model predictions can be calculated as the entropy of the softmax layer outputs before binarization. This entropy can be considered as the prediction inter-rater variability. *An ideal model should preserve the inter-rater variability during inference and produce prediction entropies similar to the GT entropy.* To quantify the perseverance of inter-rater variability in a model, we calculated the class-wise Brier score [32] as:

$$BrierScore = \frac{1}{N_{image}} \sum_{k=1}^{N_{image}} (\frac{1}{N_{voxel}} \sum_{i=1}^{N_{voxel}} (y_{i,k} - \hat{y_{i,k}})^2) \tag{10}$$

where N_{image} is the total number of images in the test set, N_{voxel} is the number of voxels in each image, $y_{i,k}$ is the average GT, and $\hat{y_{i,k}}$ is the prediction for voxel *i* of image *k* (the softmax probability). A Brier score close to zero indicates perfect preservation of inter-rater variability in model predictions [32].

4.2.2 Aleatoric and Epistemic Uncertainty Assessment

In this study, we compare TTD and deep ensemble for epistemic uncertainty assessment, and use TTA for measuring aleatoric uncertainty for both the UNet and TransUNet. In each experiment, we acquire 10 samples. In TTA and TTD, these samples are obtained through 10 forward passes through the model for each image, and in deep ensembles, each image is fed to 10 independently trained models of the same architecture. The obtained samples are then averaged to produce a final prediction for each image. This prediction is then used to calculate the entropy based on 9. To compare the quality of the resulting uncertainty maps of TTD and deep ensembles, we measure the association between uncertainties and misclassifications, using the framework provided by Mobiny et al. [40]. Conventionally, in a confusion matrix the term "positive" is used for "capturing a target label/class", but here, we define it as "capture a high uncertainty". An ideal model should be uncertain only when making a wrong prediction. Thus, a "false positive" means that the model is highly uncertain about a correct classification. Identifying a voxel as "uncertain" requires thresholding an uncertainty map. We perform the operation at multiple values, calculate the corresponding precision and recall metrics, and finally use the area under the precision-recall curve (AUC-PR) as an indicator of the quality of an uncertainty map.

4.2.3 Network Architectures and Label Fusion

Lemay et al [32] showed that the method used for label fusion in a multi-rater dataset can affect the model calibration and how well it preserves inter-rater variability. In this study, we used two different methods for integrating multi-rater annotations into our training framework: 1) The majority vote of all raters is used as the GT for each image; 2) The annotation of one rater is randomly selected at each epoch during training (refer to as random sampling). In order to explore the impact of self-attention modules on uncertainty, we train two sets of models, using TransUNet and UNet as the base model architectures. The TransUNet model has four upsampling layers with two convolution blocks at each layer. For a fair comparison, we used a UNet model with the same number of layers as the TransUNet (4 layers). With these two architectures and two label fusion methods, we trained four models for TTA and TTD-related analysis. To achieve deep ensembles for

measuring epistemic uncertainty, we also trained a set of 10 UNets and a set of 10 TransUNets with majority vote GT.

4.2.4 Dataset

Our dataset consists of a total of 673 lumbosacral T2-weighted (T2w) MR images of 119 patients (59 male, age=30∼59y) from the European research consortium project, Genodisc, on commonly diagnosed lumbar pathologies. The subjects were selected with the factors of sex and age roughly equally distributed. Our study was approved by local research ethics board. The MRI scans are from 6 different disc levels. However, due to imaging artifacts and cropping, not all patients have usable axial slices at all spinal levels. All axial MR images were processed with non-local means denoising [11] and N4 inhomogeneity correction [56] to improve image quality. Then, the left and right multifidus (MF) and erector spinae (ES) muscles were manually segmented for all patients independently by three different raters to study low back pain [64], resulting in four segmentation classes for each image (see Fig. 4.1). All raters had two training sessions to ensure the quality and consistent protocol of the segmentation.

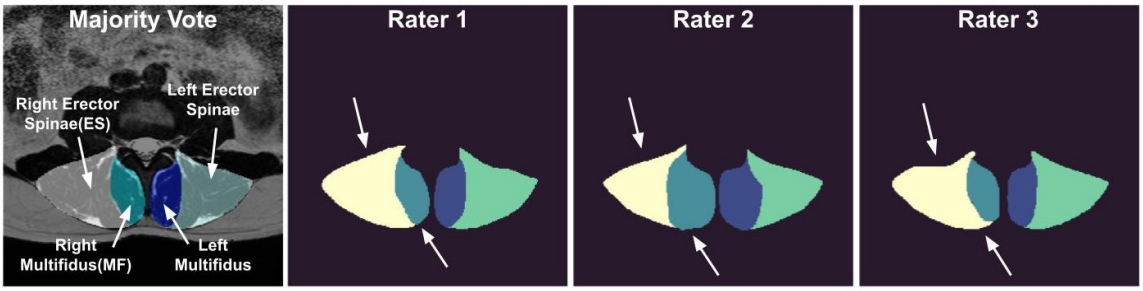

Figure 4.1: Left to right: axial MRI of paraspinal muscles, along with the majority vote label and the individual rater annotations. The arrows indicate the differences among rater annotations.

