# Advances in Uncertainty Quantification for Deep Learning-Based Medical Image Analysis

By

Brayden John Schott

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Medical Physics)

at the
UNIVERSITY OF WISCONSIN-MADISON
2025

Date of final oral examination: June 25, 2025

The dissertation is approved by the following members of the Final Oral Committee:

Robert Jeraj, Professor, Medical Physics
Diego Hernando, Associate Professor, Medical Physics and Radiology
Oliver Wieben, Professor, Medical Physics and Radiology
Sharon Yixuan Li, Assistant Professor, Computer Sciences
Scott Perlman, Professor, Medicine

© Copyright by Brayden John Schott 2025

All Rights Reserved

# Acknowledgements

First and foremost, it is imperative that I begin by acknowledging my advisor, Dr. Robert Jeraj. Thank you for seeing potential in me as an aspiring PhD student, for guiding me as I stumbled my way through research, and for supporting me as I embark on new horizons. I must similarly thank my current research group members, who are among the brightest and most thoughtful minds I know and span across multiple continents: Dr. Ali Deatsch, Dr. Victor Santoro-Fernandes, Dr. Žan Klaneček, Katja Strašek, Luciano Rivetti, Sebastian Salgado-Maldonado, Ilam Balasubramanian, and Zahra Alirezaei. Former group members—Dr. Tim Perk, Dr. Amy Weisman, Dr. Alison Roth, Dr. Peter Ferjančič, and Dr. Daniel Huff—were also instrumental in shaping my early PhD formation. To the undergraduate students who I had the privilege of mentoring—Dmitry Pinchuk, Angela Galindo-Santos, and Nicholas Schense—thank you for bringing a renewed sense of curiosity to our lab.

Next, I must thank my committee members for helping me shape my tentative ideas into a coherent and meaningful research story. I am especially grateful to Dr. Oliver Wieben, Dr. Diego Hernando, and Dr. Sharon Li for their exceptional technical feedback. I also thank Dr. Scott Perlman for his insightful clinical insights, which added a valuable perspective on my work. More broadly, I am thankful to the Department of Medical Physics at the University of Wisconsin—Madison for fostering an outstanding academic and research environment. To the many support personnel within the department—including Carol Aspinwall, Taylor Hartung, Dennis Commerford, and Erik Deitz—thank you for minimizing the technical and logistical chaos along the way.

Reaching further back, I would like to thank those who were formative in my academic journey leading to my PhD. I am grateful to Dr. Darren Craig at Wheaton College for his guidance throughout my

undergraduate years and for mentoring me during my early research endeavors. I also thank Dr. Heather Whitney for first introducing me to the field of medical physics. I am similarly thankful to Dr. Baozhou Sun, who taught me the clinical aspects of medical physics in Washington University's Department of Radiation Oncology and entrusted me with critical responsibilities, which instilled within me a deep appreciation for the field.

I would also like to thank my friends who helped to balance my composure. Within the department, I am grateful to Dr. Ethan Nikolau, Dr. Xin Tie, and Aidan Tollefson. Outside of the department, my heartfelt thanks go to Magnus Zaunmueller, Nicholas Gulachek, Father Nektarios Akkawi, and the Dinkov family. To my parents, sister, and in-laws, thank you for your constant support and belief in me throughout this journey. To my wife Angela, I am deeply grateful for your unwavering encouragement and faith in our endeavors—this work would not have been possible without you. And finally, to my dear Johnny, thank you for being an unfailing source of contagious giggles, playful affection, and boundless joy.

# Table of Contents

# List of Abbreviations and Acronyms

| | |
|---|---|
| AC-CT | Attenuation correction computed tomography |
| AI | Artificial Intelligence |
| AUC | Area under the receiver-operator curve |
| BNN | Bayesian neural network |
| CE-CT | Contrast enhanced computed tomography |
| CNN | Convolutional neural network |
| CT | Computed tomography |
| FN | False negative |
| FP | False positive |
| FPR95 | False positive rate at 95% sensitivity |
| ID | In-distribution |
| MCDO | Monte Carlo dropout |
| mCRPC | Metastatic castrate resistant prostate cancer |
| ME | Model ensembling |
| MIP | Maximum intensity projection |
| MRI | Magnetic Resonance Imaging |
| MSP | Maximum softmax probability |
| NET | Neuroendocrine tumor |
| OOD | Out-of-distribution |
| PET | Positron emission tomography |
| PFS | Progression Free Survival |
| QA | Quality Assurance |
| ROC | Receiver Operator Curve |
| SUV | Standardized uptake value |
| TP | True positive |
| TTA | Test time augmentation |
| UQ | Uncertainty quantification |
| UWCCC | University of Wisconsin Carbone Cancer Center |

# List of Figures

# List of Tables

# Abstract

The predictive outputs of deep learning models inherently contain uncertainty, which is often imperceivable to the user and is not readily available from standard model outputs. Without uncertainty information, the reliability of model predictions cannot be ensured, and model failures may not be appropriately accounted for. This poses a severe threat within clinical settings, where deep learning models are increasingly studied with the intent to inform and enhance patient care. To facilitate the responsible implementation of clinical deep learning models, where the reliability of predictive outputs is informed by uncertainty information, uncertainty quantification (UQ) methods are essential, yet remain largely underexplored. The main goal of this work is to investigate existing UQ methods for the metastatic lesion segmentation task, consider their limitations, and propose novel UQ approaches to overcome these limitations and improve UQ utility.

To that end, we have implemented, developed, and validated UQ methods which target uncertainty arising from two types of data distribution—out-of-distribution (OOD) uncertainty and in-distribution (ID) uncertainty. OOD uncertainty arises from discrepancies between image features present in the train and test data. On the other hand, ID uncertainty arises from various sources that impede robust model fitting, such as train data noise, despite shared similarities between the train and test data. Capturing both OOD and ID uncertainties is essential for comprehensive UQ. Therefore, this work considers and addresses both types of uncertainty, with methods tailored specifically to the medical image domain.

To capture OOD uncertainty, we implemented established methods and compared these to an introduced approach that utilizes a post hoc information bottleneck optimization process to quantify the discrepancy between train and test images via shared mutual information of embedded train and test

features. We then assessed and quantitatively compared the performance of several established ID UQ methods for the critical clinical task of metastatic lesion segmentation on full-body PET/CT images. Finally, we introduced a novel ID UQ method that leverages localized gradient information to capture the sensitivities of local model outputs to trained model parameters.

The novel, information-based OOD uncertainty measure outperformed established OOD measures in detecting computerized tomography (CT) test images with simulated artifacts. The introduced OOD uncertainty measure also correlated significantly with model segmentation performance metrics. When investigating established ID UQ measures for metastatic lesion segmentation, we found that the test time augmentation method achieved superior results across quantitative evaluations, such as false negative region recovery and correlations with segmentation performance metrics. Lastly, the novel localized gradients-based ID UQ demonstrated superior performance compared to standard model outputs for detecting artificially perturbed data, lower quality clinical data, detecting false positive predicted regions, and correspondence with physician disease likelihood ratings.

In summary, this work offers new solutions for the UQ of deep learning models and assesses the performance of existing methods. These uncertainty measures hold many practical applications, including providing robust reliability measures of model outputs, acquiring reliability measures for downstream clinical tasks via uncertainty propagation, and enhancing model robustness through specialized training frameworks such as active learning. Ultimately, this work contributes to the responsible deployment of clinical deep learning models where the reliability of predictive outputs is effectively characterized via uncertainty quantification to avoid undue harm to patient care.

# Introduction

## 1.1 Overview

Deep learning models have demonstrated strong performance across a wide variety of automatic medical image analysis tasks, including structure segmentation (Minaee et al., 2022), pathology diagnosis (Rana & Bhushan, 2023), image registration (Fu et al., 2020), and more (Chen et al., 2022). Not only are deep learning models frequently superior to traditional statistical and machine learning-based computer-assisted image analysis tools, but they have also achieved comparable performance to expert human readers across various medical image analysis tasks (X. Liu et al., 2019). Given this level of effectiveness, it is apparent that deep learning models hold strong potential to greatly impact medical settings, and their large-scale clinical deployment appears imminent. Most deep learning approaches, however, are missing a critical component for safe and responsible clinical use—uncertainty quantification (UQ).

Deep learning model outputs inevitably contain associated uncertainty, which is not apparent from standard model outputs (Abdar et al., 2021). Without uncertainty information, the trustworthiness of model outputs cannot be known. This poses a severe threat within the medical image analysis domain, where deep learning models are being studied with the intent to aid clinical decision-making. The lack of uncertainty information could potentially induce unseen errors in clinical workflows that will negatively impact patient care. It is therefore critical to integrate UQ into clinical models to provide clinicians with more reliable output information and to avoid consequential errors.

Fundamental to the need for UQ is the presence of poor model probability calibration. Standard model outputs for most deep learning tasks include a binarized class prediction preceded by an associated class-wise prediction probability. It is conceivable that these probability outputs may capture the prediction uncertainty, however, these outputs are insufficient in capturing predictive uncertainty. This is because the probability outputs of contemporary models are inherently poorly calibrated (Guo et al., 2017), meaning the predicted probability values deviate from a meaningful representation of task accuracy. As a result, deep learning models yield uninterpretable probability values, where very high probabilities may be assigned to incorrect predictions. Consequently, a model's probability output should not be used to infer a level of trust in the model output. Instead, auxiliary UQ methods are needed.

UQ methods yield uncertainty measures that supplement the standard model outputs (i.e., predicted binary classifications) and indicate some level of reliability associated with those outputs (Abdar et al., 2021). Some uncertainty measures are specifically designed to be calibrated probability values, ranging between 0 and 1, and can be interpreted as such. Most uncertainty measures, however, cannot be interpreted in terms of the probability or likelihood of the associated model output classification. Instead, the relative magnitude of an uncertainty measure is used to gauge output reliability. Most measures are bound by some low uncertainty limit (e.g., 0), yet have no upper limit, extending to infinity. Unless stated otherwise, uncertainty measures follow an indirect relationship with the associated output reliability, where higher uncertainty measures are associated with lower reliability model outputs and vice versa.

In this chapter, we will describe the sources of uncertainty in deep learning models, discuss the relationship between uncertainty and data domains, provide an overview of current UQ approaches, and introduce the contributions of this thesis work to advance UQ for medical image analysis applications.

## 1.2 Uncertainty Sources

The *predictive uncertainty* of deep learning models is categorized by its two underlying sources: *aleatoric* (or *data*) *uncertainty* and *epistemic* (or *model*) *uncertainty* (Abdar et al., 2021). Aleatoric uncertainty arises from noise or errors in the training data, either in the measured data (e.g., image acquisition) or in the labeled data (e.g., manual segmentations). This type of uncertainty is considered irreducible since it is unavoidable, and it establishes a hard constraint to the model fitting process. On the other hand, epistemic uncertainty is caused by the inadequacies of a model to map a given input to the desired output, either because of insufficient knowledge of the data space, poor training procedures, or insufficient model architecture. Epistemic uncertainty can be minimized with enhanced training approaches (e.g., more data, better model architecture, etc.), and therefore, is characterized as reducible. A third source of predictive uncertainty known as *distributional uncertainty* may also be defined. This uncertainty arises from a mismatch between image features present in the model's train and test data. Since this feature mismatch is a type of "insufficient knowledge of the data space", this work considers distributional uncertainty as a subset of epistemic uncertainty. Additional details about this type of uncertainty and how it can be managed are provided in Section 1.4.1.

These technical categorizations of uncertainty can be described using a toy example. In Figure 1, a sinusoidal curve was fitted to train data, which roughly followed a sinusoidal pattern. Four distinct areas which incite predictive uncertainty of the model are highlighted. In region (a), high aleatoric uncertainty is induced due to noise in the training data. In region (b), high epistemic uncertainty is present because the parameters of the fitted model (i.e., sinusoidal parameters) are not sufficient to fit the data in this region, which diverges from the sinusoidal assumption. In region (c), high epistemic uncertainty, particularly high distributional uncertainty, is present due to a gap in the train data. In this

region, it is not certain if the fitted model should continue to be sinusoidal. The same cause induces high epistemic uncertainty in region (d), where the extrapolation of the fitted model beyond the train data domain is not guaranteed.

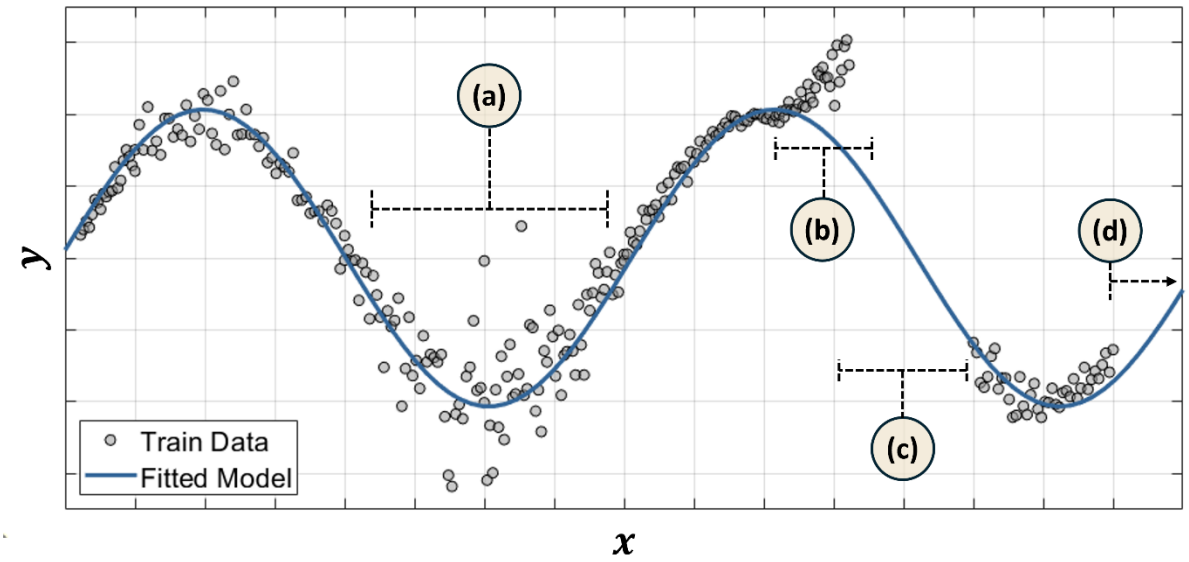



*Figure 1 A toy model describing the different sources of predictive uncertainty. Region (a) contains noisy train data, which induces higher aleatoric uncertainty. Region (b) contains train data that diverges from the parametric model assumption, which induces higher epistemic uncertainty. Regions (c) and (d) have missing train data where it is unclear if the sinusoidal fit is appropriate, which induces higher distributional uncertainty (a subtype of epistemic uncertainty).*

For deep learning problems within the medical image domain, predictive uncertainty arises from a variety of practical and unique factors. For instance, disease pathologies often exhibit ambiguous presentations, where it is difficult for a human observer to provide definite data labels (Kumbhar et al., 2021). Medical image datasets also regularly contain inconsistencies due to differences in image acquisition protocols, such as variations in contrast enhancements, MRI pulse sequences, image reconstruction algorithms, patient populations, and more. Clinical data is also prone to degraded image quality due to image artifacts and drifts in scanner output over time (Chow & Paramesran, 2016; Sahiner et al., 2023). Differences in images acquired from different scanners may also increase a model's

predictive uncertainty. Moreover, medical image datasets often suffer from class imbalance, making models biased toward specific output classes, potentially driving poor model calibration (Leevy et al., 2018). These datasets also contain large, multi-dimensional data, which necessitates the use of large-parameter models, and in turn, increases the risk of model over-fitting (Li et al., 2019). Lastly, medical image datasets frequently contain small numbers of samples due to laborious image labeling requirements and strict data-sharing policies. Small datasets accentuate the sources of predictive uncertainties, which could otherwise be smoothed out with larger datasets. Together, all these uncertainty sources contribute to noise in the model fitting process, force the model to form ambiguous decision boundaries, and predispose the model to unfamiliar data upon deployment.

## 1.3 Uncertainties and Data Distributions

It is useful to consider predictive uncertainty as it relates to different data distributions. It should be noted that the term 'data distributions' is occasionally used interchangeably with 'data domains'. For consistency with the majority of the literature, this thesis will use the term 'data distribution' throughout. In-distribution (ID) and out-of-distribution (OOD) data refer to data that lie within and outside of a model's train data feature distribution, respectively, where a feature distribution describes the span of data features present within a dataset (Yang et al., 2024). These features describe image properties such as intensities, contrast, edges, noise, and more. Fundamentally, OOD data is test data that inherently differs from the model's train data, whereas ID data is test data that is similar to a model's train data. OOD data can be drastically different, such as data from semantically different classes (e.g., an abdominal CT for a model trained to classify brain MR images). Alternatively, OOD data can be more subtly different from the train data, where an image belongs to the same semantic class as the train data yet exhibits some feature differences that push it outside the training distribution (e.g.,

excessive image blurring, noise, artifacts). This latter example of OOD data describes covariate-shifted data. These types of shifts likely contribute to the alarming lack of generalizability of medical image deep learning models (Kelly et al., 2019) due to the trained model's unfamiliarity with the test image's feature representation.

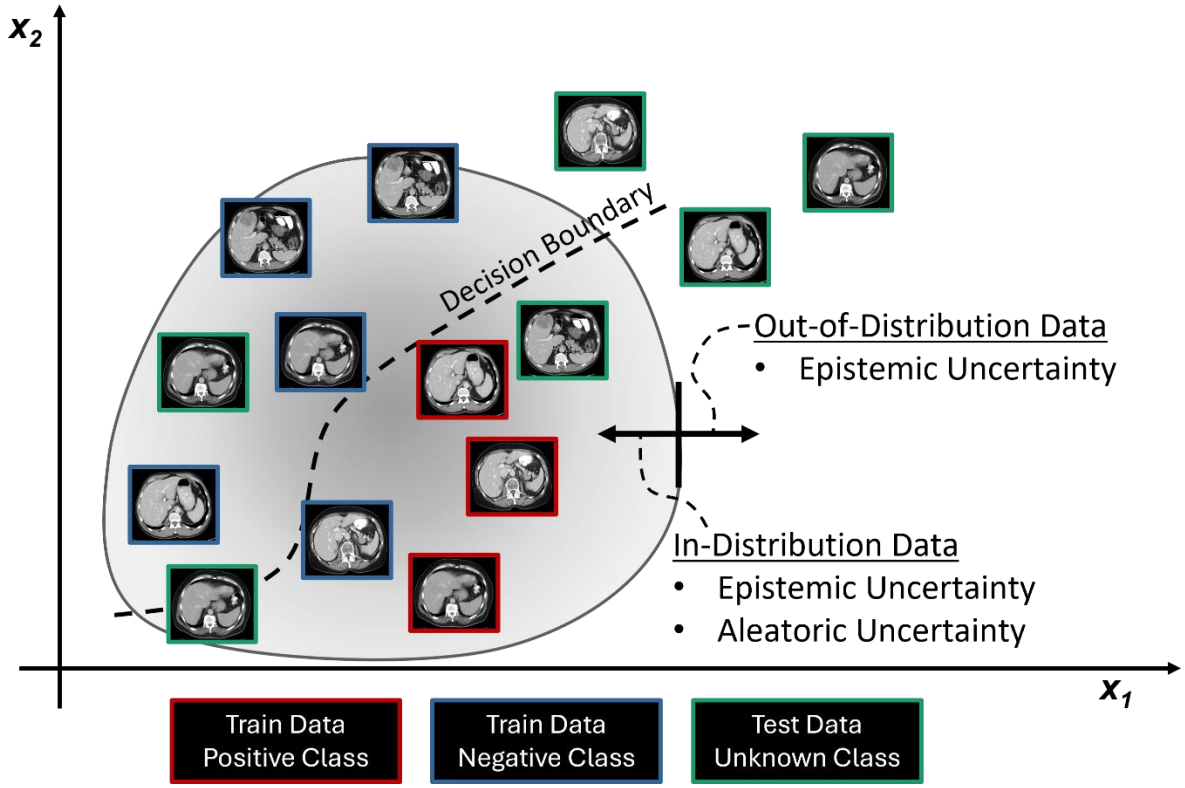



*Figure 2 Schematic depicting data distributions and uncertainty sources. Both epistemic and aleatoric uncertainty are present in the in-distribution data (greyed-out region), while only epistemic uncertainty is present in the out-of-distribution data.*

Figure 2 demonstrates different uncertainty sources in the two data distributions. Consider a simple classification problem where a deep learning model is trained to distinguish between positive and negative samples. In so doing, the model implicitly defines the ID data distribution. Predictions on test samples that fall within this data distribution (i.e., ID data) are subject to both aleatoric and epistemic uncertainty since concerns such as limitations in model fitting or ambiguity caused by noisy train data

may induce greater uncertainty. On the other hand, predictions on test samples that fall outside of the train data distribution (i.e., OOD data) are subject to epistemic uncertainty because the model lacks knowledge about how to handle such data (Abdar et al., 2021).

## 1.4 Uncertainty Quantification Methods

When seeking to quantify the uncertainty of predictive outputs, it is considered good practice to distinguish between methods that target uncertainty arising from OOD and ID sources. This is due to the concern over extrapolating decision boundaries, which are learned during training using the train data, to OOD data. ID UQ methods generally capture uncertainty that lies along learned decision boundaries. Since the model did not see OOD data during training, the learned decision boundaries and the associated decision uncertainty cannot be used for either prediction or UQ on OOD data. This intuitive rationalization is supported by experimentation, where ID UQ methods have been shown to fail to capture OOD uncertainty (Lambert et al., 2022; Ovadia et al., 2019; Schwaiger et al., 2020; Thagaard et al., 2020). Consequently, comprehensive UQ of deep learning models requires separate approaches to capture OOD and ID uncertainty.

### 1.4.1 Out-of-Distribution Uncertainty Quantification

OOD uncertainty is accounted for by invoking OOD detection algorithms. These algorithms attempt to determine whether the features from a given test image belong to the train data distribution. If not, the test image is considered OOD, and the associated model prediction should not be trusted since the model is not familiar with the image's features. Consequently, the level of prediction reliability scales inversely with OOD measure magnitude, where high OOD test samples are considered to have low

reliability and vice versa. OOD detection algorithms typically follow one of four approaches (Yang et al., 2024) and are briefly summarized in Figure 3.