4.3 Experiments and Results

4.3.1 Experimental Set-up and Implementation Details

Each model is trained to segment four paraspinal muscles. We divide the dataset subject-wise into training, validation, and test sets with respect to age group and sex. Specifically, 24 subjects

are selected for testing, 76 for training, and 19 for validation. For TTD, we add a dropout layer to each of the convolution blocks in the upsampling layers of UNet and TransUNet, resulting in the addition of a total of 8 dropouts in each model. Both models are trained for 250 epochs with early stopping applied to avoid overfitting. We applied random rotation, translation, intensity shift, and Gaussian noise during the training and applied the same augmentations at test time for TTA [59]. As mentioned in the previous section, we use entropy for measuring uncertainties in this study (Eq. 9). In order to explore the correlation between aleatoric and epistemic uncertainties with inter-rater variability, we calculate the Pearson correlation coefficients and conduct a variance partitioning analysis to verify the percentage of variance explained by the uncertainties for inter-rater variability (i.e., GT entropy).

4.3.2 Results

The Brier scores for inter-rater variability preservation assessment are listed in Table 4.1 for all the models and all segmented muscle groups. Additionally, we also plotted the prediction entropy against GT entropy with the associated correlations in Fig 4.2(a). These results indicate that compared with UNet, TransUNet better preserves inter-rater variability with lower Brier scores and higher "prediction entropy vs. GT entropy" correlations. Also, we observe that training with random sampling results in lower Brier scores than using majority vote for the same models. To evaluate the quality of epistemic uncertainty estimation, the AUC-PR results are detailed in Table 4.2, where TransUNet outperforms its UNet counterpart in both TTD and deep ensemble approaches. Here, random sampling, TTD, and deep ensemble show similar performance so for the rest of the experiments to compare random sampling and majority vote, we only trained the deep ensemble model with majority vote GT. Table 4.3 contains the correlations of inter-rater variability with aleatoric and epistemic uncertainties, and the evidence shows that in our case study, inter-rater variability is more strongly associated with epistemic uncertainty than the aleatoric one, and the phenomenon is stronger for TransUNet. The superiority of TransUNet is further demonstrated in Table 4.4, where the models' performance is evaluated with Dice score. According to the scatter plots of Fig 4.2(b), we also see that higher correlation is produced with random sampling training, and the UNet model contains higher epistemic uncertainties. Finally, from the variance partitioning analysis (see Table

4.5), we observe that epistemic uncertainty accounts for $\sim 35\%$ of the GT entropy variance with TransUNet and $\sim 12\%$ with UNet. Additionally, we observe that aleatoric uncertainty only explains a very small portion of the variance.

Table 4.1: Quantitative assessment of inter-rater variability preservation in the trained models.

	Average Brier Score ($\times 10^3$)			
	Right MF	Left MF	Right ES	Left ES
TransUNet (majority vote)	1.324	1.209	1.936	1.890
UNet (majority vote)	2.071	1.973	3.286	3.146
TransUNet (random sampling)	1.179	1.103	1.742	1.758
UNet (random sampling)	1.769	1.670	2.866	2.912

Table 4.2: AUC-PR for epistemic uncertainty. Each column shows a method for measuring the uncertainty and the training method, while the rows indicate the utilized models.

	AUC-PR		
	TTD-majority vote	TTD-random sampling	Deep Ensemble
TransUNet	0.3753	0.3737	0.3831
UNet	0.3387	0.3337	0.3265

Table 4.3: Correlation of epistemic and aleatoric uncertainties with inter-rater variability. Majority vote is shown as "Maj" while random sampling is shown as "Rand".

	Pearson Correlation Coefficient				
	TTD–Maj	TTD-Rand	Deep Ensemble	TTA-Maj	TTA-Rand
TransUNet	0.5874	0.6165	0.5985	0.0500	0.1028
UNet	0.3380	0.3635	0.3574	0.0719	0.0163

4.4 Discussion

With the experiments, the results of the Brier scores and "prediction entropy vs. GT entropy" correlations indicate that both the TransUNet architecture and random sampling have positive impacts on preserving inter-rater variability. Furthermore, as Fig 4.2(b) shows, TransUNet produces lower epistemic uncertainty with tighter distribution. Our results confirm the conclusion of Lemay et al. [32] on the benefit of random sampling in preserving inter-rater variability. Furthermore, we observe that it also results in higher prediction entropies (Fig. 4.2(a)) with more uncertain results. When assessing AUC-PR, TransUNet offers a better quality of epistemic uncertainty estimation while the advantages of different estimation techniques and training strategies are not clear. When comparing the correlations of inter-rater variability (i.e., GT entropy) with aleatoric and epistemic uncertainties, the results in Table 4.3 demonstrate that epistemic uncertainty has a stronger association in our segmentation task and database while no significant correlations with aleatoric uncertainties were found, regardless of the DL model choice. This may be partially explained by the fact that manual segmentations were performed based on pre-processed images with similar noise levels, reducing the chance of inter-rater variability being affected by image noise. Future studies to explore the impact of image noise levels and artifacts (e.g., bias fields) can further verify this hypothesis, but require a more extensive and costly experimental setup with human raters. In addition, there is also a higher correlation between inter-rater variability and epistemic uncertainty with TransUNet and random sampling as shown in Table 4.3, proving that model uncertainty can be network-dependent and better preservation of inter-rater variability leads to a stronger link to the model uncertainty. Although better "epistemic uncertainty vs. inter-rater variability" correlation and preservation of inter-rater variability are desirable as they can result in more effective reduction of uncertainty through lowering inter-rater variability, the overall higher prediction entropy and epistemic uncertainty may be a price to pay in the case of random sampling compared to majority vote. As a final evaluation, we used variance partitioning (Table 4.5) to quantify the contributions of aleatoric and epistemic uncertainties toward inter-rater variability. This way, we leverage the DL models to understand the source of inter-rater variability, which is difficult to quantify from human raters [20]. The results indicate a partial influence of epistemic uncertainty that may be due to the

factors of anatomical variability (common in pathological paraspinal muscles) and difference in visual perception, and minimum contribution from aleatoric uncertainty. This suggests the benefit of preprocessing and systematic training for expert labeling. For all experiments, the incorporation of Transformers has positive impacts in lowering the uncertainty and encoding inter-rater variability, potentially leading to better segmentation accuracy. Their ability to encode long-range content over the image may play a key role in the observed behaviors.

4.5 Conclusion

In this paper, we explored the relationship of inter-rater variability with aleatoric and epistemic uncertainties, using two DL models and two label fusion methods. Our case study indicated that inter-rater variability has a high correlation with epistemic uncertainty and no significant correlation with aleatoric uncertainty. Moreover, we showed that TransUNet better preserves inter-rater variability and its correlation with epistemic uncertainty, and it also has lower epistemic uncertainty and prediction entropy than UNet, potentially explaining its segmentation accuracy. Finally, our results showed that the label fusion method not only affects the preservation of inter-rater variability but it also affects epistemic uncertainty as well.

Table 4.4: Model performance measured by Dice Score. The superior performance of TransUNet compared to the UNet counterpart is indicated by $^{**}(p < 0.01)$ and $^{*}(p < 0.05)$.

	Average Dice Score (%)			
	Right MF	Left MF	Right ES	Left ES
TransUNet (majority vote-TTD)	**94.36±2.74	**94.77±2.37	**94.29±3.35	**94.18±3.77
UNet (majority vote-TTD)	92.41±8.9	92.77±8.52	92.21±9.52	92.06±9.31
TransUNet (random sampling-TTD)	94.15±3.14	94.44±2.53	*94.08±3.38	*93.87±3.94
UNet (random sampling-TTD)	92.76±8.86	93.07±8.67	92.49±9.06	91.92±10.04
TransUNet (ensemble)	*94.72±2.38	*94.80±2.49	**94.75±3.26	**94.50±3.18
UNet (ensemble)	92.83±8.73	93.23±8.56	92.79±8.96	92.05±10.12

Table 4.5: Variance partitioning analysis for inter-rater variability. The values show the percentage of inter-rater variability variation related to the epistemic and aleatoric uncertainties.

		Epistemic (%)	Aleatoric (%)	Joint (%)
TransUNet	TTA+TTD(majority vote)	34.506	0.251	34.896
	TTA+TTD(random sampling)	38.007	0.251	38.360
	TTA+Ensemble	35.803	0.251	35.986
UNet	TTA+TTD(majority vote)	11.422	0.517	11.517
	TTA+TTD(random sampling)	12.772	0.517	12.919
	TTA+Ensemble	13.215	0.517	13.385