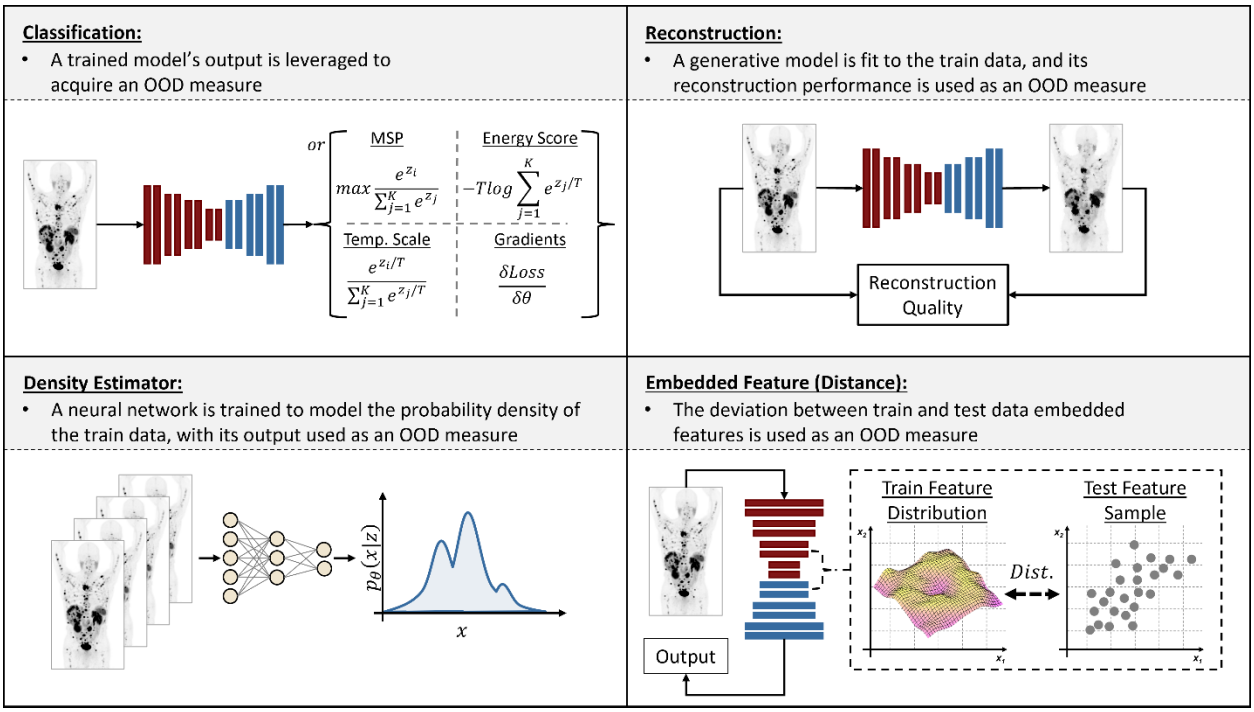



*Figure 3 Schematic describing the different families of OOD detection algorithms.*

**Classification-based** methods utilize a trained classification model's output for OOD detection purposes. The most basic of these approaches utilize a model's maximum softmax probability (MSP) output as an OOD measure, where it is assumed that high confidence probabilities will be assigned to ID distribution data and vice versa (Hendrycks & Gimpel, 2017). Building from this basic approach, methods have been proposed that alternatively process a trained model's logit output (i.e., the output directly before sigmoid activation for softmax probability acquisition) through temperature scaling (Liang et al., 2018a) or energy scoring (Liu et al., 2020) to enhance the spread between measures on ID and OOD data. Classification-based methods have also incorporated gradient information computed from a model's output space for OOD detection. For example, (Liang et al., 2018a) perturbed input data using

input gradients to enhance the separability between ID and OOD data when using temperature-scaled outputs as an OOD measure. Another work directly used the gradients of a classification model's output with respect to intermediary model layer weights for OOD detection (R. Huang et al., 2021). Classification-based OOD detection methods are typically post hoc, meaning they can be employed on a previously trained model and do not require additional training.

**Density-based** methods seek to approximate the probability density of a model's train data, and, in turn, flag test data with low probability likelihood as OOD. Generative models such as variational autoencoders (Goodfellow et al., 2014), generative adversarial networks, or flow-based models (Rezende & Mohamed, 2016) may be employed, where the loss function approximates probability likelihood. These methods advantageously provide a statistically interpretable OOD measure that theoretically aligns with the likelihood of a test image belonging to the probability density distribution defined by the train data. Unfortunately, these methods have been shown to unexpectedly assign high likelihoods to OOD data (Nalisnick et al., 2019). Due to this non-intuitive and sometimes unexplainable behavior, density estimator-based methods are not yet favored for OOD detection in medical image applications. However, work is being done to better understand the application of density-based methods for OOD detection (Kirichenko et al., 2020; Nalisnick et al., 2019), and alternative statistical measures that demonstrate the more favorable OOD detection behavior, such as typicality (Abdi et al., 2025; Nalisnick et al., 2019), are being proposed.

**Reconstruction-based** approaches train a generative model (e.g., an autoencoder) to encode and reconstruct train data, where the reconstruction performance is used as an OOD metric. High OOD data is expected to yield poor reconstructions since the trained model is unfamiliar with the data's feature representations. Due to voxel-wise image reconstruction, these methods have the unique benefit of localizing certain image regions, which may drive the overall OOD measure. For this reason,

reconstruction-based approaches are often employed for anomaly detection (Tschuchnig & Gadermayr, 2024). Reconstruction-based methods, however, tend to lack sensitivity for detecting subtle OOD shifts, especially the covariate shifts expected for medical image OOD detection problems (Denouden et al., 2018; Meissen et al., 2022).

**Embedded feature-based** methods generally have greater detection sensitivity than reconstruction-based methods (Denouden et al., 2018) and operate by invoking some similarity measure, such as a distance measure, between a set of embedded train and test features (i.e., model activations), where large dissimilarities are associated with OOD data. Embedded featured methods are typically post hoc and require lower computational resources. However, the interpretability of these methods suffers as the measurements are derived from information in the abstract embedded feature space. Despite this limitation, embedded feature-based OOD detection methods, particularly distance-based methods, are preferred for medical image OOD detection due to the reported sensitivity gains over alternative approaches.

## Implementation in Medical Image Analysis

Several types of distance-based embedded feature-based OOD measures have been previously investigated for medical image applications. For example, the *Euclidean distance* of embedded features was successfully used to detect OOD data in the context of organ segmentation on both MRI and CT images (Karimi & Gholipour, 2020). Importantly, this distance measure assumes that the features are normally distributed. The *Mahalanobis distance* operates under the same parametric assumptions (i.e., feature normality) but also accounts for possible correlations between embedded features (Mahalanobis, 1936). The Mahalanobis distance demonstrated utility for OOD uses across a variety of deep learning tasks, including COVID-19 lesion segmentation on CT (González et al., 2022), liver organ

segmentation on MRI (Woodland et al., 2023), and cardiac segmentation on MRI (Arega et al., 2025). The cosine similarity distance is a non-parametric approach that treats embedded test and train features as vectors, for which the cosine angle between is calculated. The *cosine similarity distance* demonstrated superior detection of OOD data compared to the Mahalanobis distance in cases of images with unexpected orientations (e.g., axial vs. coronal/sagittal views of CT images) (Zamzmi et al., 2024) and in the context of optical coherence tomography (Araújo et al., 2023). In these medical image OOD detection investigations, the embedded feature methods were generally superior to alternative OOD methods and other, more generalized UQ methods (e.g., Monte Carlo dropout).

Despite the reported performance gains, current embedded feature-based measures hold several limitations. For example, they lack statistical interpretation because distance measures are typically applied on unitless and abstracted feature embeddings. Another limitation is that their implementation may require feature dimensionality reduction, which potentially removes valuable information. Lastly, they may lack sensitivity since all embedded features are incorporated in the measurement, despite the understanding that some features may be encoded as noise (Samek et al., 2017). These limitations of embedded feature-based OOD detection methods will be further discussed and technically addressed by the new approach presented in Chapter 2.

### 1.4.2 In-Distribution Uncertainty Quantification

ID uncertainty targets the uncertainty present in test samples that resemble the model's train data and arises from various aleatoric and epistemic sources. Deep learning model outputs with high ID uncertainty should not be trusted because these outputs are associated with poor model fitting and/or high input data noise. Broadly speaking, ID uncertainty approaches can be categorized into four main categories: Deterministic-, Bayesian-, Ensemble-, and Test-Time Augmentation-based methods. Figure 4

provides a schematic of each family of ID UQ methods. While some of these methods claim to specifically target either aleatoric or epistemic uncertainty, these uncertainty sources inevitably become entangled during the training process (Valdenegro-Toro & Mori, 2022). For instance, a large amount of data noise (aleatoric uncertainty) will impact model fitting accuracy (epistemic uncertainty). While some uncertainty disentanglement methods have been proposed (Valdenegro-Toro & Mori, 2022), these studies have been restricted to the natural image domain and may not directly translate to medical image problems.

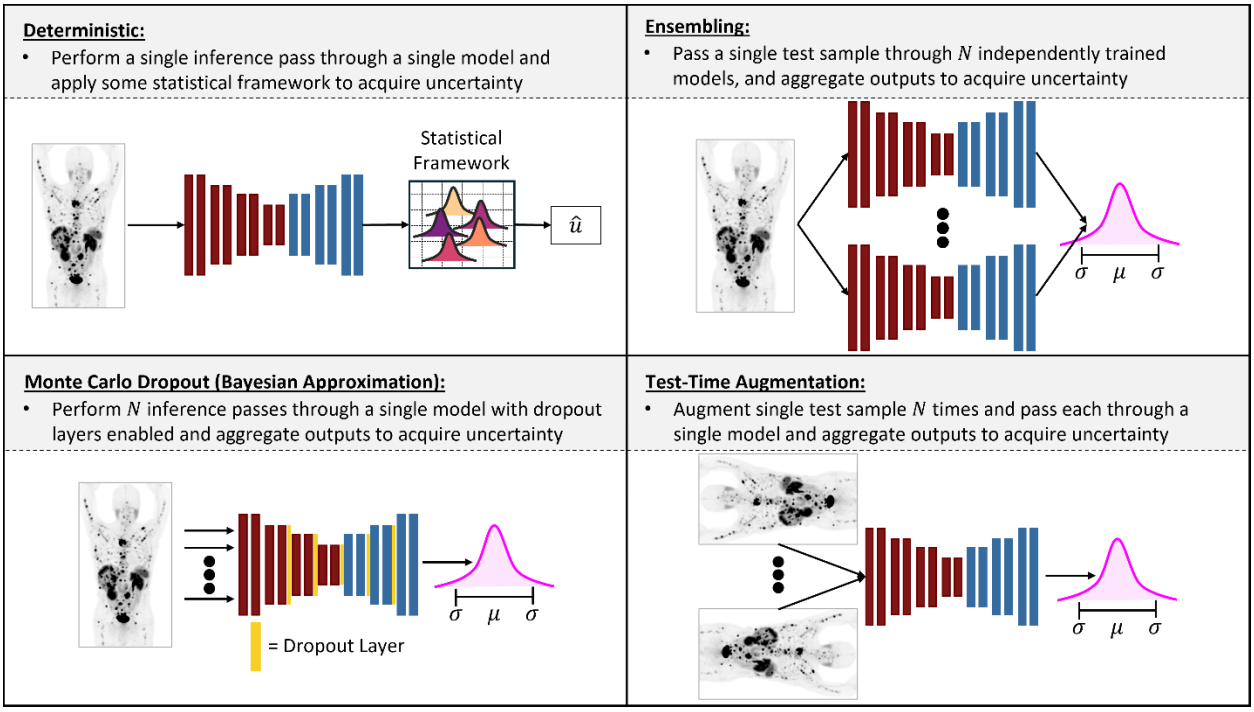



*Figure 4 Schematic describing the different families of ID UQ approaches.*

**Deterministic methods** derive an uncertainty measure from a single inference pass through a single model. This broad category encompasses a wide range of approaches, many of which are computationally efficient, beginning with simple approaches based on standard model outputs. For example, using the maximum softmax probability (MSP) output from a trained model is a deterministic UQ method, where it is assumed that low probability outputs are associated with higher uncertainty and

vice versa. MSP often serves as a baseline measure to test other measures against (DeVries & Taylor, 2018; González et al., 2022). Other methods further process the MSP outputs into alternative uncertainty measures, such as probability entropy across classes (Jungo et al., 2020).

Probability calibration is a slightly more involved deterministic approach that seeks to address the main catalyst for UQ methods: poor model calibration. Most commonly, this is accomplished through a post hoc model re-calibration method, where outputs from a trained model are re-scaled to achieve better calibration and used directly for UQ (Rousseau et al., 2021). Alternatively, calibration-aware training adds a calibration regularization term in the model's loss function during training to dynamically obtain a better-calibrated model (Murugesan et al., 2023).

Conformal prediction is another post-hoc deterministic-based approach that adopts a statistical framework to produce output prediction sets. These sets are defined such that they are guaranteed to include the true outcome according to a user-defined confidence level (Shafer & Vovk, 2008). However, a limitation of both conformal prediction and other post hoc calibration-based approaches is the requirement for a separate, hold-out calibration dataset.

Beyond post hoc approaches, there are deterministic UQ methods that require specialized model and training strategies. For example, learned uncertainty approaches augment the standard model loss with an uncertainty term that maximizes the correlation between itself and model error (Kendall & Gal, 2017). In turn, this learned uncertainty term can be used directly for UQ. Another example is evidential deep learning, which frames the learning process using the Dempster-Shafer Theory of Evidence (Dempster, 1967; Shafer, 1976). For a single prediction, this framework simultaneously assigns belief masses to each prediction class and a single uncertainty mass based on the predicted evidences, which are derived from the model's logit output. Prediction, belief, evidence, and

uncertainty masses are configured such that when evidence across all prediction classes is 0, then overall uncertainty is maximized to be 1 (Sensoy et al., 2018).

**Bayesian methods** seek to learn the probability distribution of individual model parameters. Specifically, Bayesian Neural Networks (BNN) attempt to learn the posterior distribution of the model parameters given the train dataset. Knowing this distribution allows for the expression of the probability distribution of model predictions, from which the uncertainty measure can theoretically be derived. For many deep learning models, especially those for medical image analysis, it is often infeasible to describe either of these distributions exactly due to the requirement of incorporating intractable integrals across parameter space. However, approximations can be used to mimic the output behavior of BNNs.

The most common Bayesian approximation method is the **Monte Carlo dropout (MCDO) method**, which inserts stochastic layers within a model that randomly allows information to pass according to a set probability value (i.e., dropout layers) (Gal & Ghahramani, 2016). Forward passing a single test sample multiple times through a trained model with dropout layers engaged has previously been shown to approximate the output behavior of a BNN (Gal & Ghahramani, 2016). These multiple forward passes construct a prediction distribution from which an uncertainty measure can be aggregated. The MCDO method requires relatively low computation at inference time; however, it typically requires the model to be trained with dropout layers engaged in the same manner as at inference time, which may prompt the need for model re-training. Challenging this constraint, recent work has demonstrated that the use of inference time-only dropout layers yields comparable uncertainty performance when compared to using dropout layers during both model training and inference (Ledda et al., 2023). However, it is unclear whether this can be directly translated to medical image problems.

**Model ensembling (ME) methods,** originally presented in (Lakshminarayanan et al., 2017a), train multiple, randomly initialized models on the same dataset. Given the random parameter initialization and the complex loss landscapes of deep learning models, it is assumed that each independently trained model will optimize to a unique loss optimum, resulting in a diverse set of predictions across models. Some aggregation across model outputs (e.g., variance or class-wise entropy) is then used for predictive uncertainty. Ensembling methods are attractive due to their simplicity and minimal amount of parameter tuning. However, a drawback of this approach is the high computation cost of training and inferring across multiple models. This is especially concerning for medical image tasks, where the standard use of high-dimensional input data imposes high training time burdens.

**Test-time augmentation (TTA) methods** utilize a single trained model and pass several augmented versions of a single test sample through the model, each lending a unique model prediction (Seçkin Ayhan & Berens, 2018; G. Wang et al., 2019). The distribution across model predictions is aggregated into a predictive uncertainty measure. Investigations of this method are almost entirely restricted to medical image applications where data augmentation methods are commonly used during model training to mitigate overfitting. Test time augmentation holds several practical advantages, such as low computational demand and post hoc implementation. Moreover, since it does not require any model modifications, it can even be applied to commercial products where access to the model may be restricted.

## Implementation in Medical Image Analysis

Several studies have implemented ID UQ methods across various medical image analysis tasks. Among these medical image studies, many employ deterministic-based approaches. Reporting the MSP as a baseline is common (Berger et al., 2021; DeVries & Taylor, 2018; González et al., 2022). The entropy

of predicted probabilities has also been used for UQ in structure delineation tasks for PET, CT, and MRI data (Diao et al., 2022; Mehrtash et al., 2020). Introducing model calibration regularization during model training yielded better-calibrated probability outputs across various tasks, such as automated cardiac diagnosis from cardiac MRI volumes and MRI brain tumor segmentation (Murugesan et al., 2023). A learned uncertainty framework was investigated to segment challenging brain tumors on MRI, and the associated uncertainty measure demonstrated a strong correlation with segmentation quality (McKinley et al., 2021). Ghesu and colleagues applied a novel model derived from the Dempster-Shafer theory of evidence that was formulated to output uncertainty values for pathology classification on chest x-ray images, metastasis classification on kidney ultrasound images, and tumor detection on brain MRIs (Ghesu et al., 2021).

The MCDO technique, as an approximation to a BNN, is the most widely implemented UQ method within the medical image domain and has been investigated in various image modalities and analysis tasks (Lambert et al., 2024b). For example, MCDO has been used to flag poorly segmented structures in mammogram images (Klanecek et al., 2023b), brain MRIs (McClure et al., 2019; Mehrtash et al., 2020, Nair et al., 2020b), and prostate and cardiac MRIs (Mehrtash et al., 2020). A very limited number of sensitivity studies on MCDO parameters have been investigated, such as the ideal placement of dropout layers for natural image problems (Kendall et al., 2015) and the optimal number of Monte Carlo samples (Murase et al., 2024) and dropout probability (Camarasa et al., 2020; Murase et al., 2024) for medical image problems. However, it is unclear how the results of these sensitivity studies may translate to different image modalities and analysis tasks.

ME techniques have also been extensively studied for medical image applications. Interestingly, in addition to acquiring uncertainty measures derived from the prediction distribution across ensemble samples, ME has been shown to acquire better calibrated probability outputs when segmenting organs

on MRI (Mehrtash et al., 2020) and detecting the presence of COVID-19 on chest radiographs (Asgharnezhad et al., 2022). The computational demand for ME is a major obstacle in its implementation in medical imaging, where training multiple models may not be practically possible. Several works attempt to mitigate this concern. For example, a model configured with multiple decoder paths for histology image classification and predictions across paths were aggregated into an uncertainty measure (Linmans et al., 2020). Another work, termed *Layer Ensembles*, configured a tumor segmentation model with prediction outputs placed at each decoder layer (Kushibar et al., 2022). An uncertainty measure was defined using the stability of predicted segmentations across layers, where segmentations that demonstrated minimal change across outputs were considered to have low uncertainty, and vice versa. However, since an uncertainty value can be derived from these modified ME methods (i.e., using multiple decoder paths or *Layer Ensembles*) in a single, deterministic inference pass, they may also be defined as deterministic uncertainty methods. Still, like other ME methods, the specialized model architectures used in these approaches may introduce additional and costly memory constraints.

Despite being less studied than MCDO and ensembling techniques, the TTA method was specifically developed for and has been explicitly studied within the medical image domain. In (G. Wang et al., 2019), spatial transformation augmentations were applied at test time to generate uncertainty measures that corresponded better with the segmentation performance of brain structures in MRI data than dropout-based techniques. Applying intensity shifts in addition to spatial shifts has also been explored, and the resulting uncertainty measure demonstrated utility in detecting incorrectly classified voxels for brain tumor segmentation on MRI (Ballestar & Vilaplana, 2021). The TTA technique is very natural to implement for medical image tasks because data augmentation is widely employed for these tasks during training to avoid overfitting. However, it remains unknown what type and how many augmentation transformations are optimal for generating uncertainty measures.

Despite this rich array of UQ work for medical image applications, several limitations remain. For example, most previous studies lack strong quantitative comparisons between uncertainty measures. In part, this is due to a lack of quantitative assessments that could facilitate such comparisons. Another limitation is the high computational cost associated with most methods, where additional trained models (i.e., for ME) or multiple inference passes (i.e., for MCDO and TTA) may be necessary. Another concern with the presented methods is that they inevitably will change the output of a previously trained model, which is undesirable if one seeks to augment a currently trained and validated model with UQ. Consequently, the model accuracy cannot be ensured upon implementing current commonly used UQ methods, and either model re-training and/or re-evaluation is required. Lastly, UQ remains unexplored for a variety of essential clinical tasks. Dedicated UQ studies should be performed for specific configurations of clinical tasks, pathologies, and imaging modalities to ensure the benefits of UQ extend to all clinical scenarios.

## 1.5 Deep Learning-based Medical Image Segmentation

### 1.5.1 Technical Overview

Anatomical and pathological structure segmentation is a critical medical image analysis task that involves assigning a structure label to each voxel within a region of interest for a given image. The resulting voxel-wise label map may be referred to as a segmentation mask. Attaining accurate structure segmentation masks is necessary for a variety of clinical workflows, such as defining tissues to target and to avoid during radiation therapy (Boldrini et al., 2019). Segmentations also enable a broad array of advanced imaging analytics, including the extraction of volume-specific imaging biomarkers to advance personalized care. For example, the extraction of biomarkers from tumor masks has been used to predict and assess patient (Lokre et al., 2024; Santoro-Fernandes, et al., 2024b; Schott et al., 2023; Weber et al.,

2021) and tumor (Santoro-Fernandes et al., 2025) response to treatment. Other applications include the use of segmentation masks to enhance the diagnosis and assessment of cardiac dysfunction (Song et al., 2022) or for volumetric body composition analysis to improve preventative care (Mai et al., 2023). Despite their widespread clinical utility, structure segmentations may not be routinely incorporated into clinical workflows primarily because they are tedious and time-consuming to acquire manually. However, advances in computational medical imaging methods have introduced automated segmentation methods, which hold great potential to enable more extensive adoption of medical image segmentation in clinical care.

Current automated segmentation is largely driven by deep learning-based methods. These methods have demonstrated superior segmentation performance over preceding approaches (Y. Xu et al., 2024), including Otsu thresholding (Chan & Vese, 2001), active contours (Chan & Vese, 2001), level sets (Cremers et al., 2007), and atlas-based (Kalinic, 2009) approaches. The performance of deep learning methods has even met or surpassed that of expert human image readers across a variety of segmentation tasks and exhibits less variability (Shin et al., 2020; Webb et al., 2021). Thus, deep learning-based methods offer automatic, accurate, and standardized medical image structure segmentation.

The workhorse for deep learning-based segmentation methods is the convolutional neural network (CNN) (Minaee et al., 2022). The long-standing foundational CNN for biomedical segmentation is the U-Net architecture (Ronneberger et al., 2015). Key components involved with training a U-Net CNN are depicted in Figure 5. An image is fed into the CNN, which consists of a series of convolutional blocks. Each convolutional block contains several learnable convolutional kernels, which are trained to extract relevant image features. Often, a CNN consists of two main architectural parts—an encoder and a decoder. The encoder is a series of convolutional layers that extract image features (via kernels) while

sequentially downsampling the features to a smaller spatial dimension. The last layer in the encoder has the smallest spatial dimension and is referred to as the bottleneck layer. The decoder follows the bottleneck layer and is similarly a series of convolutional layers; however, after each layer, the extracted features are upsampled until the spatial dimensions match that of the input image, where the number of features matches the number of segmentation label classes. The extracted features between layers are referred to as embedded features. In a U-Net architecture, skip connections are inserted between opposing encoder and decoder layers, which preserve semantic detail through the model and support gradient backpropagation.

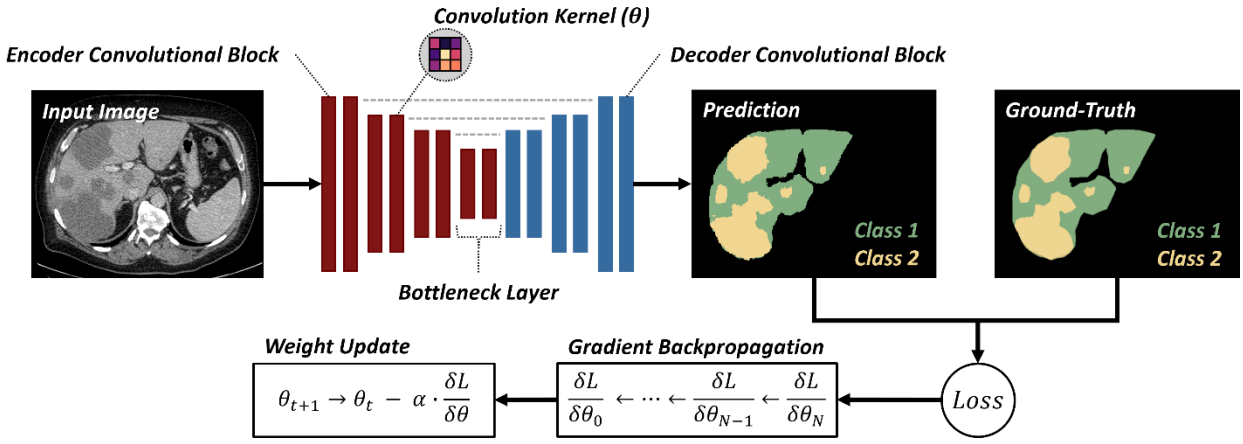



*Figure 5 Schematic describing critical steps and components for training a deep learning medical image segmentation model.*

Model predictions are made from the final model features, sometimes referred to as logits. These are then compressed into probability values via the *softmax* activation function. Class-wise probabilities are converted into one-hot encoded predicted segmentation masks via the *argmax* function, which assigns each voxel the segmentation label with the highest associated probability.

The predicted mask is then compared to the ground-truth mask via a loss function, which is often defined using some metric that quantifies the amount of binary region overlap, such as the Dice

coefficient (Dice, 1945), and/or some metric that evaluates the agreement between predicted probabilities and binary ground-truth labels, such as cross-entropy (Jadon, 2020). The gradients of each learnable parameter in the CNN are then acquired by backpropagating the loss backward into the network using the chain rule. Each learnable parameter, or weight, is then updated according to the gradient with respect to the loss and some set learning rate ($\alpha$). This process is repeated until adequate model performance is achieved or the maximum number of learning iterations (i.e., epochs) is reached.

### 1.5.2 Specific Task: Metastatic Lesion Segmentation

Over 90% of cancer patients die due to metastatic disease (Chaffer & Weinberg, 2011), and the management of these patients is often challenged by high and heterogenous disease burdens. Despite these challenges, recent research has demonstrated the critical importance of monitoring individual lesion behavior for more effective patient management (Harmon et al., 2017; Kyriakopoulos et al., 2020). To accomplish this, whole body molecular imaging, such as positron emission tomography (PET), is often employed because it delivers a comprehensive overview of disease extent and allows for the extraction of functional and biological disease information (Schwenck et al., 2023). PET images are acquired by injecting a radioactive tracer that binds to specific biological targets, often overexpressed in cancer cells. The tracer decays and emits photons that are detected and are used to reconstruct an image. The result image intensities are converted to standardized uptake values (SUV), which reflect the radioactivity concentration normalized by the injected radioactivity and patient body weight. Consequently, the function information extracted from PET scans can be summarized by a series of statistics (e.g., *mean*, *summation*, *variance*, etc.) drawn from the SUV data, termed *SUV metrics*. Example SUV metrics on an individual lesion are shown in Figure 6.

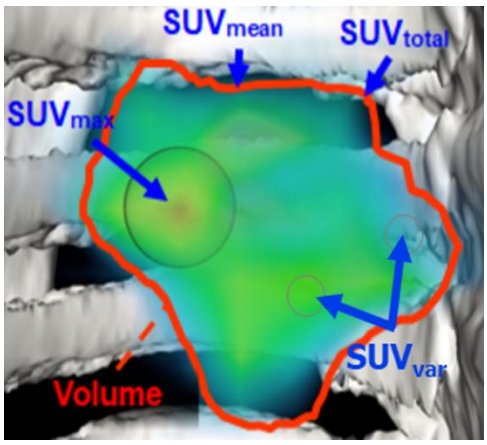



*Figure 6 Schematic summarizing SUV metrics for an isolated lesion on a PET image overlaid on a 3D skeletal rendering from the corresponding CT image.*

Individual lesion SUV metrics have demonstrated valuable clinical utility for the management of patients with metastatic lesions. For example, the extraction of lesion SUV metrics enables the evaluation of disease heterogeneity and the assessment of lesion- and patient-wise treatment response (Lokre et al., 2024; Santoro-Fernandes et al., 2025). Moreover, lesion SUV metrics have been used to build predictive models of patient response to treatment (Lokre et al., 2024; Santoro-Fernandes, et al., 2024b; Schott et al., 2023; Weber et al., 2021). However, the accuracy of SUV metric extraction and subsequent clinical analyses is contingent upon accurate and comprehensive metastatic lesion segmentation.

Several studies have investigated the utility of deep learning models for automatic whole body metastatic lesion segmentation on molecular imaging (Bilic et al., 2023; Li et al., 2020; X. Liu et al., 2021; Schott et al., 2023; Weber et al., 2021; Weisman et al., 2020). However, none of these previous studies have considered the uncertainty of the predicted lesion segmentations. This uncertainty will inevitably propagate and affect SUV metric extraction accuracy and further clinical evaluation. Thus, deep learning-based lesion segmentation uncertainty should be quantified to enhance the reliability of computational methods in the management of patients with metastatic diseases.

In addition to the clinical need for UQ for the metastatic lesion segmentation task, UQ is also critical due to the unique image interpretation challenges involved in this task. Examples of each interpretation challenge are highlighted in Figure 7. First, due to a variety of biological factors, disease presentation can vary greatly within and across patients (Ozaki et al., 2022; Perk, Chen, et al., 2018; Sica et al., 2000). As a result, even tumors in close proximity to each other may differ in presentation and conspicuity (Figure 7a and Figure 7b). Second, healthy physiology may mimic diseased regions, depending on the disease type and imaging modality (Figure 7b) (Kuyumcu et al., 2013; Reinking & Osman, 2009). Lastly, certain benign pathologies may appear very similar to malignant tumors, which are often the target of the delineation task (Figure 7c) (Delbeke et al., 1998; Even-Sapir et al., 2006). These image interpretation challenges introduce added complexity and uncertainty to the learning process of a deep learning model. Thus, UQ methods are especially needed for this task.

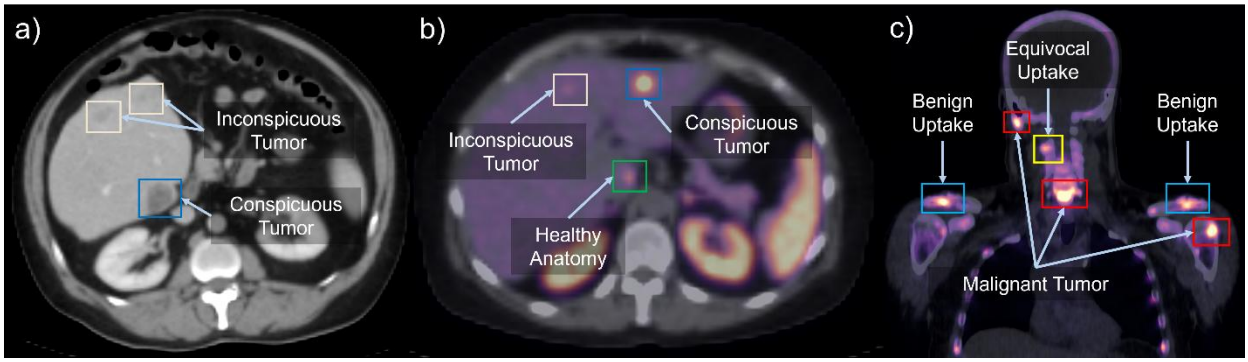



*Figure 7 Example images of patients with metastatic malignancies exhibiting challenging interpretation factors that drive prediction uncertainty. (a) A CT scan of a patient with metastatic NETs contains tumors of varying conspicuity in close proximity. (b) A $^{68}$Ga-DOTATATE PET/CT scan of a patient with metastatic NETs which also vary in conspicuity in addition to a region of healthy anatomy with high SUV uptake (i.e., adrenal gland), (c) A $^{18}$F-NaF PET/CT scan of a patient with mCRPC with both malignant and benign tumor with elevated SUV uptake.*

Only a limited number of studies have previously investigated UQ for metastatic lesion segmentation tasks, however, these have been limited to specific anatomical regions such as the brain (Mehta et al., 2022; Nair et al., 2020). Studies on UQ for primary lesion segmentation are more common

and include tasks such as lung cancer segmentation on CT (Maruccio et al., 2024) and PET/CT (Kang et al., 2024), head and neck cancer segmentation on PET/CT (Huynh et al., 2024; Ren et al., 2024; Sahlsten et al., 2024) and MRI (Ren et al., 2024), or liver cancer segmentation on CT (Hu et al., 2024) and MRI (Hu et al., 2024). Thus, UQ for segmentation of multiple lesions within a single patient volume spread throughout the body has not been previously studied. Due to its under-exploration in this space, in addition to the unique clinical and imaging challenges that demand UQ implementation, the whole body metastatic malignant lesion segmentation task is the primary clinical application for investigating and developing UQ methods in this thesis.

## 1.6 Approach and Aims

In this thesis, **we hypothesized that UQ methods for medical image analysis applications will enhance the accuracy, reliability, and utility of deep learning model outputs intended to inform clinical decision-making**. Investigations into this hypothesis were conducted using the metastatic lesion segmentation task as the main application. Therefore, **the main goal of this work was to investigate existing UQ methods for the metastatic lesion segmentation task, consider their limitations, and propose novel UQ approaches to overcome these limitations and improve UQ utility.** Ultimately, **this work will support a generalized and comprehensive UQ framework that provides an essential, and currently lacking, foundation of trust in the use of deep learning methods to inform clinical decision-making, depicted in Figure 8 (top)**. Since ID UQ methods alone do not sufficiently capture OOD uncertainty (as described in Section 1.4), this framework proposes the evaluation of ID uncertainty independently from, and contingent upon, OOD uncertainty evaluation. Here, only images deemed to be ID are allowed to be inferred upon by the model, whereas the processing of OOD images is halted and

flagged for physician review. The work supporting the components of this comprehensive UQ framework is partitioned into the following two Specific Aims and four total sub-aims:

**Specific Aim 1:** Out-of-distribution uncertainty quantification for deep learning-based medical image analysis

**SA1a:** To assess established embedded feature-based OOD detection methods

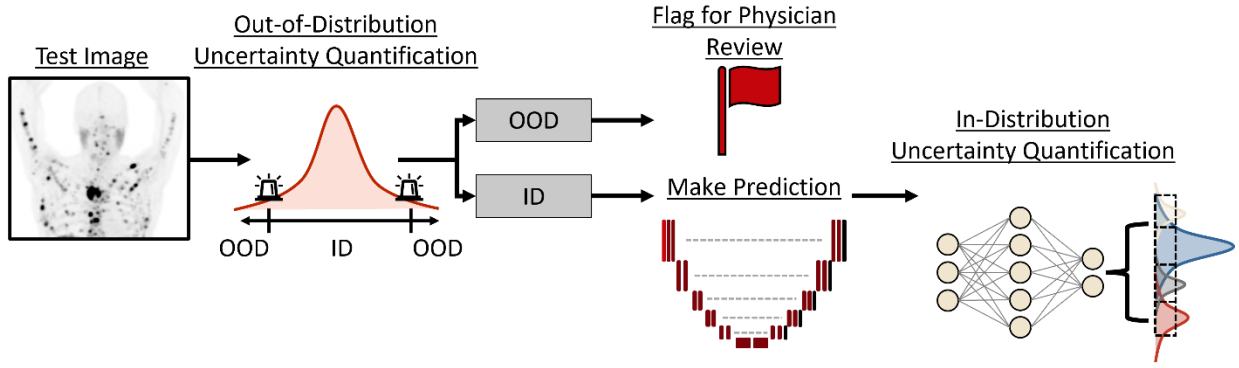
**SA1b:** To investigate a novel information theory and embedded feature-based approach for OOD UQ

**Specific Aim 2:** In-distribution uncertainty quantification for deep learning-based medical image analysis

**SA2a:** To assess established ID UQ methods for the metastatic lesion segmentation task on whole body PET/CT

**SA2b:** To investigate a novel localized gradient-based approach for ID UQ

***General Approach:***




***Thesis Aims:***

| **Specific Aim 1** Out-of-Distribution Uncertainty Quantification | **SA1a:** Assessment of established OOD detection methods<br>**SA1b:** Investigation of a novel, information theory-based approach for OOD Detection | **Specific Aim 2** In-Distribution Uncertainty Quantification | **SA2a:** Assessment of established ID UQ methods<br>**SA2b:** Investigation of a novel, localized gradient-based approach for ID UQ |
|---|---|---|---|

*Figure 8 (Top) Schematic demonstrating the overall approach of comprehensive UQ for medical image analysis tasks. (Bottom) The organization of the thesis aims which spans investigations of both OOD and ID UQ.*

This thesis makes key contributions to the proposed comprehensive UQ framework through the investigation of its Specific Aims and Sub-Aims (SA). Work dedicated to **Specific Aim 1** supports the OOD

component of the comprehensive UQ framework. As a part of **SA1a**, established embedded feature-based OOD detection methods are evaluated. As a part of **SA1b**, the feasibility and performance of a novel information theory-based approach for OOD UQ is explored and quantified. The theoretical background, implementation description, and experimental evaluations of the OOD measures associated with **Specific Aim 1** are provided in **Chapter 2.** Complementing the comprehensive UQ framework, work dedicated to **Specific Aim 2** supports the ID component. As a part of **SA2a**, established ID UQ methods are investigated specifically for the challenging and critical task of metastatic lesion segmentation on whole body PET/CT. Details describing the implementation, quantitative comparisons, and implications of ID UQ methods are provided in **Chapter 3**. Through investigating **SA2b**, an ID UQ approach utilizing a trained and localized model's gradient space was explored. The introduction, evaluation, and discussion of the associated novel ID UQ measure are provided in **Chapter 4**. Finally, the broader implications and potential future work related to these aims and technical developments of this thesis are provided in **Chapter 5**.

# 2 Development of an Information-based OOD Measure

This chapter addresses **Specific Aim 1** by first implementing and evaluating existing embedded feature-based methods for medical image OOD detection (**Specific Aim 1a**). A novel information theory-based OOD detection measure is concurrently introduced, evaluated, and compared to existing OOD measures (**Specific Aim 1b**). This novel embedding approach addresses several of the limitations of established embedded feature-based OOD detection measures; namely, they require dimensionality reduction, they encode noise in the measurement process, and, most notably, they yield insufficient OOD data detection sensitivity for relevant medical image data shifts. The introduced OOD detection measure addresses these limitations by using information bottleneck theory to optimize and quantify the amount of allowable feature information that can be shared from the model's train images. A version of this work titled "Information Bottleneck-Based Feature Weighting for Enhanced Medical Image Out-of-Distribution Detection" was originally presented and published as a conference proceeding at the *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE)* workshop at the *27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)* in Morocco (Schott et al., 2025a). The expansion of this conference work, presented in this chapter, has been submitted for journal publication in *Physics in Medicine and Biology.*

## 2.1 Introduction

The capture of OOD uncertainty via OOD detection methods is a critical component for the robust and safe deployment of clinical deep learning models. As discussed in Section 1.4.1, several OOD detection methods have been investigated for medical image applications, most of which are categorized as embedded feature-based approaches. These methods, however, suffer from several limitations

including the necessity for feature down sampling, inability to account for feature noise, and lack of statistical interpretability. An alternative and unexplored approach to OOD detection is to quantify the amount of embedded feature information from the train data that can be shared with the test data while minimally perturbing the test data output. If the test data is similar to the train data, then a large amount of sharing can take place with minimal effect on the model output. On the other hand, if the test data is dissimilar to the train data, then sharing information from the train data will deteriorate the test data output. To accommodate this approach, a process is needed to quantify both the amount of train information that can be shared with a given test data sample and the effect of that sharing on the model output.

For this purpose, we propose the adaptation of an information bottleneck optimization process. This was initially introduced as a mechanism to reduce the complexity of an information system by the removal of unnecessary information via the insertion of randomly sampled noise such that the model performance is maintained (Tishby et al., 2000). Given a deep learning model, this process can be used to reduce the complexity of selected feature embedding spaces. If noise sampled from a model's train data, rather than noise from a purely random distribution, is inserted into the embedding space, then this optimization process will facilitate quantification of the amount of allowable train information to be shared with a given test data sample. Doing so will enable post hoc OOD detection, which does not require dimensionality reduction, is statistically interpretable, and will intrinsically account for embedded feature noise. Information bottlenecks have previously been used for deep attribution mapping (Schulz et al., 2020; Zhmoginov et al., 2020) but have not yet been utilized for OOD detection.

In this work, we explored the utility of an information bottleneck optimization process for OOD detection, and we denote the associated OOD measure as the *InfoOOD measure*. We evaluated the benefit of this novel OOD detection approach for the metastatic liver tumor segmentation task, where

we assessed the InfoOOD measure's performance at detecting data with simulated CT artifacts and the correlation between the OOD measure and model segmentation performance.

## 2.2 Methods

### 2.2.1 Information Bottleneck Implementation

#### Mathematical Framework

In this work, information bottleneck optimization was implemented on a trained deep learning model in a post hoc manner to accommodate train data feature sharing. To this end, we adapted the information bottleneck implementation described in K. Schulz et al. (Schulz et al., 2020). An information bottleneck block was inserted after the selected model layer. Within this block, the layer output's feature information is disrupted via the injection of noise according to

$$Z = \lambda(\alpha)R + \big(1 - \lambda(\alpha)\big)\epsilon, \tag{1}$$

where $R$ represents the layer features, $\lambda(\alpha) = \mathrm{sigmoid}(\alpha)$, $\alpha$ is a learnable parameter of the same dimensionality as $R$ inserted at the model layer, and $\epsilon$ is replacement noise defined as $\epsilon \sim \mathcal{N}(\mu_R, \sigma_R^2)$, where $\mu_R$ and $\sigma_R^2$ are the estimated mean and variance of the layer features, sampled from the train data. Since the replacement noise is defined using the model's train data, the optimizable $\alpha$ parameter controls the degree to which the train feature distribution can be shared with the given test sample features via feature interpolation. A depiction of the information bottleneck block is provided in Figure 9.

Information sharing was optimized such that test sample output performance was retained according to the loss function:

$$\mathcal{L} = \mathcal{L}_{model} + \beta\mathcal{L}_{info}, \quad (2)$$

where $\mathcal{L}_{model}$ is the model's standard loss (e.g., cross entropy), $\mathcal{L}_{info}$ describes the mutual information between train and test features, and $\beta$ is set according to the desired trade-off between these two terms. Consequently, this loss function maximizes the shared information between train and test features by minimizing the difference in these distributions while retaining good model performance. The shared mutual information between train and test features ($\mathcal{L}_{info}$) was approximated as

$$\mathcal{L}_{info} = I[R, Z] \cong \mathbb{E}_R\big[D_{KL}[P(Z \mid R) \,||\, Q(Z)]\big], \quad (3)$$

where $P(Z \mid R)$ is the probability distribution of $Z$ given $R$, $Q(Z) = \mathcal{N}(\mu_R, \sigma_R^2)$ is a variational approximation, and $D_{KL}[\cdot]$ represents the Kullback-Leibler-divergence. A detailed derivation of equation 3 can be found in (Schulz et al., 2020).

### The InfoOOD Measure

The total loss in equation 2 holds potential as an OOD measure. Test samples with low loss can be interpreted as having embedded features that are similar to those in the train data, and thus, the train features can be shared with the test features without a deterioration in model performance (i.e., low $\mathcal{L}_{model}$). High losses, then, are attributed to the test sample having features that are dissimilar to the train features, and sharing train feature information deteriorates model performance. In this work, we utilize the total loss from equation 2 as an OOD measure and denote it as the *InfoOOD* measure.

### U-Net Model Implementation

In this work, an information bottleneck was augmented to a fully trained U-Net segmentation model at the model's architectural bottleneck. The information bottleneck was optimized in a post hoc

manner on individual test data samples, where the only learnable parameter in the model was the inserted $\alpha$ parameter. This was a self-contained optimization process, meaning that the optimization was invoked individually on test samples and did not require independent data for training purposes. This process was trained for 30 iterations using the Adam optimizer and a learning rate of 0.5. $\beta$ was set to 1E+03, and all indices in the $\alpha$ parameter were initialized to 5.0 to ensure no feature sharing occurred at the start of the optimization process. The feature-wise Gaussian distributions in equations $1$ and $3$ were sampled from the model's train dataset. To enable use in a deployed setting, the ground-truth in the model loss of equation $2$ was set as the model prediction before initiating the information bottleneck optimization. Thus, the optimization process was encouraged to maintain model performance while maximizing train feature sharing. An ablation experiment was performed on the magnitude of the set $\beta$ parameter.