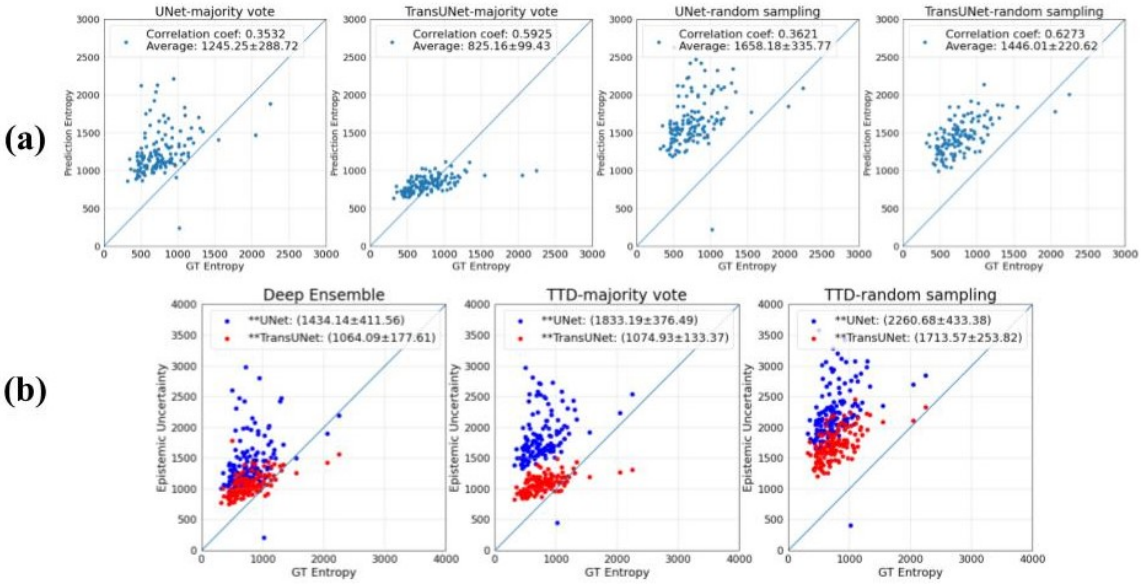

Figure 4.2: (a) Assessment of preservation of inter-rater variability, along with the average entropy and Pearson correlation coefficient shown in the graphs. (b) Comparison of epistemic uncertainty with inter-rater variability, along with the average uncertainties shown in the graphs. Significant correlation is denoted by $^{**}(p < 0.01)$.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, we first introduced a segmentation model designed not only for accurate segmentation but also capable of predicting pixle-wise inter-rater variability. Then, extending our focus on inter-rater variability, we further explored the potential correlations of inter-rater variability with aleatoric and epistemic uncertainties while investigating the impact of factors including model architecture and label fusion methods on these relationships. These studies revealed the critical role of inter-rater variability in medical image segmentation accuracy, described methods for visualizing and predicting inter-rater variability, and elucidated considerations for estimating uncertainties of DL models. In addition, through the experiments, we highlighted the advantages of integrating transformers into the CNN UNet model.

In Chapter 3, we proposed a multi-task TransUNet to simultaneously predict the paraspinal muscle segmentation maps and the inter-rater variability, presented as the pixel-wise variance of the raters' opinions. Highlighting the advantages of integrating CNN and transformer architectures for the specified tasks over the widely adopted UNet, our proposed technique showcases comparable or improved segmentation accuracy for individual paraspinal muscles compared to prior studies. Moreover, our model provides user-friendly visualization about rater disagreements and can be easily extended to other segmentation tasks.

Chapter 4 provides details of our study on the relationship of inter-rater variability with aleatoric

and epistemic uncertainties. Our case study indicated that there is a noticeable correlation between inter-rater variability and epistemic uncertainty, which could mean that the effect of extreme anatomical variations is higher in rater disagreements compared to that of data noise. Our results could also be an indicator of the positive impact of the initial denoising of the dataset on lowering rater disagreements and, subsequently, the uncertainty of DL model outcomes. Moreover, we showed that TransUNet better preserves inter-rater variability and its correlation with epistemic uncertainty, and it also has lower epistemic uncertainty and prediction entropy than UNet. Finally, our results showed label fusion methods not only influence inter-rater variability preservation but also epistemic uncertainty.

5.2 Future Work

In our first work, where we proposed the multi-task TransUNet, our experiments focused on paraspinal muscle segmentation. Expanding these experiments to include other datasets would provide a more comprehensive validation of our model's advantages over other state-of-the-art models. However, large public datasets with multi-rater segmentation are still rare and are the primary barrier for more comprehensive evaluations and extension of the proposed method.

One of the key conclusions of our second work to quantify the relationship between inter-rater variability and uncertainty was that initial denoising of the dataset (prior to manual segmentation by raters) can effectively reduce the impact of data noise on rater disagreements and, therefore, possibly lead to lowering the correlation values between aleatoric uncertainty and inter-rater variability. To confirm and generalize these findings and prove the claims about correlation decrease, future exploration could involve conducting experiments with datasets with various image noise characteristics to confirm the reproducibility of observed phenomena. An interesting avenue here is to perform experiments on ultrasound datasets such as the recently released breast cancer-related lymphedema dataset [21] because of the unique noise characteristics in ultrasound images.

One of the key fundamental factors in our uncertainty study in Chapter 4 is the methodologies employed for uncertainty measurement. To further analyze our current strategies, which rely on sampling from the output distribution for estimating aleatoric and epistemic uncertainties, it can

be beneficial to propose an investigation into the impact of increasing the number of samples (i.e., forward iterations through the network for each image) on the measured correlation between inter-rater variability and the two types of uncertainties. Additionally, expanding the study to include alternative uncertainty estimation methods could provide valuable insights into potential variations in uncertainty values. One alternative to estimating the values of aleatoric and epistemic uncertainties is using evidential deep learning [53] to directly learn the underlying distribution parameters from which the model outputs are sampled. This approach is less studied compared to the other estimation approaches due to the computational complexity. However, comparing the results of this method against more conventional estimation approaches could provide further and more comprehensive insights into the accuracy of uncertainty measurement techniques.

Convolutional neural networks are not shift invariant because of the max-pooling step common in these networks [54, 68]. Therefore, the segmentation maps, especially around the borders of the object are affected by the network structure. The impact of the network's shift-invariance on segmentation uncertainties is another area of future work.

Finally, in our second study, we only compared TransUNet and UNet models. As various variants of U-shaped models with different attention mechanisms have emerged [22, 19, 7], experimenting with other model architectures can further prove the validity of our findings. Note that when exploring different model architectures, particularly concerning the preservation of inter-rater variability, one critical aspect to investigate is model calibration, which evaluates the alignment between the probabilistic outputs of the DL model (model confidence) and its actual accuracy to best take advantage of uncertainty measures as a surrogate to safeguard unreliable outcomes from the DL algorithms. In other words, a DL model that exhibits a misalignment between accuracy and confidence levels cannot be considered satisfactory, even if it outperforms other models in preserving inter-rater variability and accuracy. To further complement our investigations in Chapter 4, we conducted a calibration analysis for the UNet and TransUNet models that we trained, and the results are provided in Appendix A.

Appendix A

Calibration of TransUNet and UNet Models Trained with Different Label Fusion Methods

This section shows the calibration of the UNet and TransUNet models trained in Chapter 4 with accuracy-confidence plots (see Fig. A.1). Following the same method as Lemay et al. [32], we divided the full confidence range of 0 to 1 into 10 intervals, calculated the average accuracy and confidence of voxels in each interval for every image separately, and then obtained the average confidence and accuracy for each bin in the entire test dataset (please refer to Chapter 4 for details). Here, confidence for each voxel is defined as the maximum of output probabilities. For example, if the output probabilities for a specific voxel in a 3-class segmentation are [0.1, 0.2, 0.7] then the model's confidence about its prediction is 0.7. For each image, voxels are divided into 10 bins according to their output confidence, and then the average accuracy and confidence were calculated for each bin. As an example, in a well-calibrated model, if the average confidence of a group of voxels is 0.82, the model's accuracy in those voxels should also be close to 80 percent. In our study, TransUNet results in better model calibration compared to UNet, and random sampling outperforms majority voting in model calibration quality. Our calibration analysis suggests that training models with majority voting results in over-confident models in general, which is in line with the previous

findings in this field [32].