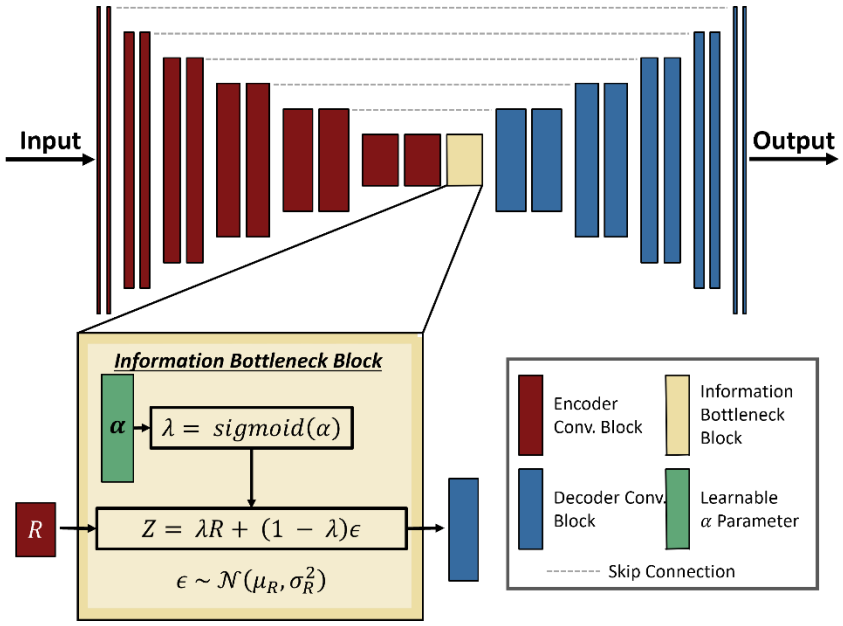



*Figure 9 A schematic, adapted from (Schulz et al., 2020), describing the implementation of an information bottleneck on a U-Net architecture. The information bottleneck block is placed after the U-Net bottleneck layer.*

### 2.2.2 Out-of-Distribution Measures

#### OOD Measures for Comparison

Several embedded feature-based OOD measures were implemented for comparison. In agreement with the InfoOOD measure, each comparison method was implemented on the segmentation model's bottleneck layer, and the reference feature distribution was defined using the model's train data. In addition, all OOD measures were post hoc and were implemented on the fully trained liver organ and tumor segmentation model.

First, the Euclidean distance ($Dist_{Eucl}$) between train and test features was implemented according to

$$Dist_{Eucl} = \|R - \mu\|_2, \tag{4}$$

where $\|\cdot\|_2$ represents the $L_2$ norm. Next, the Mahalanobis distance ($Dist_{Maha}$) was implemented according to

$$Dist_{Maha} = \sqrt{(R' - \mu')^T C^{-1} (R' - \mu')}, \tag{5}$$

where $R'$, represents the downsampled embedded test features, $\mu'$ represents the downsampled mean of the train features, $(\cdot)^T$ represents the matrix transpose operation, and $C^{-1}$ represents the inverse of the covariance matrix sampled from the sampled train features. Downsampling via average pooling of the embedded features and adding a regularization term of $1\mathrm{E}-05$ to the diagonal of $C$ was necessary to accommodate the matrix inversion operation. The $Dist_{Maha}$ accounts for possible correlations between features, and as a result, is generally assumed to yield greater utility over $Dist_{Eucl}$. Lastly, the cosine similarity distance ($Dist_{Cosine}$) was implemented according to

$$Dist_{Cosine} = 1 - \frac{R \cdot \mu}{\|R\|_2 \|\mu\|_2}. \quad (6)$$

The cosine similarly distance subtracts the cosine angle between the embedded test features ($R$) and the average of the train features ($\mu$) from one, resulting in a range between 0 and 2. A $Dist_{Cosine}$ measure of 0 implies the two feature vectors are completely aligned, while a $Dist_{Cosine}$ measure of 2 implies these two vectors are antiparallel. Thus, we assume $Dist_{Cosine}$ measures closer to 2 are associated with OOD data.

Due to the absence of computational constraints, both the $Dist_{Eucl}$ and the $Dist_{Cosine}$ measures utilized the full feature space and did not utilize dimensionality reduction like the $Dist_{Maha}$ measure. For each OOD measure (including the InfoOOD measure), larger magnitudes are associated with higher likelihood of the associated image to be OOD.

### 2.2.3 Segmentation Model

A three-dimensional segmentation model was trained in-house to segment the liver organ and metastatic liver tumors using the nnUNet repository (Isensee et al., 2021). The nnUNet model has demonstrated highly competitive results on various segmentation tasks, making it well-suited to augment with and test OOD detection algorithms. The model was trained for 1,000 epochs with a batch size of 2 and using the sum of the Dice coefficient and cross-entropy as the loss function. Instance normalization was used between each convolutional layer. The input image patch size was $[128 \times 128 \times 128]$ voxels. Data augmentation was applied during training using additive Gaussian noise, Gaussian blurring, contrast adjustment, low resolution simulation, gamma augmentation, rotation, scaling, and mirroring. Model training took place on an Nvidia RTX Titan GPU workstation with 24 GB of memory.

### 2.2.4 Datasets

Datasets utilized in this work consisted of abdominal computed tomography (CT) scans of patients with metastatic liver tumors. Selection of this dataset was motivated by the high incidence of liver metastases across many common malignancies (Clark et al., 2016) and the critical role of medical image assessment in managing the disease (Freitas et al., 2021), including deep learning-based segmentation approaches (Vorontsov et al., 2019). The base segmentation model was trained using $N = 157$ scans acquired from the publicly available Colorectal Liver Metastases (CRLM) dataset (Simpson et al., 2024). All images in this dataset were acquired following the same acquisition protocol and parameters. Consequently, deviations in the model's train feature distributions due to differences in scan type were minimized. A validation data split was defined using $N = 40$ images from the same CRLM dataset and was used as test data for segmentation model performance and OOD detection evaluations. An addition test comprising of $N = 131$ abdominal CT images from the publicly available LiTS Liver Tumor Segmentation challenge (Bilic et al., 2023) was acquired for further OOD detection evaluation. Unlike the images from the CRLM dataset, these images were acquired using non-standardized image acquisition protocols.

### 2.2.5 Data Preprocessing

Prior to model training, all images were normalized to zero-mean-unit-variance, resampled to a common 2.5 $mm^3$ voxel spacing, and cropped about the center of the liver to the model patch size. The same voxel spacing and image cropping were applied to test images for all OOD evaluations. Due to the patch-based segmentation model used in this work, cropping to the model patch size ensured only one OOD measure was computed per test image.

### 2.2.6 Experiments

#### CT Artifact Detection

In this first assessment, each OOD measure's ability to detect clinically relevant OOD test data was investigated. Here, the segmentation model's train data defined the in-distribution (ID) data, and CT artifacts were simulated on the CRLM test data to curate OOD data. Simulated artifact data was selected for this purpose because it has been shown to be challenging to detect this type of data using standard OOD measures (González et al., 2022).

#### Artifact Simulation

Using physics-based CT artifacts as OOD test data more closely resembles possible OOD data in deployed clinical settings. Three artifacts were simulated: low dose, sparse views, and rings artifacts. Each artifact was simulated at low, medium, and strong magnitudes to curate near, medium, and far OOD test data. Artifacts were simulated following a filtered back-projection image reconstruction framework (Pan et al., 2009), where simulated projection data was acquired via the Radon transform of the original images and modified for artificial simulation. The inverse Radon transform was then employed on the modified projection data to acquired images with simulated artifacts. A high-level overview of the artifact simulation process and an example image across artifacts and magnitudes is shown in Figure 10. All artifact simulation and subsequent analyses were performed on the validation data split from the segmentation model dataset (i.e., the CRLM data).

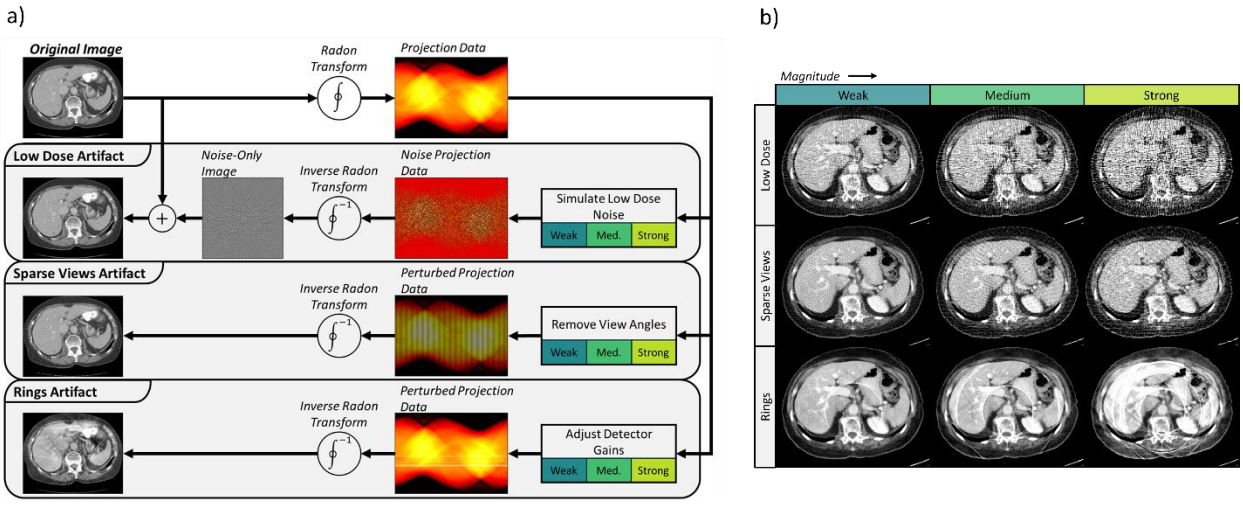



*Figure 10 (a) Process for simulating low dose, sparse views, and rings artifacts on CT test data. (b) Example CT test image with simulated CT artifacts for each artifact type and magnitude.*

Low-dose CT images were simulated using the method established in (Yu et al., 2012). Low-dose noise-only projection data was acquired according to

$$\tilde{\eta}_B \approx \sqrt{\frac{1-a}{a} \cdot \frac{exp(P_A)}{N_{0,A}}} \cdot \epsilon, \tag{7}$$

where $\tilde{\eta}_B$ is the approximate additive low-dose noise projection data, $P_A$ is the original image projection data, $a$ is the dose reduction factor, $N_{0,A}$ is the number of incident photons from the original high dose scan, and $\epsilon$ is a randomly sampled Gaussian distribution defined as $\epsilon \sim \mathcal{N}(\mu = 0, \sigma = 1)$. $N_{0,a}$ was assumed to be 1E04. The inverse radon transform was performed on $\tilde{\eta}_B$ to acquire a low-dose noise-only image, which was added to the original image to simulate the low-dose image with non-random and non-uniform additive noise. Weak, medium, and strong low dose artifacts were defined using $a = 0.75$ (three-quarter dose), $a = 0.50$ (half dose), and $a = 0.25$ (one-quarter dose), respectively.

The sparse views artifact was simulated by removing data from every $n^{th}$ index along the view angle ($\theta$) dimension of the original image's projection data. Weak, medium, and strong view angle artifacts were defined using $n = 4$, $n = 3$, and $n = 2$, respectively.

The rings artifact was simulated by adjusting positive and negative gains of $n$ indices along the detector ($x$) dimension of the original image's projection data. Gain factors were randomly sampled between integers $[g_1,\ g_2]$, where positive gains were applied using these factors and negative gains were applied using $1/[g_1,\ g_2]$. Weak, medium, and strong rings artifacts were simulated using $n = 2$, $[g_1 = 3,\ g_2 = 4]$; $n = 3$, $[g_1 = 4,\ g_2 = 5]$; and $n = 4$, $[g_1 = 5,\ g_2 = 6]$; respectively.

### Artifact Induced Shifts

Shifts induced by the simulated CT artifact data on the segmentation model performance and the InfoOOD measure were assessed. The Dice coefficient for the predicted liver organ and liver lesion segmentation masks and the corresponding InfoOOD measures were reported for each artifact type and magnitude and for the OOD test data with no artifact simulation. Statistical significance was reported between the non-artifact and artifact data using the Wilcoxon rank sum test. Segmentation performance (Dice) was expected to degrade with artifact magnitude, whereas the InfoOOD measure was expected to increase with artifact magnitude.

### OOD Detection Performance

The detection performance of each OOD measure for detecting CRLM test data with each simulated artifact type and magnitude was assessed. Performance using areas under the Receiver Operator Curves (AUCs) was evaluated for detecting each artifact type and magnitude. Bootstrapping using $n = 1{,}000$ samples was invoked to acquire 95% confidence intervals of the AUC values. OOD

detection performance was also reported using the InfoOOD measure from each iteration of the optimization process to observe the behavior of the measure across iterations, capturing the bottleneck optimization process.

## Ablation Study

Within the CT artifact detection experiment, an ablation was performed over the $\beta$ used in the post hoc information bottleneck loss function (equation 2). In this ablation, OOD detection performance was assessed on the CRML test data at the medium magnitude for each artifact using $\beta =$ [1E-01, 1E+00, 1E+01, 1E+02, 1E+03, 1E+04].

## Correlations With Segmentation Model Performance

In this second assessment, we evaluated each OOD measure's correlation with the trained segmentation model's performance. Strong correlations may indicate that an OOD measure can be used as a proxy for model performance in a deployed clinical setting in the absence of ground-truth data. This correlation analysis was performed using the $N = 131$ abdominal CT test images from the LiTS dataset. Since these images were acquired using non-standardized image acquisition protocols, they are expected to be broadly distributed with varying OOD magnitudes, making them more suitable for the correlation analysis. Evaluated model performance metrics included predicted liver lesion detection sensitivity, segmentation Dice coefficient, and the percent difference between ground truth and predicted lesion volumes. These metrics were chosen due to their clinical importance. For instance, accurate lesion segmentation is crucial for radiation therapy planning (Savjani et al., 2022), while liver lesion volume is an established biomarker in disease assessment and prediction (Assouline et al., 2023; Sahu et al., 2017). The Spearman correlation coefficient between individual OOD measures and each of these segmentation performance metrics was reported. To facilitate more homogenous analysis, the negative of the lesion

volume percent difference metric was used such that a strong negative correlation indicated better OOD measure performance for each model performance metric. To observe the benefit of the information bottleneck optimization process on the InfoOOD measure, these correlations were reported at each iteration in the InfoOOD optimization process.

## 2.3 Results

### 2.3.1 CT Artifact Detection

#### Induced Artifact Shifts

The performance of the trained segmentation model on OOD test data across all artifact types and magnitudes is summarized in Figure 11. The segmentation performance degraded with increasing artifact magnitude for each artifact type, as expected. In general, degradations were more pronounced for the liver lesion segmentation compared to the organ segmentation. However, the segmentation performance on each artifact type and magnitude for both liver organ and lesions was significantly inferior to the performance in the ID data (i.e., $p < 0.001$). When considering both liver organ and lesion segmentation performance, segmentation performance deterioration was strongest for the Rings artifact and weakest for the Sparse Views artifact.

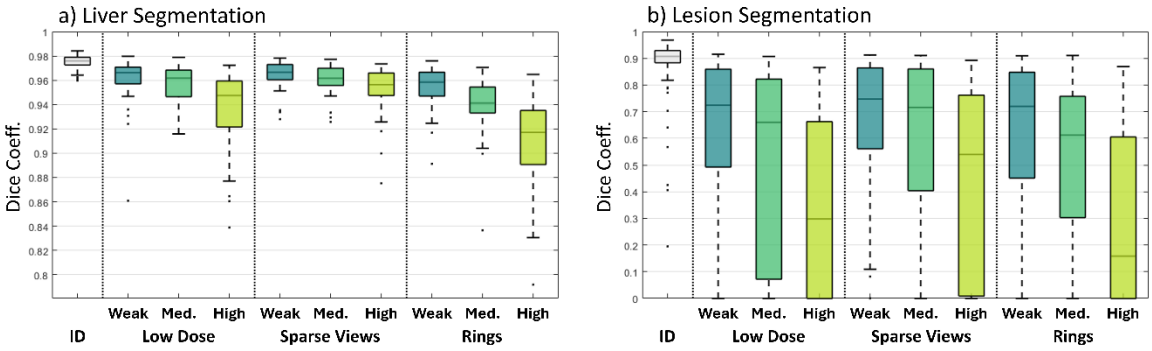



*Figure 11 Dice coefficients of predicted liver organ and liver lesion segmentations on in-distribution (ID) and out-of-distribution OOD test data for each simulated artifact type and magnitude.*

Figure 12 shows the InfoOOD measures across artifact types and magnitudes. The InfoOOD measures for the high magnitude data for each artifact type demonstrated much higher magnitudes than measures for the weak and medium magnitude data. For better visualization, the InfoOOD measures are shown across all artifact magnitudes (Figure 12a) and across only weak and medium magnitudes (Figure 12b). For each artifact type, the InfoOOD measure increased with artifact magnitude. The InfoOOD measures for each artifact type and magnitude were statistically greater ($p < 0.001$) than those in the in-distribution data.

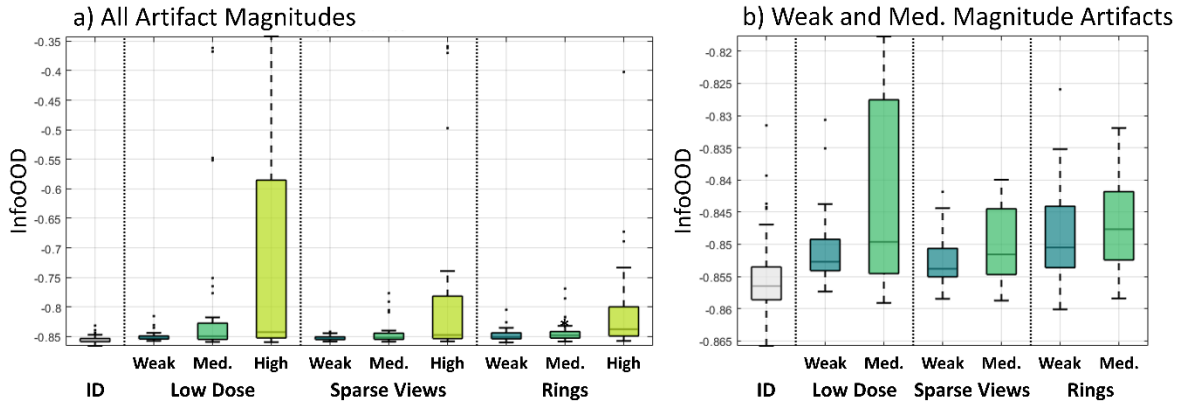



*Figure 12 (a) The InfoOOD measure on ID data and OOD test data across each artifact type and magnitude and (b) across weak and medium magnitudes, for improved visualization.*

## OOD Detection Performance

The OOD detection performance of each OOD measure across all artifact types and magnitudes is summarized in Table 1. The proposed InfoOOD measure achieved superior OOD detection performance across all artifact types and magnitudes, and detection performance increased with increasing magnitudes across artifact types. The InfoOOD measure yielded the worst detection performance for the weak sparse views artifact ($AUC = 0.74$) and the best detection performance for the strong rings artifact ($AUC = 0.93$). All comparison OOD measures failed to adequately detect OOD data, where the worst detection performance was $AUC = 0.38$ for the detecting the weak sparse views artifact data using both $Dist_{Eucl}$ and $Dist_{Cosine}$, and the best detection performance was $AUC = 0.57$ for detecting *strong rings* artifact data using each comparison OOD measure. Still, the detection performance of each comparison OOD measure increased with increasing artifact magnitudes.

*Table 1 AUC scores (↑) with 95% confidence intervals ($[\cdot]$) for detecting OOD data using each of the OOD measures. Detection performance is shown for each OOD data artifact simulation type and magnitude.*

| Artifact Type | Artifact Magnitude | OOD Measure | | | |
|---|---|---|---|---|---|
| | | $Dist_{Maha}$ | $Dist_{Eucl}$ | $Dist_{Cosine}$ | **InfoOOD** |
| Low Dose | Weak | 0.42 [0.32, 0.51] | 0.39 [0.30, 0.48] | 0.39 [0.30, 0.49] | **0.79 [0.72, 0.86]** |
| | Med. | 0.43 [0.34, 0.54] | 0.41 [0.32, 0.52] | 0.42 [0.32, 0.51] | **0.85 [0.78, 0.91]** |
| | Strong | 0.48 [0.37, 0.59] | 0.48 [0.38, 0.59] | 0.49 [0.38, 0.59] | **0.90 [0.83, 0.94]** |
| Sparse Views | Weak | 0.40 [0.30, 0.49] | 0.38 [0.29, 0.48] | 0.38 [0.29, 0.48] | **0.74 [0.67, 0.82]** |
| | Med. | 0.40 [0.31, 0.50] | 0.39 [0.30, 0.48] | 0.39 [0.30, 0.49] | **0.80 [0.73, 0.87]** |
| | Strong | 0.44 [0.34, 0.54] | 0.45 [0.35, 0.54] | 0.45 [0.35, 0.54] | **0.86 [0.79, 0.92]** |
| Rings | Weak | 0.45 [0.33, 0.51] | 0.42 [0.33, 0.51] | 0.42 [0.32, 0.52] | **0.82 [0.73, 0.89]** |
| | Med. | 0.47 [0.37, 0.57] | 0.46 [0.36, 0.56] | 0.46 [0.36, 0.56] | **0.88 [0.81, 0.93]** |

| | | | | | |
|---|---|---|---|---|---|
| | Strong | 0.57 [0.47, 0.66] | 0.57 [0.48, 0.66] | 0.57 [0.47, 0.66] | **0.93 [0.88, 0.97]** |

The OOD detection performance of the InfoOOD measure at each iteration in the information bottleneck optimization process is shown in Figure 13**.** In general, OOD detection performance increased with successive optimization iterations, and detection performance was greatly enhanced from the 0$^{th}$ to the last iteration, with a particular spike of increased performance around the 20$^{th}$ iteration. For example, the detection of images with weak and strong magnitude artifacts increased from $AUC \leq 0.46$ to $AUC \geq 0.74$ and from $AUC \leq 0.68$ to $AUC \geq 0.86$, respectively. Interestingly, the poor performance of the InfoOOD measure at the 0$^{th}$ iteration was still superior to the comparison OOD measures (e.g., $AUC = 0.69$ vs. $AUC = 0.57$ for detecting strong rings artifact data using the comparison methods). For the majority of artifact types and magnitudes, the detection performance saturated around the 25$^{th}$ iteration. The detection performance on a few of the OOD test sets, however, did not fully saturate by the last iteration (e.g., the medium and strong rings artifact data). Regardless of the saturation point, the information optimization process greatly benefited detection performance across artifact types and magnitudes.