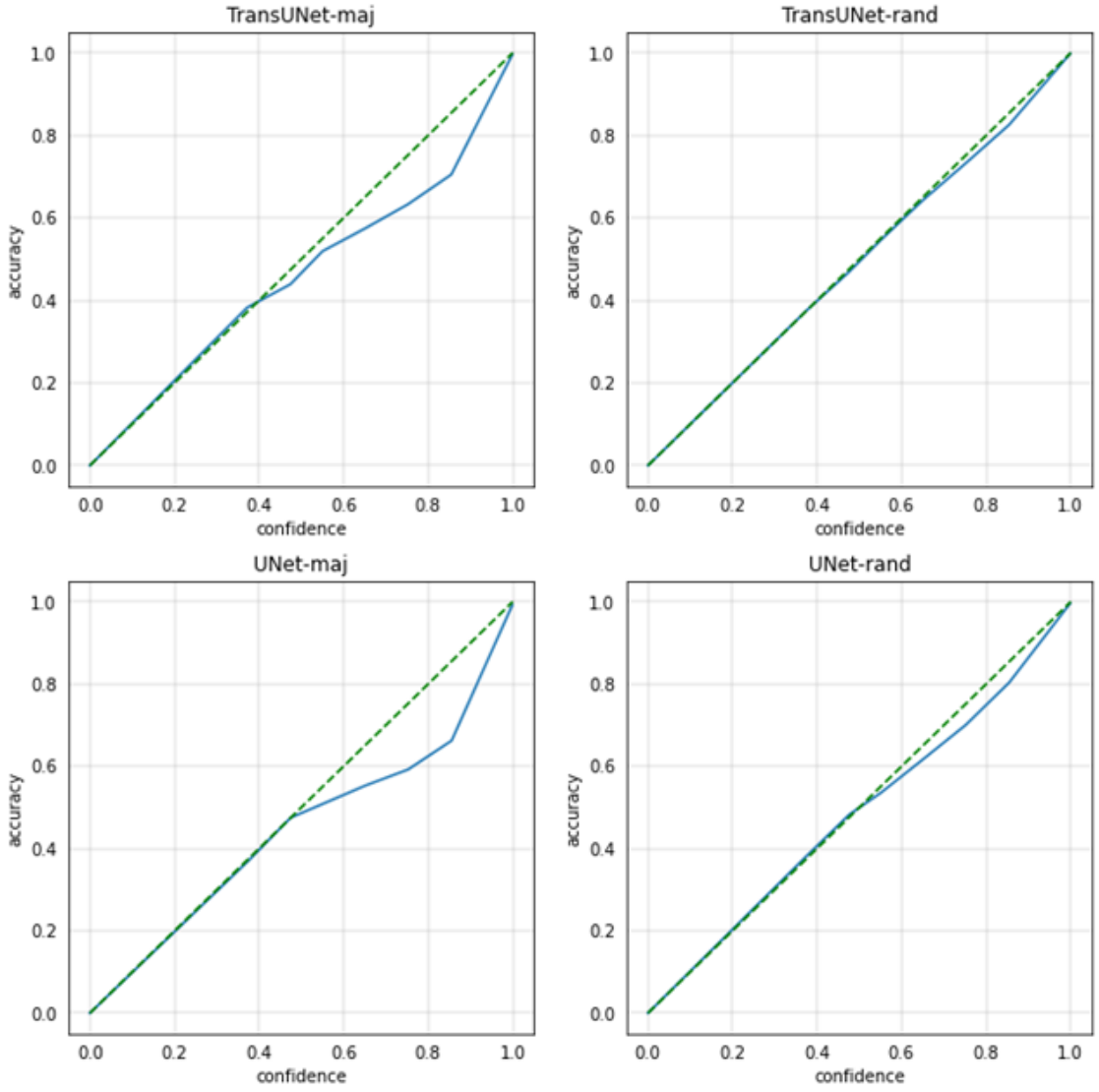

Figure A.1: Calibration of the UNet and TransUNet models trained with majority vote (maj) and randomly sampled (rand) ground truths

Bibliography

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

[2] Mina Amiri, Rupert Brooks, Bahareh Behboodi, and Hassan Rivaz. Two-stage ultrasound image segmentation using u-net and test time augmentation. *International journal of computer assisted radiology and surgery*, 15:981–988, 2020.

[3] Meagan Anstruther, Bianca Rossini, Tongwei Zhang, Terrance Liang, Yiming Xiao, and Maryse Fortin. Pillar: Paraspinal muscle segmentation project-a comprehensive online resource to guide manual segmentation of paraspinal muscles from magnetic resonance imaging. 2023.

[4] Sayantan Auddy, Ramit Dey, Min-Kai Lin, Daniel Carrera, and Jacob B Simon. Using bayesian deep learning to infer planet mass from gaps in protoplanetary disks. *The Astrophysical Journal*, 936(1):93, 2022.

[5] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.

[6] Robin Camarasa, Daniel Bos, Jeroen Hendrikse, Paul Nederkoorn, M Eline Kooi, Aad van der Lugt, Marleen de Bruijne, et al. A quantitative comparison of epistemic uncertainty maps

applied to multi-class segmentation. *Machine Learning for Biomedical Imaging*, 1(UNSURE2020 special issue):1–39, 2021.

[7] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[8] Rohitash Chandra, Mahir Jain, Manavendra Maharana, and Pavel N. Krivitsky. Revisiting bayesian autoencoders with mcmc. *IEEE Access*, 10:40482–40495, 2022.

[9] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[10] Jeffrey R. Cooley, Bruce F. Walker, Emad M. Ardakani, Per Kjaer, Tue S. Jensen, and Jeffrey J. Hebert. Relationships between paraspinal muscle morphology and neurocompressive conditions of the lumbar spine: a systematic review with meta-analysis. *BMC Musculoskeletal Disorders*, 19(1), September 2018.

[11] Pierrick Coupé, Pierre Yger, Sylvain Prima, Pierre Hellier, Charles Kervrann, and Christian Barillot. An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images. *IEEE transactions on medical imaging*, 27(4):425–441, 2008.