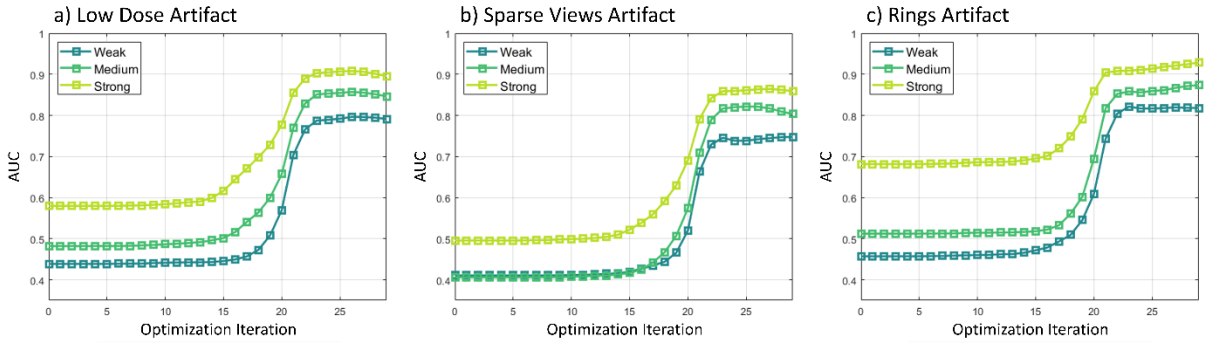



*Figure 13 AUC (↑) detection performance for the InfoOOD measure at each iteration in the post hoc information bottleneck optimization process for detecting a) Low-Dose Artifact data, b) Sparse Views Artifact data, and c) Rings Artifact data, each at weak, medium, and strong magnitudes.*

**Ablation Study**

The detection performance of the InfoOOD measure when using different $\beta$ values in the information bottleneck loss function (equation 2) is shown in Table 2**.** This detection was performed using the medium magnitude data from each artifact type. Using $\beta = 1E+03$ offered superior detection performance across artifact types at this magnitude. Both $\beta$ values less than and greater than this yielded inferior detection performance. However, when considering the confidence intervals, detection performance was not substantially different across $\beta$ values.

*Table 2 AUC scores (↑) with 95% confidence intervals ([·]) for detecting each artifact at the medium magnitude when using different $\beta$ values in the information bottleneck loss function (InfoOOD measure).*

| | Artifact (Medium Magnitude) | | |
|---|---|---|---|
| $\beta$ | Low Dose | Sparse Views | Rings |
| 1E-01 | 0.74 [0.64, 0.83] | 0.70 [0.59, 0.80] | 0.85 [0.78, 0.91] |
| 1E+00 | 0.74 [0.64, 0.84] | 0.70 [0.59, 0.80] | 0.85 [0.78, 0.92] |
| 1E+01 | 0.74 [0.64, 0.83] | 0.70 [0.58, 0.80] | 0.85 [0.79, 0.92] |
| 1E+02 | 0.76 [0.66, 0.84] | 0.71 [0.61, 0.81] | 0.86 [0.79, 0.92] |
| 1E+03 | **0.85 [0.78, 0.91]** | **0.80 [0.73, 0.87]** | **0.88 [0.81, 0.93]** |
| 1E+04 | 0.81 [0.73, 0.87] | 0.75 [0.67, 0.82] | 0.69 [0.60, 0.78] |

## 2.3.2 Correlations With Segmentation Model Performance

The Spearman correlation coefficients ($\rho$) between each OOD measure and segmentation model performance metrics are shown in Table 3**.** The InfoOOD measure achieved superior correlation performance across segmentation performance metrics. For this measure, the strongest correlation was shown for the lesion sensitivity ($\rho = -0.52$) metric, followed by lesion Dice coefficient ($\rho = -0.49$), and

then lesion volume accuracy ($\rho = -0.47$). All comparison OOD measures demonstrated inferior correlation performance across all segmentation metrics. These correlations were relatively poor such that no meaningful trend between OOD measure and segmentation performance was observed (i.e., $\rho \geq -0.09$ for all). Scatter plots depicting the relationship between the InfoOOD measure and each segmentation performance metric are shown in Figure 14**.**

*Table 3 Spearman correlation coefficients (ρ) between each OOD measure and segmentation performance metric. Larger negative numbers indicate stronger correlations. Bold text indicates the strongest correlation coefficient.*

| | Segmentation Metric | | |
|---|---|---|---|
| OOD Measure | Sensitivity | Dice Coeff. | Volume Accuracy |
| $Dist_{Maha}$ | -0.06 | 0.02 | -0.09 |
| $Dist_{Eucl}$ | 0.06 | 0.08 | -0.04 |
| $Dist_{Cosine}$ | -0.05 | -0.08 | 0.04 |
| InfoOOD (Ours) | **-0.52** | **-0.49** | **-0.47** |

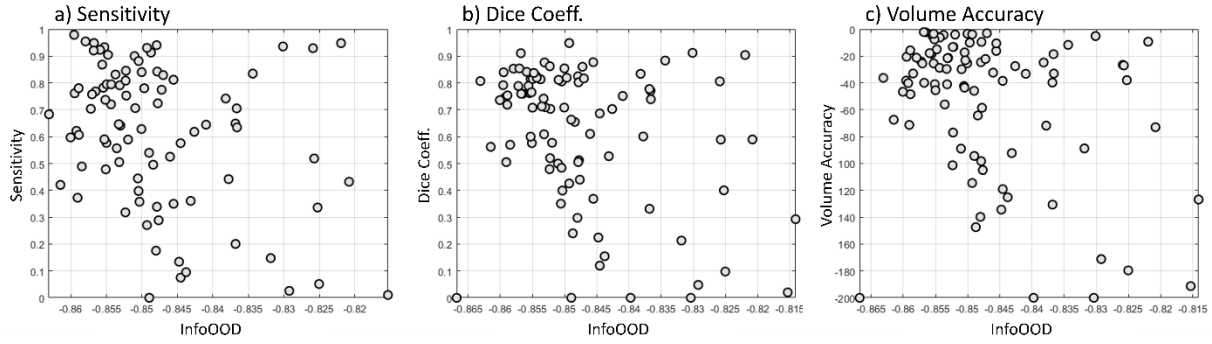



*Figure 14 Scatter plots between the InfoOOD measure and each segmentation performance metric: (a) lesion sensitivity, (b) lesion Dice coefficient, and (c) lesion volume accuracy. Outlier data defined as test image with greater than the 75$^{th}$ percentile InfoOOD measure are omitted from the scatter plots for visualization.*

Figure 7 shows examples of test images that exhibit the strong correlation between the InfoOOD measure and the segmentation metrics (i.e., low InfoOOD measures are associated with good segmentation performance and vice versa). The segmentation model achieved strong performance for

the test image in Figure 15a, where, considering the whole test dataset, the percentiles of the performance metrics were 98.3%, 91.1%, and 74% for lesion sensitivity, Dice coefficient, and volume accuracy, respectively (i.e., very high segmentation performance). At the same time, this test image yielded a low InfoOOD measure which was under the $10^{th}$ percentile (i.e., very low OOD) of all InfoOOD measures in the test data. On the other hand, Figure 15b shows a test image on which the segmentation model performed poorly and yielded segmentation metrics equating to percentiles of 14.5%, 18.5%, and 16.1% for lesion sensitivity, Dice coefficient, and volume accuracy, respectively (i.e., very low segmentation performance). Meanwhile, the InfoOOD measure percentile on this image was 98.4% (i.e., very high OOD).

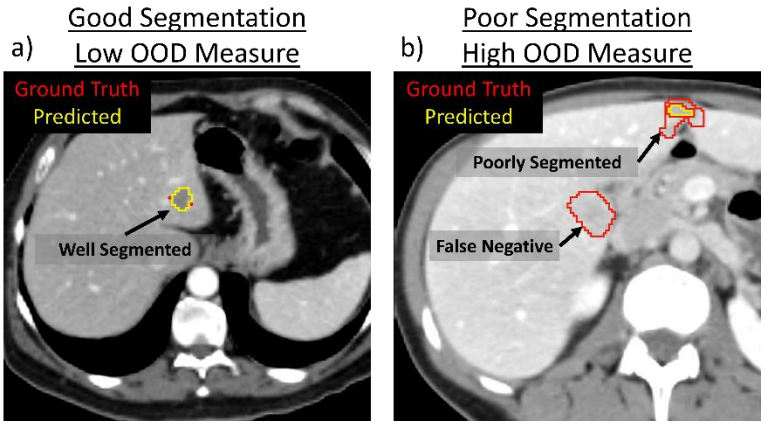



*Figure 15 Example test images that demonstrated a strong correlation between the InfoOOD measure and predicted segmentation performance. The test image in (a) shows an image with good segmentation performance and a low InfoOOD measure. Conversely, the test image in (b) shows an image with poor segmentation performance and a high InfoOOD measure.*

Figure 16 shows the Spearman correlation coefficients between the InfoOOD measure and each segmentation performance metric at each iteration in the information bottleneck optimization process. Overall, these correlations were enhanced with successive iterations, indicating the importance of the optimization process in determining the amount of allowable shared information from the train features with the test features. The optimization process strengthened each correlation from $\rho \geq -0.11$ to $\rho \leq$

$-0.47$. Thus, the information bottleneck optimization process was essential to acquire strong correlations with model performance.

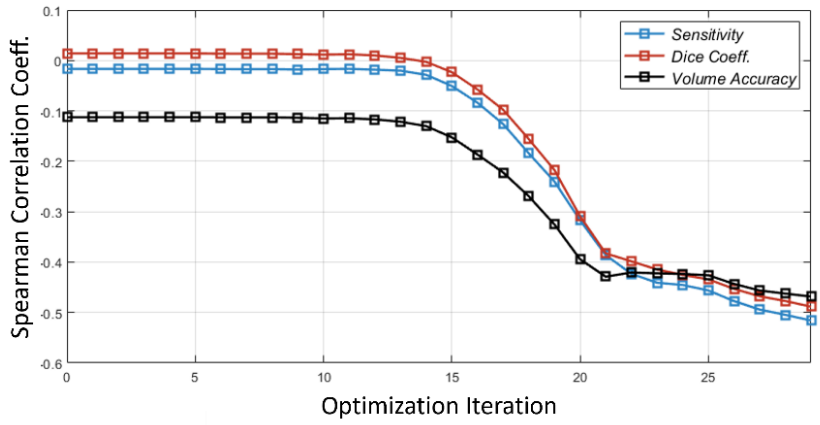



*Figure 16 Spearman correlation coefficients between the InfoOOD measure and each segmentation performance metric across information bottleneck optimization iterations.*

## 2.4 Discussion

In this work, we introduced a novel approach to OOD detection for deep learning-based medical image analysis tasks, which leveraged a post hoc information bottleneck optimization. Specifically, this optimization process was used to quantify the amount of train data embedded feature information that can be shared with test data feature information without deteriorating model outputs. Importantly, this approach is post hoc, meaning, it can be implemented on any fully trained deep learning model, and the optimization is self-contained, relying only on the input test image and the trained deep learning model without requiring separate training data. We showed that the OOD measure associated with this approach, InfoOOD, achieved superior performance over established, feature embedding-based OOD measures for detecting OOD test data, which was defined using simulated image artifacts. We also demonstrated how this approach was necessary to obtain strong correlations between the OOD

measure and model performance. Overall, this work introduces and demonstrates the benefit of an information-based approach for medical image OOD tasks over more standardized feature embedding-based approaches.

The InfoOOD measure offers an alternative OOD measure which acts on the embedded feature space of a trained model. Most previous embedding space methods are categorized as distance-based methods. For example, the work of (González et al., 2022) used the Mahalanobis distance of embedded features for OOD detection. This distance measure is generally preferred over others, such as the Euclidean distance, because it accounts for possible correlations between features. However, due to the high computational demand of this distance measure, the dimension of the embedded feature space is often reduced via average pooling. This dimensionality reduction may remove feature information relevant to OOD detection. In contrast, the InfoOOD measure does not require dimensionality reduction, allowing for all feature information to be considered for OOD detection. Alternative dimensionality reduction techniques, such as Principal Component Analysis, have been proposed for OOD detection (Woodland et al., 2023). While these approaches can be interpreted as isolating representative features for OOD detection, these downsampled features are not directly linked to the model output. The InfoOOD approach can similarly be interpreted as isolating important features, but this "importance" is directly tied to the effect on model output via the post hoc loss function (equation 2). For example, high InfoOOD measures will be dominated by features critical to the prediction output yet different from the train features. Consequently, the InfoOOD measure intrinsically accounts for possible noise in the embedded feature space, which could otherwise decrease the sensitivity of the OOD measurement.

In our first evaluation, we defined OOD data using simulated image artifacts based on image acquisition physics to represent OOD data that may more realistically be found in clinical settings. In contrast, previous works have defined OOD data using less clinically realistic perturbations, such as by

omitting image slices (Zimmerer et al., 2022), by applying affine transformations (González et al., 2022), by applying random parameterized noise perturbation (Schott et al., 2024), or by applying non-random noise from adversarial attacks (Schott et al., 2025a). In part, these perturbations are advantageous due to their implementation simplicity, whereas acquiring actual clinical OOD data can be challenging. However, simulating physics-based image artifacts, as was done in this work, is both straightforward and enables more clinically realistic and relevant OOD quantitative assessments. Moreover, our image artifact simulation approach allows for simulations of varying magnitudes, which enables the user to test the sensitivity of OOD measures on images spanning near and far OOD data.

The performance of the trained segmentation model degraded with each introduced image artifact and magnitude (Figure 11). At the same time, the introduced InfoOOD measure responded in an inverse manner, where measures were greater for stronger artifact magnitudes (Figure 12). This verified the assumption that simulated image artifacts should both degrade model performance and yield higher OOD distances. In agreement with this pattern, the OOD detection performance of the InfoOOD measure scaled directly with artifact magnitude (Table 1 and Figure 13). The detection performance of the comparison OOD measures similarly scaled with artifact magnitude, however, none of these measures achieved adequate detection performance. When evaluating the correlation between each OOD measure and the three segmentation performance metrics, we found that the comparison OOD measures failed to yield useful correlations. Meanwhile, the correlations from the InfoOOD measure were more meaningful (e.g., $\rho = -0.52$ for lesion sensitivity). Figure 15b provides an example where the InfoOOD measure was appropriately high for a test image that the segmentation model failed on, possibly due to the differences in contrast enhancement between this image and the segmentation model train data. While the correlations from the InfoOOD measure were the strongest amongst the tested OOD measures, fluctuations in the data were still present (Figure 14). However, a perfect

correlation between an OOD measure and model performance is not expected since OOD is only one factor that could induce model failures, including other sources of predictive uncertainty (Lambert et al., 2024). Because of this limitation in correlation assessments, it is important to evaluate any OOD measure using multiple assessments, as we have done in this work. Future work should be dedicated towards combining different uncertainty measures (including OOD detection and traditional, ID uncertainty quantification) to acquire better correlations with clinical endpoints.

The comparison OOD measures failed to yield meaningful results for artifact detection and correlation to segmentation performance. In the work of (González et al., 2022), however, the Mahalanobis distance obtained superior OOD detection and correlation performance when compared to other OOD measures. Another work similarly found the Mahalanobis distance yielded meaningful results, however, these were comparable to alternative OOD measures (Anthony & Kamnitsas, 2023). In both of these works, batch normalization was used between model layers. In contrast, our segmentation model utilized instance normalization. It is possible that batch normalization is more conducive to OOD detection using standard distance measures than instance normalization, where embedded features are more tightly constrained. It is critical, however, to investigate OOD detection using models with instance normalization due to the associated performance boost this normalization provides, especially when using small batch sizes during training (Kolarik et al., 2020), as is often done for medical image tasks. Alternative feature normalization strategies or training approaches to better conform the feature space (Zamzmi et al., 2024) should be investigated for OOD detection. The poor results of the comparison OOD measures may also be attributable to the type of OOD data targeted in this work. Non-clinical image artifacts (e.g., spatial transformations) have been shown to be difficult to detect using OOD measures (González et al., 2022). Our work was the first to define OOD data using clinically realistic and physics-based image artifacts. Our results indicate that established, distance-based OOD measures may similarly

not be sensitive to this type of OOD data. Nevertheless, like the other embedded feature methods, the InfoOOD measure did not achieve meaningful results without optimization (at iteration 0). Thus, the optimization process appeared to be critical for these OOD detection tasks.

The InfoOOD measure holds potential with and without optimization, where each use offers a different rationale as an OOD measure. Without optimization, this measure can be interpreted as quantifying the amount of shared information between train and test features. With optimization, the InfoOOD measure can be interpreted as quantifying the amount of shareable information from the train features with the test features such that model performance is not significantly degraded. While using the InfoOOD measure without optimization shows slight promise for OOD detection (e.g., $AUC = 0.68$ for the strong rings artifact), the optimization was crucial to achieve strong results across OOD evaluations. Using the InfoOOD measure with optimization, however, introduced the need for hyperparameter selection, such as the learning rate or $\beta$ parameter. A hyperparameter search was performed for the $\beta$ parameter, which found $\beta =$ 1E+03 to be optimal. Performing additional searches for the remaining parameters was not critical because OOD performance was expected to saturate within the optimization, and changing the hyperparameters will likely only change the iteration at which this saturation occurs. However, this saturation point may be different across different OOD evaluations. For example, OOD detection performance saturated at approximately the 25$^{th}$ iteration for most artifact types and magnitudes. For the medium and strong magnitude rings artifact, however, it appeared the saturation point was not yet reached at the final iteration. This was also the case in the correlation evaluations (Figure 16). Studies investigating optimal stopping point strategies for this optimization process are needed to address this methodological limitation.

Several additional challenges are present in this work, aside from the undetermined optimal stopping point criteria for OOD purposes. A primary challenge was the absence of an OOD benchmark

dataset for medical image OOD detection, which, to our knowledge, does not currently exist and therefore limited our ability to directly compare our work to others. Ideally, our detailed explanation of the OOD curation process using simulated image artifacts will invoke others to assess OOD measures in a similar manner and may contribute to the standardization of defining medical image OOD data from which to benchmark OOD measures. Another challenge of this work was only investigating one category of OOD data. In addition to image artifacts, OOD data could be categorized in several ways, such as differences in image scanners manufacturers and models, stark anatomical differences, and the presence of pathological anomalies. However, our approach using simulated image artifacts enabled us to curate a large and comprehensive test set from a single test data set. Acquiring OOD data from these alternative categories would be very challenging and would require the use of real clinical data. Building public data repositories of these types of clinical data in addition to sharing data on which high-quality models fail to generalize (i.e., performs poorly) will help researchers build reliable OOD detection and uncertainty quantification methods. Our data pre-processing introduced another challenge to this work. Images were resampled to 2.5 $mm^3$ voxel spacing due to computer memory constraints. This resampling process may have removed features important for OOD detection, and sensitivity may be enhanced using higher resolution images. This work was further challenged by limiting the assessment to a single disease type and modality. While the selected modality (i.e., CT) is routinely used for the clinical care of patients with metastatic liver tumors, other modalities also play a role (Fusai & Davidson, 2003). Still, it is expected that the InfoOOD method will perform well in other modalities due to an overlap of defining liver organ and tumor image features (e.g., shape, edges, etc.) between modalities. Lastly, the introduced method incurs additional computational time at inference due to the post hoc optimization process. Unless real-time inference is a necessity for the application at hand (e.g., in real-time motion monitoring), we found this additional time constraint to be negligible.

The InfoOOD measure introduced in this work could readily be applied to a variety of deep learning tasks. Segmentation tasks, especially ones that employ a U-Net model architecture, could benefit from this work where the implementation should be similar. However, it may be imperative to assess the InfoOOD measure's performance in other modalities and pathologies. This can be done in a similar manner to what we have done for CT images. For example, OOD data can be constructed using artificial MRI artifacts such as spike artifact, under sampling, or low field image acquisition. For classification tasks, it may be necessary to investigate the optimal target layer for placing the information bottleneck block and acquiring the OOD distance measure. For each new application, it will likely also be important to determine the optimal $\beta$ parameter value through an ablation experiment and observe the OOD evaluation performance across information bottleneck optimization iterations. The applications of this work for contemporary deep learning architectures, such as vision transformers and multi-modal models, may be less straightforward, but they should be explored as future research.

## 2.5 Conclusion

We demonstrated the importance of adapting an information bottleneck optimization process for medical image OOD detection by introducing and evaluating the InfoOOD measure. This approach outperformed established embedded feature-based methods in detecting OOD medical images generated with simulated CT artifacts and in correlating OOD measures with segmentation model performance metrics. Our results encourage the adoption of post hoc information bottleneck optimization for OOD detection in clinical settings, paving the way for safer and more reliable deployments of deep learning-based medical image analysis.

# 3 Comparison of Uncertainty Quantification Methods for Metastatic Lesion Segmentation

---

In this chapter, **Specific Aim 2a** is addressed by implementing and quantitatively comparing the performance of established UQ measures for the metastatic lesion segmentation task on whole body PET/CT images. UQ has not previously been studied for this clinical task, despite its pressing need (described in Section 1.5.2). Thus, the optimal UQ method for this task has not been established. Moreover, quantitative comparisons across UQ methods are not typically employed for medical image applications. In this chapter, we introduce and assess several quantitative comparisons to determine the optimal UQ methods for this understudied task. This work has been published in *Physics in Medicine and Biology* under the title "Uncertainty Quantification for Deep Learning-based Metastatic Lesion Segmentation on Whole Body PET/CT" (Schott et al., 2025b). This work was also published as a scientific abstract and presented at the *American Physical Society's (APS)* annual meeting in 2024 (Minneapolis, Minnesota).

## 3.1 Introduction

As outlined in Section 1.5.2, the metastatic lesion segmentation task is critically in need of UQ due to the associated image interpretation challenges and a variety downstream clinical tasks which rely on the predicted segmentations. Not only will UQ help clinicians understand when a predicted segmentation may be incorrect, but it will also enable the propagation of uncertainty into these downstream tasks such as patient outcome prediction and lesion-wise response-to-treatment assessment. In this chapter, we share our work which investigated state-of-the-art uncertainty

quantification methods to determine the optimal approach for the deep learning-based metastatic lesion segmentation task. Uncertainty measures were acquired using four of the most prominent UQ methods for medical image tasks including maximum softmax probability, Monte Carlo dropout, model ensembling, and test time augmentation techniques. The UQ methods were implemented on a 3D U-Net model trained for metastatic lesion segmentation on $^{68}$Ga-DOTATATE PET/CT scans of patients with metastatic neuroendocrine tumors (NETs). The resulting uncertainty outputs were evaluated across three quantitative evaluations.

## 3.2 Methods

### 3.2.1 Data

A dataset of $N = 59$ whole body $^{68}$Ga-DOTATATE PET/CT images from 59 patients with metastatic neuroendocrine tumors (NETs) was used in this work. All images were acquired prior to the start of a systemic radionuclide therapy. Patient demographic information is provided in Table 4. This data cohort was well suited to investigate UQ methods due to the high disease burden with a total of 2322 lesions (per patient median: 22, range: [1, 220]) and due to its use in outcome prediction models to guide patient care (Santoro-Fernandes et al., 2024b). Moreover, images were acquired across a variety of PET scanners, including GE – Discovery MI, GE – Discovery 710, GE – Discovery ST, Philips Vereos, and Siemens Biographs mCT, which contributed to greater uncertainty in this dataset. Additional scanner information is provided in Table 5. All images were acquired under standard clinical acquisition protocols, which included patient fasting for four hours prior to a $^{68}$Ga-DOTATATE injection dose of 2 MBq per kilogram of bodyweight, where the upper and lower dose limits were 111 MBq and 200 MBq, respectively. Image acquisition started 60 minutes after injection, and the acquisition time was 5 minutes per bed position. Image intensities were converted to standardized uptake values (SUV) by dividing the

measured activity concentration by the injected dose per bodyweight. The same patients underwent $^{177}$Lu-DOTATATE peptide receptor radiation therapy. All data in this study was retrospectively acquired from the University of Wisconsin Hospital and Clinics (UWHC) following all ethical guidelines and internal review board authorization.

*Table 4 Patient demographic information from the $N = 59$ images and patients.*

| Sex | | Age | | |
|---|---|---|---|---|
| Male | Female | <50 | 50 to 60 | ≥60 |
| 34 | 25 | 9 | 20 | 30 |

*Table 5 Image acquisition scanner information and instances. VPFXS = Vue Point FX system; VPHD = Vue Point HD ViP; OSEM = ordered-subset expectation maximization; i3s15 = 3 iterations, 15 subsets; OSEM3D = 3-dimensional ordered-subset expectation maximization; TOF = time of flight; 2i21s = 2 iterations, 21 subsets; N/a = not applicable.*

| Scanner Manufacturer | Scanner Model | Reconstruction Technique | Reconstruction Diameter (mm) | *N* |
|---|---|---|---|---|
| GE Healthcare | Discovery 690 | VPFXS | 700 | 2 |
| | Discovery 710 | VPFXS | 700 | 28 |
| | Discovery LS | OSEM | 500 | 1 |
| | Discovery MI | VPFXS | 700 | 23 |
| | | VPHD | 700 | 2 |
| Philips | Vereos | OSEM 3is15 | 576 | 1 |
| | | | 676 | 1 |
| Siemens | Biograph mCT | OSEM3D with TOF 2i21s | N/a | 1 |

For segmentation model training purposes, NET lesions were identified and manually segmented under physician guidance by four radiographers with a minimum of 10 years of experience using 3D Slicer and ITK-SNAP. Prior to model input, all images were resampled to 2.5 mm isotropic voxel spacing and normalized to zero-mean-unit-variance. Example coronal view maximum intensity projection (MIP) images from this dataset with corresponding ground-truth labels are shown in Figure 17.

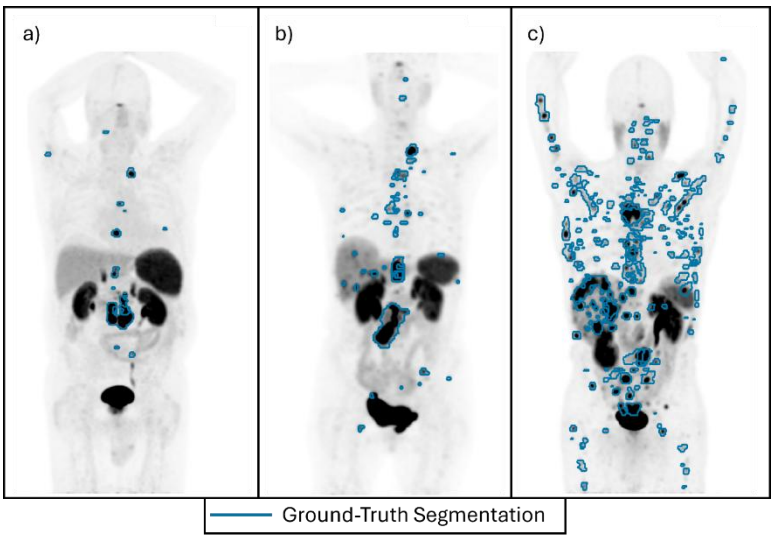



*Figure 17 Coronal view maximum intensity projection (MIP) images from $^{68}$Ga-DOTATATE PET/CT scans of three example patients with (a) low (N=16 lesions), (b) medium (N=43 lesions), and (c) high (N=220 lesions) disease burdens. MIP images are shown with a SUV window between 0 and 12.*

### 3.2.2 Segmentation Model

A 3D convolutional neural network (CNN) was trained in-house to segment NETs on the baseline $^{68}$Ga-DOTATATE PET/CT images in a binary fashion using the nnUNet repository (Isensee et al., 2021). This is a medical image segmentation tool which follows a standard U-Net architecture (Ronneberger et al., 2015) yet automatically selects an optimal subset of training hyperparameters given data properties and available computer infrastructure. The nnUNet model has achieved strong performance across a variety of medical image segmentation tasks, making it well suited to test and compare multiple UQ methods because the training model error should be minimized. The nnUNet model was also selected due to its widespread adoption for whole body metastatic lesion segmentation and quantification on PET/CT (Blanc-Durand et al., 2021; Carlsen et al., 2022; B. Huang et al., 2024; Leung et al., 2024; Moreau et al., 2022; Schott et al., 2023; Xue et al., 2023).

Five-fold cross validation was invoked for model training with 80%/20% train/test data splits to acquire model predictions and uncertainty measures in each image in the dataset. The model outputs on the test data from each fold were used for all segmentation and UQ measure performance evaluations. Each model fold was trained for 500 epochs and 250 minibatches per epoch using the sum of cross entropy and Dice coefficient as the loss and using the Adam optimizer. Data augmentation was applied during training with spatial, color, noise, and cropping augmentations. The model was trained on a workstation with 31 GB of system RAM, an 8-core CPU with 16 threads (2 threads per core), and an NVIDIA RTX Titan GPU with 25 GB of memory.

### 3.2.3 Uncertainty Quantification Methods

Four uncertainty quantification methods were implemented in this work: (1) probability entropy, (2) Monte Carlo dropout, (3), model ensembling, and (4), test time augmentation. These were selected due to their predominate use in previous UQ studies (Lambert et al., 2024). A summary of each method is provided in Figure 18.

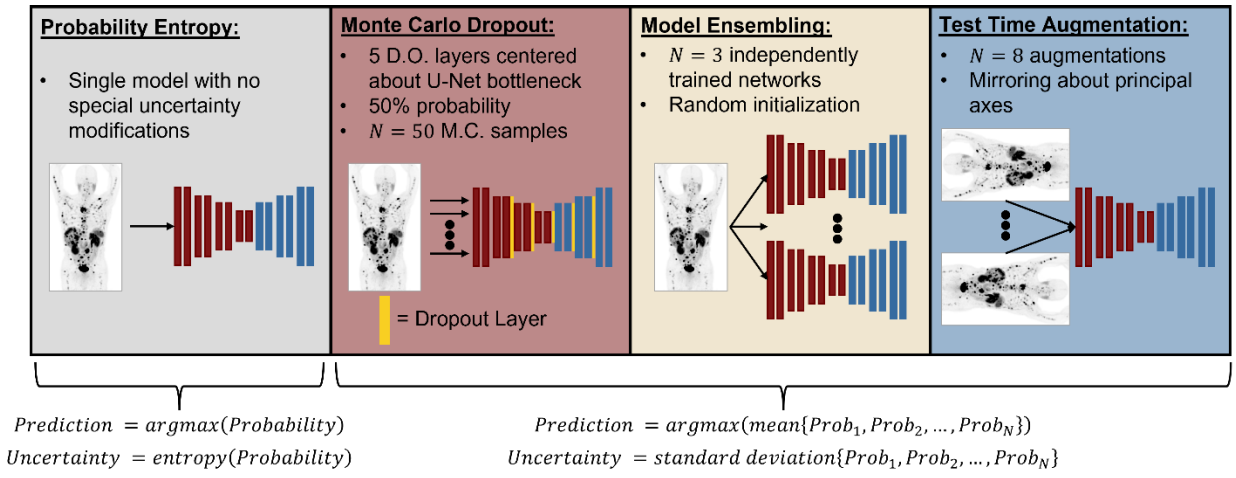



*Figure 18 Schematic depicting a high-level summary of each implemented uncertainty quantification method.*

The **probability entropy UQ measure ($U_{PE}$)** was implemented using the predicted class probability output within the predicted foreground from the standard segmentation model (described in Section 3.2.2). Probability entropy quantifies the amount of information in a probability distribution and is maximized when probabilities represent equivocal prediction classification (i.e., the probability across prediction classes is uniform). Since this approach can be implemented without any model modifications, it serves as a reference method to compare against more involved UQ measures. The probability entropy uncertainty measure was defined as

$$U_{PE.,i} = -p_i \cdot \log(p_i) - (1 - p_i) \cdot \log(1 - p_i), \quad (8)$$

where $p_i$ represents the model's *softmax* foreground probability for voxel $i$ and $\log(\cdot)$ represents the natural logarithm. Segmentation masks were defined using the voxel-wise *argmax* operation applied across prediction classes of the model's probability output.

The **Monte Carlo dropout UQ measure ($U_{MCDO}$)** was implemented using a separate segmentation model trained according to the description of Section 3.2.2, except with dropout layers engaged. In this model, dropout layers were inserted on the last two layers of the model's encoder, on the model's bottleneck layer, and on the first two layers of the model's decoder. This configuration aligns with those used in previous segmentation UQ studies (Bhat et al., 2021; Nair et al., 2020), and it also has been shown to be the optimal dropout layer placement for UQ purposes (Kendall et al., 2015). Moreover, each dropout layer had a dropout probability of 50%, which was previously found to be optimal for UQ applied to lesion analysis tasks (Bhat et al., 2021). During test time, each image was passed through the trained model $N = 50$ times with dropout layers engaged.

The **model ensembling UQ measure ($U_{ENS}$)** was implemented by training three independent segmentation models with random weight initialization (following the description in Section 3.2.2). Each

test scan was separately passed through each model to obtain three predictions per scan. Including more than three ensembles was computationally expensive due to the necessity of following a five-fold cross validation training approach for each model in the ensemble.

The **test time augmentation UQ measure ($U_{TTA}$)** was implemented using the model described in Section 3.2.2. Following the test time augmentation strategy in (González et al., 2022), each test image was augmented by applying mirroring along each combination of image axes, resulting in eight predictions per image.

Following the methodology in (Klanecek et al., 2023), the voxel-wise mean followed by *argmax* binarization and standard deviation across predicted probabilities were used to acquire the predicted segmentation and uncertainty, respectively, for $U_{MCDO}$, $U_{ENS}$, and $U_{TTA}$. For all uncertainty measures, high magnitudes corresponded to high predictions and vice versa. Normalization of the uncertainty measures was not performed to preserve the unbounded upper bound of the standard deviation for $U_{MCDO}$, $U_{ENS}$, and $U_{TTA}$, and the quantitative comparisons did not require the measures to be on the same scale.

### 3.2.4 Segmentation Model Performance

The implementation of each UQ measure investigated in this work inevitably yields a different segmentation output from a standard model (except for $U_{PE}$). Hence, we investigated the segmentation model performance of each UQ method to better understand how to interpret UQ performance assessments. Prediction performance of each UQ method was stratified by segmentation, lesion detection, and biomarker extraction performance metrics. Segmentation performance was reported using the mean Dice coefficient ($\pm$ standard deviation) across test patients. Lesion detection performance was reported using the mean sensitivity ($\pm$ standard deviation) and the mean number of

false positive predicted regions per patient ($\pm$ standard deviation) across test patients. Lastly, biomarker extraction performance was evaluated using the Pearson correlation coefficient between the $SUV_{mean}$ and $SUV_{total}$ biomarkers extracted using predicted and ground-truth segmentations. 95% confidence intervals acquired via bootstrapping using $n = 1{,}000$ random samples with replacement were reported for biomarker extraction performance.

Within the segmentation performance analysis, we also reported the segmentation model train and inference times for the model configurations associated with each uncertainty measure. Model train times were reported as summations across the five training sessions necessary for cross validation. Model inference times were reported as the average computation time per image.

### 3.2.5 Uncertainty Assessments

#### Image Degradation Detection

The first uncertainty assessment consisted of using uncertainty measures to detect artificial degradations added to test images. Here, we assume that images with added degradations should illicit predicted segmentations with higher uncertainty. Within a clinical setting, degraded images may be caused by factors such as differences in scanner properties or acquisition parameters (e.g., shorter acquisition time). We simulated degraded image quality using additive Gaussian noise. Noise was added to the PET components of each PET/CT scan, $x_{PET} \in \mathbb{R}^{n \times m \times z}$, according to

$$\tilde{x}_{PET} = x_{PET} + \mathcal{N}(\mu = 0, \sigma), \tag{9}$$

where $\mathcal{N}(\mu = 0, \sigma)$ is an $n \times m \times z$ Gaussian noise matrix with 0-mean and $\sigma$-standard deviation. Image noise was inserted on the test data using three magnitudes, $\sigma = [2.5, 5, 7.5]$, constructing three sets of noised test data of low, medium, and high magnitudes. It was assumed that uncertainty should increase

with an increase in noise perturbation magnitude. An example of the noised images for a single test scan is shown in Figure 19.

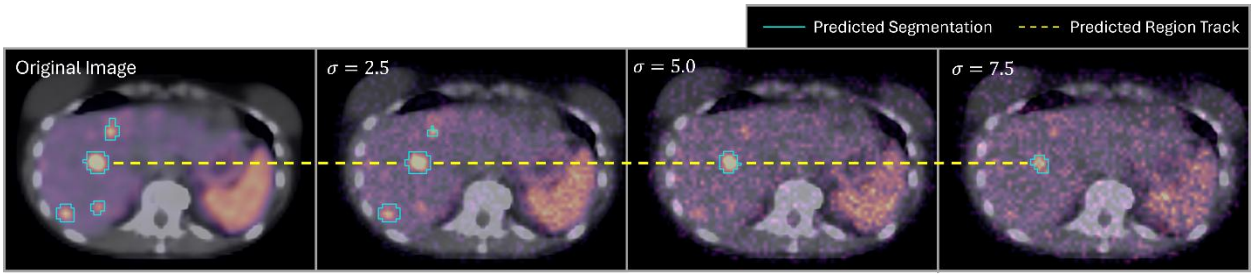



*Figure 19 An example abdominal slice from a $^{68}$Ga-DOTATATE PET/CT test scan with additive Gaussian noise at each inserted noise magnitude and overlaid predicted lesion segmentations (solid cyan line) from the standard segmentation model (Described in Section 3.2.2). The dashed yellow line indicates the persistent predicted region track from which the uncertainty-based image degradation data is derived.*

Model predictions and uncertainty measures were acquired for each test image across additive noise magnitudes. Individually predicted lesion regions for a given test image were matched and tracked across noise magnitudes using a previously developed automatic lesion matching tool (Santoro-Fernandes et al., 2024a) to establish predicted lesion tracks (an example is shown in Figure 19). Within each track and for each noise magnitude, the uncertainty measure magnitude was used to distinguish the predicted segmentation on the noised image from the predicted segmentations on the non-noised image. As some predicted segmentations disappear with increasing noise magnitudes, the noise detection assessment was confined to lesions with tracked predicted segmentations across all noise magnitudes to ensure the same number of regions was used for detection in each noise category. Voxel-wise uncertainty measures were aggregated into region-wise measures using the mean voxel-wise uncertainty within a predicted region as:

$$U_{region} = \frac{1}{N}\sum_{i=1}^{N} u_i \,, \tag{10}$$

where $u_i$ is the $i^{th}$ voxel uncertainty in a predicted region with $N$ total voxels. We reported the distributions of each uncertainty measure across noise magnitudes (including no noise) and the statistical significance between each noised and non-noised distribution (using Wilcoxon signed rank test). Detection metrics including the area under the receiver operator curve (AUC) and the false positive rate at the 95% true positive rate threshold were used to evaluate the performance of each uncertainty measure. Bootstrapping over each set of region-wise uncertainty values was performed using $N = 1{,}000$ random bootstrap samples with replacement to acquire 95% confidence intervals on the detection metrics.

### False Positive Region Detection

Subsequently, we assessed each uncertainty measure's ability to detect false positive predicted regions. In this assessment we followed the assumption established by Nair and colleagues (Nair et al., 2020) that false positive predicted regions should contain high levels of predictive uncertainty, and that a good uncertainty measure should be able to detect false positive predicted regions. The false positive detection performance of each uncertainty measure was assessed using receiver-operator curve (ROC) metrics including the area under the ROC curve (AUC) and the false positive rate at the 95% true positive rate threshold. Bootstrapping was employed in the same manner as in the Image Degradation Detection section (directly above) to acquire 95% confidence intervals on these metrics. False positive predictive regions were determined using the same metastatic lesion matching algorithm as in the Image Degradation Detection (Santoro-Fernandes et al., 2024a) to match predicted and ground-truth lesions. Lastly, we reported the uncertainty measure threshold which maximized Youden's J statistic for detecting FP regions. This threshold indicated the best balance between sensitivity and specificity of distinguishing FP from TP regions. Statistical significance between TP and FP distributions was tested using the

Wilcoxon rank sum test. Again, voxel-wise uncertainties were aggregated into region-wise uncertainties using the mean voxel-wise uncertainty within a predicted region (as in equation 10).

### False Negative Region Recovery

Next, we investigated the potential to use each uncertainty measure for false negative (FN) region recovery. FN regions may be recoverable using uncertainty information if a region exhibits probability values below the probability binarization threshold for segmentation yet contains uncertainty values which are higher than those found in the true negative (TN) space. Since the TN space represented the majority of the image (i.e., true non-lesion region), it was not feasible to distinguish FNs from TNs using regional uncertainty magnitude. Instead, the recall rate of detecting FN regions was explored as a function of intra-image uncertainty thresholds.

To accomplish this, each FN region within an image was determined through the same lesion matching algorithm as in the above sections (Santoro-Fernandes et al., 2024a). The FN region uncertainties were computed following equation 10, where the voxel localization was defined by the ground-truth segmentation. The FN region recall rate within an image was defined as a function of uncertainty thresholds. These thresholds were defined as factors of the median TP regional uncertainty measure within an image, linearly spaced in decreasing order from 1 to 0.05. The FN recall rate at each threshold was computed as the number of FN regions with an uncertainty that is equal to or greater than the threshold divided by the total number of FN regions. Thus, a higher recall rate was expected as the threshold decreased. Moreover, uncertainty measures with higher recall rates across thresholds were assumed to demonstrate greater potential to recover FN regions. The mean FN recall rate as a function of intra-image uncertainty thresholds across images was reported for each uncertainty

measure. For quantitative comparison, we reported the FN recall rate for each uncertainty measure at 0.5 and 0.1 threshold factors.

## Correlations to Model Segmentation Performance

Lastly, we assessed each uncertainty measure's correlation with segmentation model performance. Here, we hypothesize that a reliable uncertainty measure should correspond with model performance such that higher uncertainty predictions are associated with poorer model performance. Model performance was defined in two manners: biomarker extraction accuracy and segmentation accuracy.

The biomarkers included in this assessment were the image-wise mean ($SUV_{mean}$) and total ($SUV_{total}$) standardized uptake value. These biomarkers were selected due to their previously established use in both predictive modeling of patient outcome and patient response assessments (Harmon et al., 2017; Hartrampf et al., 2023; Lokre et al., 2024; Swiha et al., 2024). The predicted lesion segmentation masks within each image were used to extract corresponding predicted lesion PET uptake data. The average and summation of this data was computed and defined the image-wise $SUV_{mean}$ and $SUV_{total}$ biomarkers, respectively. Accuracy for biomarker $x$ (i.e., either mean or total) was defined as the following relative difference:

$$Accuracy(SUV_x) = \frac{\left|SUV_{x,pred} - SUV_{x,GT}\right|}{\left(SUV_{x,pred} + SUV_{x,GT}\right) / 2}, \tag{11}$$

where $SUV_{x,pred}$ and $SUV_{x,GT}$ indicates the selected SUV biomarkers extracted using predicted and ground-truth lesion segmentations, respectively. Segmentation accuracy was defined as *1 – Dice coefficient* and *cross entropy*, such that positive correlations were expected between each performance metric and uncertainty measure. Spearman correlation coefficients were reported between each

uncertainty measure and model performance metric. Bootstrapping was employed using $N = 1{,}000$ random bootstrap samples with replacement to acquire 95% confidence intervals on these correlation coefficients.

Each selected segmentation performance metric is an image-wise metric. Thus, to perform the correlations, the voxel-wise uncertainty measures were aggregated into image-wise uncertainty measures using the total uncertainty in the image, normalized by the number of lesion-predicted voxels within the image:

$$U_{image} = \frac{1}{M}\sum_{j=1}^{K} u_j \,, \tag{12}$$

where $u_j$ is the uncertainty in the $j^{th}$ image voxel, $K$ is the total number of voxels in the image, and $M$ is the total number of predicted voxels within an image.

## 3.3 Results

### 3.3.1 Segmentation Model Performance

The model segmentation performance from each model configuration associated with each UQ measure is summarized in Table 6. Segmentation performance in terms of the Dice coefficient was comparable across all UQ method implementations, where the lowest Dice coefficient was 0.69 (for $U_{PE}$ and $U_{MCDO}$) and the highest Dice coefficient was $0.71$ (for $U_{ENS}$). Lesion detection sensitivity was also comparable across UQ methods, where the lowest sensitivity was 0.75 (for $U_{PE}$) and the highest sensitivity was $0.77$ (for $U_{MCDO}$ and $U_{ENS}$). The number of false positive predicted regions per patient were comparable for $U_{MCDO}$ ($n = 8.5$), $U_{ENS}$ ($n = 8.8$), and $U_{TTA}$ ($n = 8.8$) methods, however, the

standard deviation for this metric was large across UQ methods. Lastly, the Pearson correlation coefficients between predicted and ground-truth based biomarkers was 0.95 across UQ methods for $SUV_{mean}$. The lowest and highest performing correlation for $SUV_{total}$ was 0.96 (for $U_{MCDO}$) and 0.98 (for $U_{TTA}$), respectively.