[12] Zhijie Deng, Yucen Luo, Jun Zhu, and Bo Zhang. Measuring uncertainty through bayesian learning of deep neural network structure. *arXiv preprint arXiv:1911.09804*, 2019.

[13] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] Nikita Durasov, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Masksembles for uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13539–13548, 2021.

[16] Maryse Fortin, Àron Lazáry, Peter Paul Varga, and Michele C Battié. Association between paraspinal muscle morphology, clinical symptoms and functional status in patients with lumbar spinal stenosis. *European spine journal*, 26:2543–2551, 2017.

[17] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

[18] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[19] Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 61–71. Springer, 2021.

[20] Asma Ghandeharioun, Brian Eoff, Brendan Jou, and Rosalind Picard. Characterizing sources of uncertainty to proxy calibration and disambiguate annotator and data bias. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4202–4206. IEEE, 2019.

[21] Sobhan Goudarzi, Jesse Whyte, Mathieu Boily, Anna Towers, Robert D. Kilgour, and Hassan Rivaz. Segmentation of arm ultrasound images in breast cancer-related lymphedema: A database and deep learning algorithm. *IEEE Transactions on Biomedical Engineering*, 70(9):2552–2563, 2023.

[22] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[23] Jiawei Huang, Haotian Shen, Bo Chen, Yue Wang, and Shuo Li. Segmentation of paraspinal muscles at varied lumbar spinal levels by explicit saliency-aware learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, pages 652–661. Springer, 2020.

[24] Martin Holm Jensen, Dan Richter Jørgensen, Raluca Jalaboi, Mads Eiler Hansen, and Martin Aastrup Olsen. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 540–548. Springer, 2019.

[25] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12341–12351, 2021.

[26] Craig K Jones, Guoqing Wang, Vivek Yedavalli, and Haris Sair. Direct quantification of epistemic and aleatoric uncertainty in 3d u-net segmentation. *Journal of Medical Imaging*, 9(3):034002–034002, 2022.

[27] Nirmala Devi Kathamuthu, Shanthi Subramaniam, Quynh Hoang Le, Suresh Muthusamy, Hitesh Panchal, Suma Christal Mary Sundararajan, Ali Jawad Alrubaie, and Musaddak Maher Abdul Zahra. A deep transfer learning-based convolution neural network model for covid-19 detection using computed tomography scan images for medical applications. *Advances in Engineering Software*, 175:103317, 2023.

[28] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[29] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6129–6138, 2017.

[30] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[31] Max-Heinrich Laves, Sontje Ihler, Jacob F Fast, Lüder A Kahrs, Tobias Ortmaier, et al. Recalibration of aleatoric and epistemicregression uncertainty in medical imaging. *Machine Learning for Biomedical Imaging*, 1(MIDL 2020 special issue):1–26, 2021.

[32] Andreanne Lemay, Charley Gros, Enamundram Naga Karthik, Julien Cohen-Adad, et al. Label fusion and training methods for reliable representation of inter-rater uncertainty. *Machine Learning for Biomedical Imaging*, 1(January 2023 issue):1–27, 2023.

[33] Haixing Li, Haibo Luo, and Yunpeng Liu. Paraspinal muscle segmentation based on deep neural network. *Sensors*, 19(12):2650, 2019.

[34] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[35] Yikuan Li, Shishir Rao, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Gholamreza Salimi-Khorshidi, Mohammad Mamouei, Thomas Lukasiewicz, and Kazem Rahimi. Deep bayesian gaussian processes for uncertainty estimation in electronic health records. *Scientific reports*, 11(1):20685, 2021.

[36] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[37] Raghav Mehta, Angelos Filos, Ujjwal Baid, Chiharu Sako, Richard McKinley, Michael Rebsamen, Katrin Dätwyler, Raphael Meier, Piotr Radojewski, Gowtham Krishnan Murugesan, et al. Qu-brats: Miccai brats 2020 challenge on quantifying uncertainty in brain tumor segmentation-analysis of ranking scores and benchmarking results. *The journal of machine learning for biomedical imaging*, 2022, 2022.

[38] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[39] Zahra Mirikharaji, Kumar Abhishek, Saeed Izadi, and Ghassan Hamarneh. D-lema: Deep learning ensembles from multiple annotations-application to skin lesion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1837–1846, 2021.

[40] Aryan Mobiny, Pengyu Yuan, Supratik K Moulik, Naveen Garg, Carol C Wu, and Hien Van Nguyen. Dropconnect is effective in modeling uncertainty of bayesian deep networks. *Scientific reports*, 11(1):5458, 2021.

[41] Jishnu Mukhoti, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty for semantic segmentation. *arXiv preprint arXiv:2111.00079*, 2021.