*Table 6 Segmentation model performance metrics for each UQ method implementation. Segmentation performance is reported as the mean Dice Coefficient (± standard deviation) across test patients. Lesion detection performance is reported as the mean (± standard deviation) sensitivity and number of false positive predicted regions (Num. FPs). Biomarker extraction accuracy was reported as the Pearson correlation coefficient between predicted and ground truth-based biomarkers, where ([·]) indicates 95% confidence intervals derived from bootstrapping.*

| | Segmentation | Lesion Detection | | Biomarker Accuracy | |
|---|---|---|---|---|---|
| | Dice Coeff. ↑ | Sensitivity ↑ | Num. FPs ↓ | $SUV_{mean}$ ↑ | $SUV_{total}$ ↑ |
| $U_{PE}$ | 0.69 (±0.16) | 0.75 (±0.16) | 9.6 (±13.4) | 0.95 [0.92, 0.97] | 0.97 [0.94, 0.99] |
| $U_{MCDO}$ | 0.69 (±0.15) | 0.77 (±0.16) | 8.5 (±11.9) | 0.95 [0.91, 0.97] | 0.96 [0.92, 0.99] |
| $U_{ENS}$ | 0.71 (±0.14) | 0.77 (±0.15) | 8.8 (±12.6) | 0.95 [0.92, 0.97] | 0.97 [0.96, 0.99] |
| $U_{TTA}$ | 0.70 (±0.16) | 0.76 (±0.16) | 8.8 (±12.3) | 0.95 [0.92, 0.97] | 0.98 [0.96, 0.99] |

The segmentation model training and inference times are summarized in Table 7. For the uncertainty measures requiring only one trained model, all the train times were comparable. $U_{ENS}$ required three independently trained models; therefore, model training time was approximately three times greater than for the other measures. $U_{PE}$ had the shortest inference time followed by $U_{ENS}$ then $U_{TTA}$. $U_{MCDO}$ had the longest inference time by a large margin.

*Table 7 Computation times for deploying each uncertainty measure. Train times include the 5 training sessions necessary for cross validation training. Inference times are reported as average (± standard deviation) times per image.*

| Associated Uncertainty Measure | Total Train Time (hours) | Inference Time per Image (seconds) |
|---|---|---|
| $U_{PE}$ | 115.3 | $0.9 \pm 0.4$ |
| $U_{MCDO}$ | 115.2 | $19.2 \pm 3.4$ |
| $U_{ENS}$ | 346.9 | $2.8 \pm 2.6$ |
| $U_{TTA}$ | 115.3 | $3.7 \pm 2.0$ |

## 3.3.2 Uncertainty Assessments

### Image Degradation Detection

The distributions of each uncertainty measure across noise magnitudes are shown in Figure 20. Each uncertainty measure responded positively to the added noise, where an increase in additive noise magnitude incited an increase in uncertainty. All distributions from the noised images were statistically different from the non-noised distributions, where all had a p-value of $p < 0.001$, except for the comparison between the $\sigma = 2.5$ and the 'No Noise' distribution for $U_{PE}$.

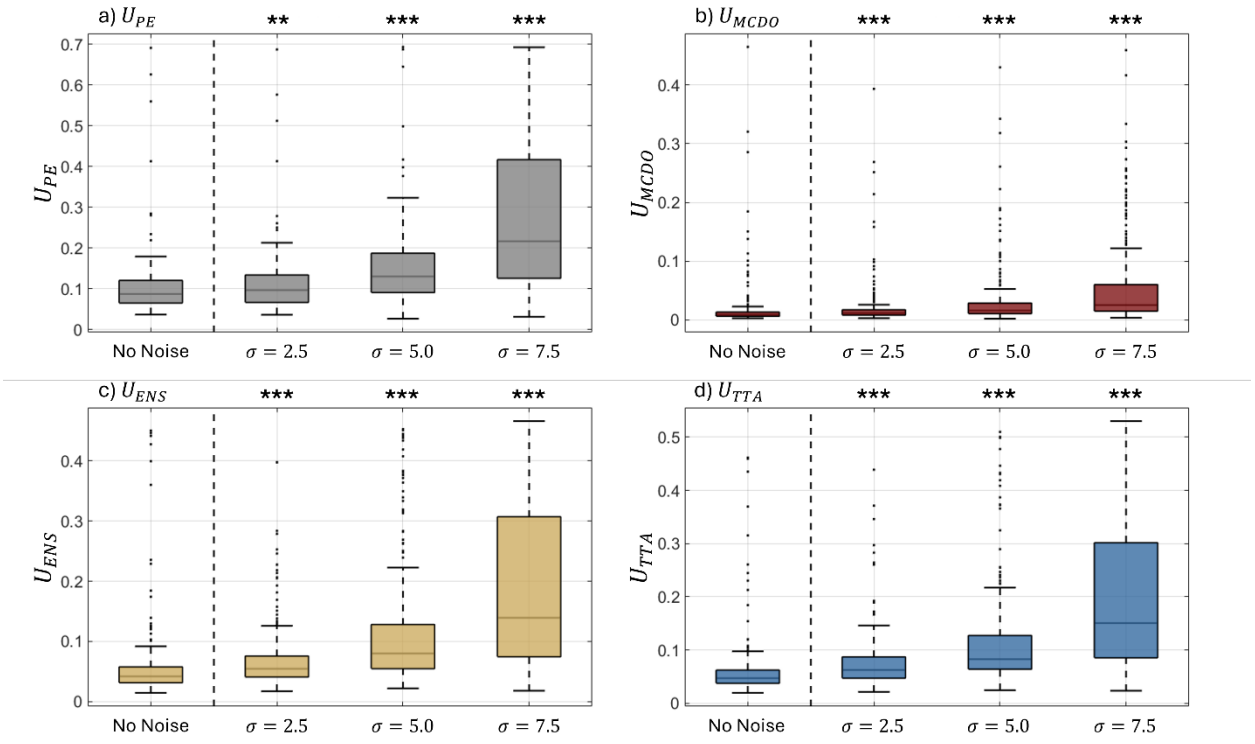



*Figure 20 Distributions of regional uncertainty values across additive noise magnitudes. Statistical significance testing was performed between each noised distribution to the non-noised distribution (shown above), where $**$ implies $p < 0.01$ and $***$ implies $p < 0.001$.*

The ROC curves for distinguishing predicted segmentations from noised and non-noised images for each uncertainty measure and noise magnitude are shown in Figure 21. The detection statistics derived from these ROC curves are summarized in Table 8. Each non-baseline UQ measure outperformed the $U_{PE}$ for this detection task across noise magnitudes. Across all noise magnitudes, the $U_{MCDO}$ measure yielded inferior AUC values when compared to the $U_{ENS}$ and $U_{TTA}$ measures. The $U_{TTA}$ measure outperformed all other measures across noise magnitudes and detection metrics except for FPR95 for the least noised images, where the $U_{MCDO}$ measure yielded a better result of 0.78 (compared to 0.86 for $U_{TTA}$). The $U_{ENS}$ measure performed comparably to $U_{TTA}$ when considering the overlapping of the 95% confidence intervals.

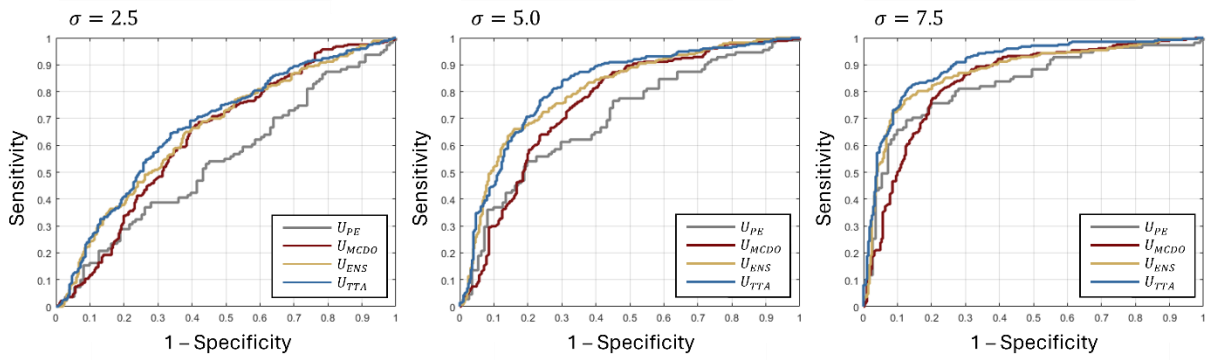



*Figure 21 ROC curves for using each UQ measure to distinguish between lesions with and without additive noise for each noise magnitude ($\sigma = [2.5, 5.0, 7.5]$. Each predicted lesion was matched and tracked across noised images. Within each track, the change in lesion-wise uncertainty from the original (non-noised) image to each noised image was used to distinguish between prediction from noised and non-noised images.*

*Table 8 Detection statistics, including the area under the ROC curve (AUC) and the false positive rate at the 95% sensitivity threshold (FPR95), for each UQ measure for distinguishing predicted lesion segmentations in noised images from matched predicted lesion segmentations in non-noised images. Each metric is reported with the mean and 95% confidence intervals ([·]) from bootstrapping. Bolded and underlined text indicates the best and second-best performing result, respectively.*

| $\sigma$ | | $U_{PE}$ | $U_{MCDO}$ | $U_{ENS}$ | $U_{TTA}$ |
|---|---|---|---|---|---|
| 2.5 | AUC ↑ | 0.54 [0.47, 0.62] | 0.64 [0.59, 0.69] | <u>0.66</u> [0.61, 0.70] | **0.68** [0.63, 0.72] |
| | FPR95 ↓ | 0.96 [0.86, 1.00] | **0.78** [0.72, 0.87] | 0.87 [0.81, 0.92] | <u>0.86</u> [0.77, 0.93] |
| 5 | AUC ↑ | 0.70 [0.63, 0.77] | 0.75 [0.71, 0.80] | <u>0.81</u> [0.77, 0.84] | **0.82** [0.78, 0.85] |

| | | | | | |
|---|---|---|---|---|---|
| | FPR95 ↓ | 0.89 [0.69, 0.95] | 0.72 [0.64, 0.80] | <u>0.71</u> [0.58, 0.79] | **0.67** [0.51, 0.84] |
| 7.5 | AUC ↑ | 0.83 [0.77, 0.88] | 0.84 [0.80, 0.87] | <u>0.87</u> [0.84, 0.90] | **0.90** [0.88, 0.93] |
| | FPR95 ↓ | 0.68 [0.49, 0.98] | <u>0.61</u> [0.38, 0.77] | 0.64 [0.42, 0.80] | **0.41** [0.27, 0.56] |

### False Positive Region Detection

The ROCs for detecting false positive predicted regions across UQ measures are shown in Figure 22. The TP and FP distributions were statistically different (i.e., p<0.001) for each uncertainty measure (Wilcoxon rank sum test). The detection statistics for the FP detection assessment are summarized in Table 9. All uncertainty measures yielded strong performance in this assessment, where the worst performing AUC was 0.77 ($U_{PE}$) compared to the best performing AUC of 0.81 ($U_{TTA}$). The $U_{ENS}$ and $U_{TTA}$ achieved comparable FPR95 values at 0.65 and 0.66, respectively. Meanwhile, the FPR95 values achieved by the $U_{MSP}$ and $U_{MCDO}$ were slightly inferior at 0.83 and 0.76, respectively. The regional uncertainty measure which maximized Youden's J Statistic was 0.15 for $U_{PE}$, 0.04 for $U_{MCDO}$, 0.12 for $U_{ENS}$, and 0.16 for $U_{TTA}$.

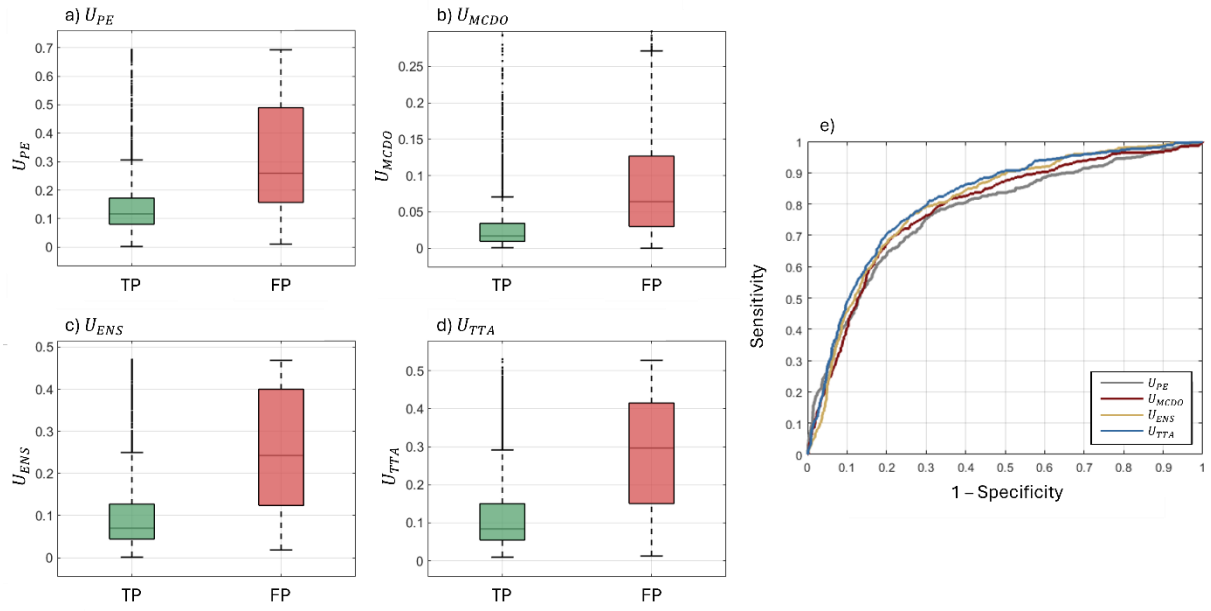



*Figure 22 FP detection results for each uncertainty measure. The difference in uncertainty measures between true positive and false positive predicted regions for (a) $U_{PE}$, (b) $U_{MCDO}$, (c) $U_{ENS}$, (d) $U_{TTA}$ are shown as box plots. (e) The ROC curves for detecting false positive from true positive predicted regions using each uncertainty measure. The TP and FP distributions were statistically different (i.e., p<0.001) for each uncertainty measure.*

*Table 9 False positive predicted region detection statistics across uncertainty measures including area under the ROC curve (AUC) and the false positive rate at the 95% sensitivity threshold (FPR95). Ranges within brackets ([·]) indicate the 95 percent confidence interval for each statistic. Bolded and underlined text indicates the best and second-best performing result, respectively.*

| | $U_{PE}$ | $U_{MCDO}$ | $U_{ENS}$ | $U_{TTA}$ |
|---|---|---|---|---|
| AUC ↑ | 0.77 [0.75, 0.80] | 0.79 [0.76, 0.81] | <u>0.80</u> [0.78, 0.82] | **0.81** [0.79, 0.83] |
| FPR95 ↓ | 0.83 [0.75, 0.89] | 0.76 [0.66, 0.87] | **0.65** [0.61, 0.76] | <u>0.66</u> [0.53, 0.76] |

## False Negative Region Recovery

The FN recall curves as functions of intra-image regional uncertainty threshold factors are shown in Figure 23. The $U_{TTA}$ measure exhibited the strongest FN recall across thresholds factors, while the $U_{PE}$ exhibited the worst recall. Meanwhile, the $U_{MCDO}$ and $U_{ENS}$ measures demonstrated comparable FN recovery performance. FN recall at the 0.5 factor threshold was 0.02 for $U_{PE}$, 0.26 for $U_{MCDO}$, 0.25 for

$U_{ENS}$, and 0.33 for $U_{TTA}$. FN recall at the 0.1 factor threshold was 0.04 for $U_{PE}$, 0.34 for $U_{MCDO}$, 0.37 for $U_{ENS}$, and 0.49 for $U_{TTA}$.

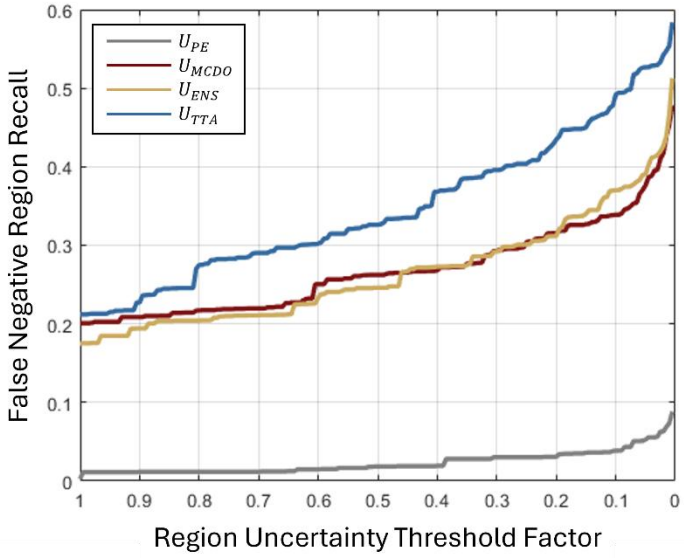



*Figure 23 FN region recall rates as functions of region uncertainty threshold factors of the median TP region uncertainty within each image.*

## Correlations to Model Segmentation Performance

The Spearman correlation coefficients between the image-wise aggregated uncertainty measures and segmentation and biomarker extraction performance metrics are shown in Table 10. Across the tested metrics, the $U_{PE}$ achieved the worst performance except for the cross-entropy metric, where it achieved the strongest correlation ($\rho = 0.96$). For the accuracy of capturing $SUV_{mean}$, the $U_{MCDO}$ and $U_{TTA}$ performed comparably with a Spearman correlation coefficient of 0.35 and 0.34, respectively. The $U_{TTA}$ measure yielded the best correlation with accurately capturing $SUV_{total}$ ($\rho = 0.57$), which was followed by $U_{MCDO}$ ($\rho = 0.51$) and $U_{ENS}$ ($\rho = 0.51$). The $U_{TTA}$ measure similarly yielded the best correlation with the *1 - Dice coefficient* metric ($\rho = 0.72$), which was followed by $U_{ENS}$ ($\rho = 0.70$). Large confidence intervals were observed across all uncertainty measures for the $SUV_{mean}$, $SUV_{total}$, and *1 – Dice*

*Coefficient* performance metrics. Strong correlations with tighter confidence intervals were observed between the cross-entropy metric, where the $U_{PE}$, $U_{ENS}$, and $U_{TTA}$ correlations were comparable given the confidence bounds. Across all segmentation model performance metrics, $U_{TTA}$ achieved the overall best correlation performance. The correlation plots between the higher performing $U_{TTA}$ measure and the segmentation performance metrics are shown in Figure 24.

*Table 10 Spearman correlation coefficients between the uncertainty measures and segmentation model performance metrics. The accuracy of the SUV metrics was quantified using log differences between predicted and ground truth-extracted SUV values. Ranges within brackets ([·]) indicate the 95 percent confidence interval for each correlation coefficient. Bolded and underlined text indicates the best and second-best performing result, respectively.*

| | $U_{PE}$ | $U_{MCDO}$ | $U_{ENS}$ | $U_{TTA}$ |
|---|---|---|---|---|
| $SUV_{mean}$ | 0.24 [-0.05, 0.49] | **0.35** [0.12, 0.57] | 0.29 [0.03, 0.52] | <u>0.34</u> [0.06, 0.58] |
| $SUV_{total}$ | 0.37 [0.10, 0.61] | <u>0.51</u> [0.30, 0.70] | <u>0.51</u> [0.27, 0.71] | **0.57** [0.36, 0.73] |
| 1 - Dice Coeff. | 0.52 [0.30, 0.72] | 0.67 [0.48, 0.81] | <u>0.70</u> [0.53, 0.83] | **0.72** [0.57, 0.84] |
| Cross Entropy | **0.96** [0.93, 0.98] | 0.84 [0.74, 0.91] | <u>0.92</u> [0.85, 0.97] | 0.91 [0.85, 0.95] |

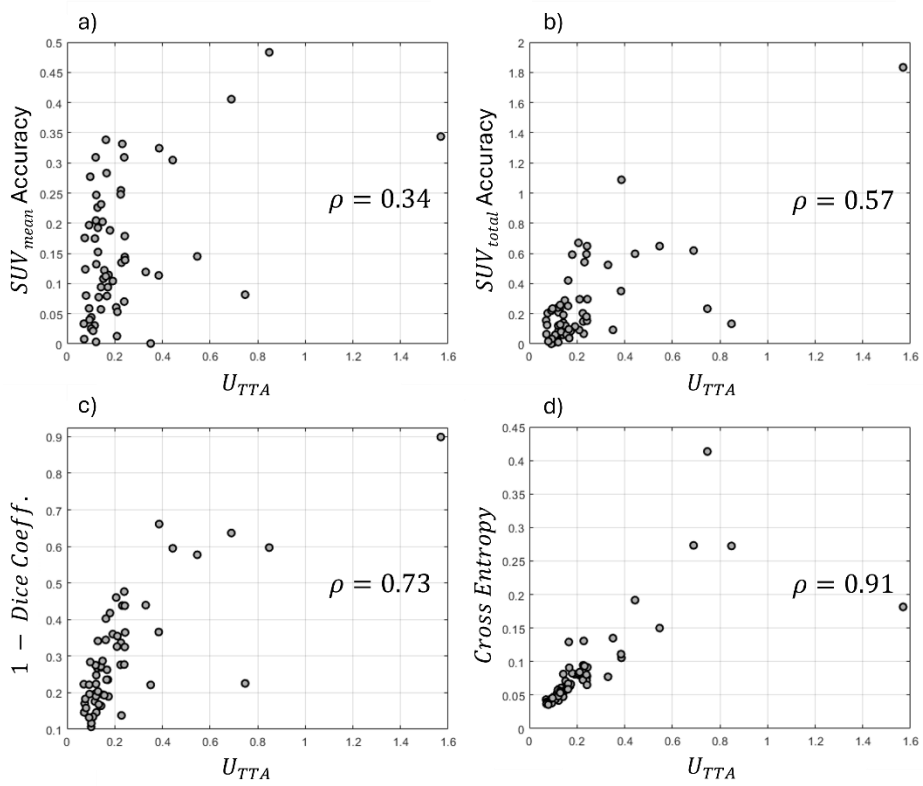



*Figure 24 Scatter plots and spearman correlation coefficients (ρ) between the image-wise $U_{TTA}$ uncertainty measure and each segmentation model performance metric—a) $SUV_{mean}$ accuracy, b) $SUV_{total}$ accuracy, c) segmentation Dice coefficient, and segmentation cross entropy. Results are shown only for the strongest performing, $U_{TTA}$ measure.*

## 3.4 Discussion

In this work, we investigated four UQ methods—probability entropy ($U_{PE}$), Monte Carlo dropout ($U_{MCDO}$), deep ensembles ($U_{ENS}$), and test time augmentation ($U_{TTA}$)—for the deep learning-based metastatic lesion segmentation task. This is the first systematic assessment of UQ methods for segmentation of metastatic malignancies on whole body imaging. Across UQ assessments, the simple $U_{PE}$ measure underperformed all the other, more involved uncertainty measures (i.e., $U_{MCDO}$, $U_{ENS}$, and $U_{TTA}$). This indicates that model probability outputs are not sufficient at quantifying uncertainty, and auxiliary UQ methods should be employed. Amongst these uncertainty measures, we found that the test

time augmentation ($U_{TTA}$) method yielded optimal UQ performance for this task, where it achieved either best or second best performance at detecting artificially degraded images (Figure 17 and Table 8), at detecting false positive predicted regions (Figure 22 and Table 9), at recovering false negative predicted regions (Figure 23), and at capturing correlations with model performance metrics (Table 10 and Figure 24). When the $U_{TTA}$ measure yielded second-best performance, it was still comparable to the best performing method given the reported confidence intervals. The performance of all other UQ measures across evaluations was more heterogeneous than $U_{TTA}$. $U_{ENS}$ yielded the second fastest inference time from $U_{PE}$, which used a single pass through a single model, followed by $U_{TTA}$ (Table 7). However, the train time for $U_{ENS}$ was much larger than any other measure. $U_{MCDO}$ yielded the slowest inference time by a large margin. Thus, on our computation infrastructure (described in Section 3.2.2), $U_{TTA}$ was the most computationally efficient measure following $U_{PE}$.