[42] Brennan Nichyporuk, Jillian Cardinell, Justin Szeto, Raghav Mehta, Jean-Pierre Falet, Douglas L Arnold, Sotirios A Tsaftaris, Tal Arbel, et al. Rethinking generalization: The impact of annotation style on medical image segmentation. *Machine Learning for Biomedical Imaging*, 1(December 2022 issue):1–37, 2022.

[43] Alex M. Noonan and Stephen H. M. Brown. Paraspinal muscle pathophysiology associated with low back pain and spine degenerative disorders. *JOR SPINE*, 4(3), September 2021.

[44] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[45] OpenAI. Gpt-4 technical report, 2023.

[46] Kilian M Pohl, John Fisher, Sylvain Bouix, Martha Shenton, Robert W McCarley, W Eric L Grimson, Ron Kikinis, and William M Wells. Using the logarithm of odds to define a vector space on probabilistic atlases. *Medical Image Analysis*, 11(5):465–477, 2007.

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[48] Parinaz Roshanzamir, Hassan Rivaz, Joshua Ahn, Hamza Mirza, Neda Naghdi, Meagan Anstruther, Michele C Battié, Maryse Fortin, and Yiming Xiao. Joint paraspinal muscle segmentation and inter-rater labeling variability prediction with multi-task transunet. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 125–134. Springer, 2022.

[49] Parinaz Roshanzamir, Hassan Rivaz, Joshua Ahn, Hamza Mirza, Neda Naghdi, Meagan Anstruther, Michele C Battié, Maryse Fortin, and Yiming Xiao. How inter-rater variability relates to aleatoric and epistemic uncertainty: a case study with deep learning-based paraspinal muscle segmentation. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 74–83. Springer, 2023.

[50] Soorena Salari, Amirhossein Rasoulian, Hassan Rivaz, and Yiming Xiao. Focalerrornet: Uncertainty-aware focal modulation network for inter-modal registration error estimation in ultrasound-guided neurosurgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 689–698. Springer, 2023.

[51] Gabriele Scalia, Colin A. Grambow, Barbara Pernici, Yi-Pei Li, and William H. Green. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of Chemical Information and Modeling*, 60(6):2697–2717, April 2020.

[52] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019.

[53] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi,

and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[54] Mostafa Sharifzadeh, Habib Benali, and Hassan Rivaz. Investigating shift variance of convolutional neural networks in ultrasound image segmentation. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 69(5):1703–1713, 2022.

[55] Ali KZ Tehrani, Ivan M Rosado-Mendez, and Hassan Rivaz. Homodyned k-distribution: parameter estimation and uncertainty quantification using bayesian neural networks. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2023.

[56] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[58] Olivier Vincent, Charley Gros, and Julien Cohen-Adad. Impact of individual rater style on deep learning uncertainty in medical imaging segmentation. *arXiv preprint arXiv:2105.02197*, 2021.

[59] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

[60] Samuel R Ward, Choll W Kim, Carolyn M Eng, Lionel J Gottschalk IV, Akihito Tomiya, Steven R Garfin, and Richard L Lieber. Architectural analysis and intraoperative measurements demonstrate the unique design of the multifidus muscle for lumbar spine stability. *The Journal of bone and joint surgery. American volume.*, 91(1):176, 2009.

[61] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[62] Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

[63] Wenyao Xia, Maryse Fortin, Joshua Ahn, Hassan Rivaz, Michele C Battié, Terry M Peters, and Yiming Xiao. Automatic paraspinal muscle segmentation in patients with lumbar pathology using deep convolutional neural network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 318–325. Springer, 2019.

[64] Yiming Xiao, Maryse Fortin, Joshua Ahn, Hassan Rivaz, Terry M Peters, and Michele C Battié. Statistical morphological analysis reveals characteristic paraspinal muscle asymmetry in unilateral lumbar disc herniation. *Scientific Reports*, 11(1):15576, 2021.

[65] Guoping Xu, Xingrong Wu, Xuan Zhang, and Xinwei He. Levit-unet: Make faster encoders with transformer for medical image segmentation. *arXiv preprint arXiv:2107.08623*, 2021.

[66] Feng Yang, Ghada Zamzmi, Sandeep Angara, Sivaramakrishnan Rajaraman, André Aquilina, Zhiyun Xue, Stefan Jaeger, Emmanouil Papagiannakis, and Sameer K. Antani. Assessing inter-annotator agreement for medical image segmentation. *IEEE Access*, 11:21300–21312, 2023.

[67] Nina Youneszade, Mohsen Marjani, and Chong Pei Pei. Deep learning in cervical cancer diagnosis: Architecture, opportunities, and open research challenges. *IEEE Access*, 11:6133–6149, 2023.

[68] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019.