Prior to the uncertainty measure assessments, segmentation model performance was first assessed with the predicted segmentations from each UQ method. The segmentation performance of each UQ method was comparable to each other across segmentation, lesions detection, and biomarker extraction metrics. Consequently, the uncertainty measures from each UQ method can readily be compared because discrepancies in UQ performance due to differences in segmentation performance were minimized. Moreover, the segmentation performance across UQ methods was comparable to previously reported segmentation models for NET segmentation on full-body $^{68}$Ga-DOTATATE PET/CT. For example, in the work by Santilli and colleagues, an nnUNet segmentation model achieved a Dice coefficient of 0.64 and a detection sensitivity of 0.67 (Santilli et al., 2023). The same study reported an inter-observer Dice coefficient on a subset of patient to be 0.68. The lowest Dice score achieved by our UQ methods was 0.69 (the highest was 0.71). Thus, our results are comparable to inter-observer

variability. As a result, any prediction uncertainty due to insufficient model implementation was minimized.

In the image degradation detection assessment, the detection performance of all implemented uncertainty measures increased with increasing image degradation magnitude, where each uncertainty measure distribution from the noised (i.e., degraded) images was statistically different than the uncertainty measures from the non-noised data. This indicates a level of reliability in each measure's ability to detect image data with degraded quality. The $U_{PE}$ measure underperformed all other measures, where the best performance was achieved by the $U_{TTA}$ measure. Consequently, certain uncertainty measures (e.g., $U_{TTA}$) appear to better detect predictions from degraded image data and may be more sensitive at detecting model predictions from lower quality data within clinical settings. Few previous works have assessed UQ performance by detecting artificially degraded data (Schott et al., 2024). However, some works have implemented this assessment for out-of-distribution (OOD) detection evaluations, which captures the uncertainty related to predicting on test data with dissimilarities to the train data. For instance, previous work implemented adversarial attacks at varying magnitudes to assess OOD detection performance (Schott et al., 2025a). In another work by González and colleagues (González et al., 2022), a series of spatial augmentations at increasing magnitudes were applied to test an OOD measure for lung lesion segmentation on CT images. In agreement with our results, this work showed that using an uncertainty measure based on the model's probability output was inferior to both Monte Carlo dropout and test time augmentation at detecting such degraded image data, however, a distance-based OOD measure achieved the strongest results. These image-wise and distance-based OOD measures were not directly transferable to our work which evaluated detecting image degradations at the lesion-level rather than at the image-level. Moreover, OOD detection is generally treated separately from predictive UQ (Y. Liu et al., 2021; Ovadia et al., 2019; Schwaiger et al., 2020). However, since OOD

data may be subtle within medical image settings, more work should be done to explore the relationship between OOD detection and predictive UQ measures in this domain.

In the false positive detection assessment, all tested uncertainty measures yielded strong performance, where the $U_{PE}$ and $U_{TTA}$ measures achieved the lowest ($AUC = 0.77$) and highest ($AUC = 0.81$) AUCs, respectively. Despite demonstrating the lowest performance, the simple $U_{PE}$ measure was still able to distinguish TP and FP predicted regions. This indicates that most of the false positive predicted regions were dominated by equivocal probability values (i.e., probabilities near 0.5 for binary classification). A previous study similarly demonstrated that Monte Carlo-based uncertainty measures performed better than a model's probability values for false positive detection in the multiple sclerosis segmentation task, however, the performance difference between the probability-based and auxiliary UQ measure was larger. (Nair et al., 2020). This discrepancy may be due to differences in model calibration, where better calibrated models have more informative probability outputs (Guo et al., 2017). Therefore, the utility of a model's probability output for false positive detection varies across tasks and model implementations. Another work utilized uncertainty quantification to enhance lesion detection where uncertainty information was used as input into a machine learning model to classify true and false positive regions (Bhat et al., 2022). Consequently, this introduces an additional model that inevitably has its own uncertainty to quantify and limits a user's ability to implement other uncertainty assessments such as correlations to model performance. The use of false positive detection as an UQ assessment is limited. While this assessment is important for lesion, especially metastatic lesion, detection tasks, it is less relevant to other tasks such as primary tumor or organ segmentation. However, other task-specific error detection assessments would be important and implementable for each application (Lambert et al., 2024). To investigate the optimal uncertainty threshold for FP detection, we reported the uncertainty measure which maximizes Youden's J Statistic. This threshold corresponds with the optimal tradeoff

between FP detection sensitivity and specificity. Depending on the clinical use-case, FP detection specificity or sensitivity may be more important. In this case, it will be critical to adjust the uncertainty threshold to meet the clinical need.

In the FN recovery assessment, the $U_{TTA}$ measure displayed the highest FN recall rate across uncertainty thresholds. The $U_{MCDO}$ and $U_{ENS}$ measure exhibited comparable FN recall rates, while the $U_{PE}$ exhibited the lowest FN recall rate. This implies the $U_{TTA}$ uncertainties within FN regions were relatively high and could more readily be used to recover FN regions compared to other measures. In general, the uncertainties in FN regions were small relative to those in TP regions. For example, at the 0.5 factor threshold, the recall rate was only 0.33 for $U_{TTA}$, meaning, only approximately one-third of the FN regions had higher uncertainty than half of the median of TP uncertainties. At a factor of 0.1, approximately one-half of FN regions had greater uncertainty than 10% of the median value for TP uncertainties. Still, the recall rates for the $U_{PE}$ measure were much smaller compared to the other measures, which indicates the need for auxiliary uncertainty methods for FN region recovery. FN recovery analysis using region-wise uncertainties has not previously been reported. However, previous work assessed the distributions of voxel-wise uncertainties found within TP and FN voxels (Huynh et al., 2024). In this work, FN voxels exhibited higher uncertainties than TP voxels. Conversely, our analysis demonstrated FN regions generally had smaller uncertainty than TP regions across uncertainty measures. This discrepancy may be due to the aggregation of voxel-wise uncertainties into region-wise uncertainties, where voxels dominated by low uncertainties within the ground-truth-defined FN regions may have greatly reduced the region-wise uncertainty. Since the detection of FN regions may be critical for certain clinical tasks for managing patients with metastatic malignancies, such as treatment selection and response assessment, future work should be dedicated towards developing uncertainty measures which more accurately detect these regions.

In the correlations with model performance assessment, $U_{PE}$ achieved the best correlation with the cross-entropy metric. This is understandable since both are measures of entropy. However, the $U_{PE}$ measure achieved the worst correlation with all other metrics. This implies that high correlation with cross entropy does not guarantee high correlation amongst other metrics. Furthermore, the low correlation of the $U_{PE}$ measure in the biomarker extraction accuracy and Dice coefficient metrics implies the use of auxiliary uncertainty measures is necessary to obtain a stronger correlation between model predictive uncertainty and performance. The $U_{TTA}$ measure outperformed or performed comparably to other measures across all other performance metrics. This result agrees with the work of Wang and colleagues where test time augmentation achieved better correlations with segmentation performance than Monte Carlo dropout for the MRI brain tumor segmentation task (G. Wang et al., 2019). This work, however, only assessed correlations between uncertainty and segmentation performance, omitting assessing correlations with more clinically relevant metrics. Most previous works only assess correlations between an uncertainty measure and segmentation Dice coefficient (Lambert et al., 2024). Conversely, we assessed the correlations between both segmentation performance metrics (i.e., Dice coefficient and cross entropy) and biomarker extraction accuracy (i.e., $SUV_{mean}$ and $SUV_{total}$), which are more clinically relevant. In general, our results showed stronger correlations for the segmentation metrics than for the biomarker metrics. For example, the lowest correlation coefficients were observed for the $SUV_{mean}$ accuracy metric ($\rho = 0.35$ was the maximum correlation coefficient for this metric). The discrepancy in the correlation performance of biomarker accuracy and segmentation performance metrics may be due to aggregating uptake information from predicted lesion voxels into a single measure. If some small yet critical uptake values (e.g., high uptake voxels) were not sufficiently captured, then this can negatively skew the correlation performance. The discrepancy further implies that strong segmentation performance does not guarantee strong biomarker extraction performance, and thus, model output and predictive uncertainty performance should be assessed with clinical endpoints in mind. For example, the

relationship between segmentation uncertainty and dose deposition has been evaluated for radiation therapy applications (Nomura et al., 2020). However, these types of assessments are not common despite the apparent imminent adoption of deep learning models in applications such as radiation therapy (Lastrucci et al., 2024). Correlations with model performance is one of the more common quantitative evaluations of UQ measures for medical image tasks (Lambert et al., 2024). Most previous works, however, report the Pearson correlation coefficient which assumes a linear relationship between uncertainty and performance. It is not clear that uncertainty should scale linearly with model performance. Therefore, we reported the Spearman correlation coefficient, which does not make a linearity assumption.

This superior performance of the $U_{TTA}$ measure across different imaging tasks may be attributable to the type of uncertainty it is targeting. While test time augmentation captures aleatoric uncertainty, Monte Carlo dropout and deep ensembles capture both aleatoric and epistemic uncertainty (Lambert et al., 2024). The strong performing of the $U_{TTA}$ measure suggests that the lesion segmentation task is dominated by aleatoric uncertainty. Likely, the majority of this aleatoric uncertainty arises from noise in the metastatic disease labels, which can be difficult and ambiguous to acquire. Still, some utility was demonstrated by the $U_{MCDO}$ and $U_{ENS}$ measures. More work should be dedicated to understanding the nuances and use cases for each UQ method. The $U_{TTA}$ measure is also advantageous due to being post hoc. As a result, it can easily be augmented to previously trained models with minimal computational overhead (Table 7).

This work establishing the optimal UQ method for metastatic lesion segmentation holds many clinical and technical implications. The $U_{TTA}$ measure could readily be propagated into segmentation-based clinical assessments to acquire assessment-specific uncertainty information. For example, doing so for assessments such as patient response assessment or outcome predictive modeling will guide

clinicians in deciding how much to rely upon automated and predictive outputs. Additionally, patients with metastatic malignancies may undergo stereotactic body radiation therapy to enhance disease burden reduction (Sharabi et al., 2017; Z. Wang et al., 2022). Deep learning-based methods may aid in metastatic lesion segmentation for radiation therapy targeting, greatly reducing labeling times. Employing UQ methods in this workflow may guide the review of automatically generated metastases segmentations. Moreover, the uncertainty information may be used for better target definition and enable novel treatment planning strategies such as robust optimization (Chu et al., 2005). Lastly, the $U_{TTA}$ measure may be used to curate more robust segmentation models using limited-sized datasets through specialized data curation frameworks such as active learning (Budd et al., 2021).

A challenge of this work was defining quantitative evaluations for comparing uncertainty measures. Standard quantitative assessments have not been established for uncertainty quantification. As a result, many studies implement different assessments that may not translate across tasks. Therefore, we implemented several quantitative assessments to more comprehensively characterize and compare each uncertainty measure. Standardizing uncertainty assessments using images with known and clinically relevant uncertainty sources will expedite progress in this field. Another challenge of this work was the lack of UQ measure investigation using different segmentation model frameworks. However, the exploration of more advanced segmentation methods, such as transformer-based models, was not as warranted at this time because many studies related to metastatic lesion segmentation utilize the nnUNet framework. Similarly, this work could be augmented to include additional UQ methods, however, the selected methods were among the most widely investigated and relevant to metastatic lesion segmentation. On the implemented UQ methods, parameter tuning was not performed. UQ method parameters were either selected from a review of the literature, or due to practical constraints for the task and data at hand (e.g., utilizing more than three trained models for the $U_{ENS}$ measure would

be very costly due to the need for five-fold cross validation). Alternative selection of UQ parameters may impact uncertainty performance. For example, in the case of the $U_{TTA}$ measure, incorporating more advanced augmentations—especially those that simulate expected image quality fluctuations, such as intensity shifts or additive noise—may more accurately capture data uncertainties and generate better uncertainty measures. Future work should explore the impact of segmentation model selection, the assessment of additional UQ methods, and the effect UQ parameter tuning on UQ measure performance. This work was further challenged by the relatively small and single modality dataset used for model training and testing. We invoked five-fold cross validation to mitigate this concern. In addition, we hypothesized that the concern regarding our limited dataset was mitigated by the large number of lesions across images ($n = 2322$). However, future work should be dedicated to expanding the dataset size, utilizing a dedicated test dataset, including data from different clinical sites, and investigations into alternative radiotracers and malignancies. In addition, only single-observer labels were available and incorporated in this study. The use of multi-observer labels may reduce data annotation noise, can establish inter-observer variability, and may be leveraged to better quantify uncertainty. Future work acquiring and utilizing multi-observer labels for segmentation and UQ evaluation should be pursued. Lastly, unlike related UQ works (Huynh et al., 2024; Nair et al., 2020), we did not include voxel-wise UQ assessments in this work. This was primarily because clinical PET assessments generally rely on either region-wise or image-wise metrics (Wahl et al., 2009), rendering voxel-wise assessments less interpretable and impactful.

While this was the first work to explore UQ for the metastatic lesion segmentation task, many of the conclusions from this work may be translated to other tasks and imaging modalities. Test-time data augmentation may similarly be the ideal uncertainty method in other applications where high aleatoric uncertainty is a concern. For instance, segmentation of the clinical target volume for radiation therapy is

likely to be subject to high aleatoric uncertainty due to a lack of visible structure definition. While uncertainty quantification has been investigated for this task (Balagopal et al., 2021), test time augmentation has not been tested. However, not all the evaluations in this work may translate into different tasks such as large structure (e.g., primary tumor or organ) segmentation, where detecting false positives may not be critical, for example. Alternatively, task-specific UQ evaluations and comparisons would be necessary. Each of the assessed uncertainty measures could also be implemented for more distant tasks such as image classification or image reconstruction. However, the relative performance of each uncertainty measure for these tasks is not known *a priori*, as each task is uniquely influenced by independent sources of uncertainty that are captured differently by each UQ method. Thus, task-specific comparison studies should be explored in addition to investigations into novel combinations of uncertainty measures for more robust quantification.

## 3.5 Conclusion

Across uncertainty quantification evaluations, the overall best-performing uncertainty method for the metastatic lesion segmentation task was test time augmentation ($U_{TTA}$). In addition to its superior UQ performance, a critical advantage of the test time augmentation method is that it is post hoc (i.e., it can be implemented on a previously trained model), and it requires a lower computational load to implement when compared to other established methods. Thus, acquiring reliable uncertainty information is relatively straightforward to implement using the test time augmentation method and should be strongly considered when evaluating and deploying deep learning models for whole body metastatic lesion segmentation.

# 4 Development of a Gradient-based UQ Measure

This chapter addresses **Specific Aim 2b** by introducing and evaluating the utility and performance of a novel, gradient-based approach for in-distribution uncertainty quantification. This approach has many features that address major limitations of previously established UQ methods: it is post hoc, computationally efficient, and does not change a trained model's output space. This is accomplished by uniquely leveraging localized gradient information from a trained model to establish the uncertainty of individually localized prediction regions. This work was previously published in *Physics in Medicine and Biology* under the title "Uncertainty quantification via localized gradients for deep learning-based medical image assessments" (Schott et al., 2024). In addition, this work was published as a scientific abstract and presented at the virtual Society for Imaging Informatics in Medicine's (SIIM) Conference on Machine Intelligence in Medical Imaging (CMIMI) in 2022.

## 4.1 Introduction

### 4.1.1 Overview

A wide range of ID UQ methods exist, and several have been investigated for medical image analysis applications (see Section 1.4.2). These previously developed UQ methods, however, largely fail to meet at least one of two constraints for streamlined clinical implementation. First, most of these methods are not post hoc, meaning that they cannot be implemented on a previously trained model. Rather, their use necessitates complete model re-training. The only method that can be considered post hoc is test-time data augmentation, as implemented in (G. Wang et al., 2019). Second, all these methods will induce changes to the predictive outputs of a previously trained model. The performance of

deployed models will need to be re-validated because of this change in model outputs. UQ methods designed to be post hoc and to not change a previously trained model's output would be clinically advantageous because they can retrospectively be augmented to deployed models without the need for model re-training or performance re-validation. In addition, many previous UQ methods, such as Monte Carlo dropout, output voxel-wise uncertainty measures. The acquisition of region-wise uncertainty is important for many tasks such as metastatic disease delineation where a priority of the image analysis is to determine if suspicious lesions are actually malignant or not. Voxel-wise uncertainty measures may be aggregated into a region-wise metric; however, the appropriate aggregation strategy may not be straight-forward.

In this work, we present and validate a novel UQ method designed for deep learning tasks with spatially representative outputs (e.g., anatomical delineations). Unlike the previously established methods listed above, this method is designed to be post hoc and will not change a trained model's output (e.g., predicted delineations). Thus, this method is more suited for the direct integration into deployed clinical models. In addition, the method directly acquires a region-wise uncertainty measure, bypassing any voxel-to-region aggregation step. The method works by targeting a model's localized gradient space to derive a regional uncertainty measure. The model's backwards computational path is redirected such that model gradients are computed with respect to individual regions, localizing gradient information to specific regions. Therefore, we refer to this novel method as the *Local Gradients UQ* method. This is the first uncertainty quantification method for deep learning-based medical image assessment which utilizes information from a model's gradient space. While the Local Gradients UQ method can be applied to many medical imaging tasks, we demonstrate its utility for metastatic disease delineation tasks, as one of the more demanding tasks. The performance of each uncertainty measure is assessed in four clinically relevant experiments: (1) response to artificially degraded image quality, (2)

comparison between matched high- and low-quality clinical images, (3) false positive (FP) filtering, and (4) correspondence with physician-rated malignant disease likelihood.

## 4.2 Methods

### 4.2.1 Local Gradients Uncertainty Quantification Method

#### Mathematical Framework

The Local Gradients UQ method was designed for deep learning models with spatially representative outputs (e.g., anatomical delineation), as opposed to models with 1D vector outputs (e.g., a classification model). We first assumed that there is a previously trained deep learning model with learned parameters, $\theta$. We also assumed that there is a mechanism to localize individual regions in the model output. For instance, a predicted label map can be used for an anatomical delineation task. The goal of the Local Gradients UQ method is to quantify the sensitivity (or change) of a given localized region with respect to the learned model parameters. Mathematically, this can be expressed as the partial derivative of the localized region ($R$) with respect to the model's learned parameters as

$$\frac{\Delta R}{\Delta \theta} = \frac{\delta R}{\delta \theta}. \tag{13}$$

To carry out this computation, it is first necessary to formulate a scalar score that sufficiently describes each localized region, referred to as the *regional target function* ($T_R$). For this, we selected the Kullback-Leibler (KL) divergence between the model's predicted class *softmax* output within a localized region ($p = \{p_i\}$) and a reference distribution ($q = \{q_i\}$), normalized by the number of voxels within a given localized region, $N$

$$T_R = \frac{1}{N} D_{KL}(p \parallel q) = \frac{1}{N} \sum_{i}^{N} p_i \log\left(\frac{p_i}{q_i}\right). \quad (14)$$

The KL divergence is a measure of similarity between $p$ and $q$. We set the reference distribution to be uniform $\boldsymbol{q} = \left[\frac{1}{C}, \frac{1}{C}, \ldots, \frac{1}{C}\right] \in \mathbb{R}^N$ of length $N$ (number of voxels in region), where $C$ is the number of output classes (e.g., $C = 2$ for the binary delineation task). This uniform distribution simulates a region made up of voxels with ambiguous classifications. Therefore, for localized regions dominated by high *softmax* probabilities (high confidence), $T_R$ will be large and vice versa.

Substituting $T_R$ from equation $14$ into $R$ from equation $13$, formulated the sensitivity of a localized region with respect to the model parameters as

$$\frac{\delta T_R}{\delta \theta} = \frac{\delta \frac{1}{N} D_{KL}(p \parallel q)}{\delta \theta} = \frac{\delta \frac{1}{N} \sum_i^N p_i \log\left(\frac{p_i}{q_i}\right)}{\delta \theta}. \quad (15)$$

$T_R$ was individually backpropagated for each localized region which populated the gradient information of $\theta$. Practically speaking, these gradients were accessed using the "backwards hook" capabilities of the PyTorch deep learning library. Rather than collecting gradient information from all model parameters, which may be computationally expensive, we selected a subset of model parameters, $\varphi$, from which to retrieve gradient information. Gradient information from the selected model parameters was aggregated into a single scalar via the $L_1$-norm. The Local Gradients score ($LG$) for a single localized region $R$ within a test image $x$ can be written as

$$LG(x)_R = \left\| \frac{\delta T_R}{\delta \varphi} \right\|_1 = \left\| \frac{\delta \frac{1}{N} \sum_i^N p_i \log\left(\frac{p_i}{q_i}\right)}{\delta \varphi} \right\|_1, \quad (16)$$

where $\|\cdot\|_1$ denotes the $L_1$-norm.

To enhance the interpretation of the Local Gradients score, we normalized it to define the Local Gradients uncertainty measure as

$$U(x)_{LG,R} = \frac{LG(x)_R}{P_{95}\{LG_{low}\}}, \tag{17}$$

where $P_{95}\{LG_{low}\}$ denotes the 95$^{th}$ percentile of a set of Local Gradients scores calculated from model outputs on a validation dataset with a priori assumed low uncertainty. At this percentile, the majority of outputs with assumed low uncertainty are captured while still accounting for potential, high uncertainty outliers.

Under this formulation, it was assumed that regions with high uncertainty will yield a large gradient response, making $U(x)_{LG,R}$ large, and regions with low uncertainty will yield a small gradient response, making $U(x)_{LG,R}$ small. This process was repeated for each localized region within a given test image. An outline of the Local Gradients UQ algorithm applied to the metastatic disease delineation task is shown in Figure 25. A key advantage of the Local Gradients UQ algorithm is that it does not require any image ground-truth data to implement as it only relies on the model's outputted probabilities ($p$) for a predicted region and a user-defined distribution ($q$) to compare the model probabilities to the *regional target function ($T_R$)*.

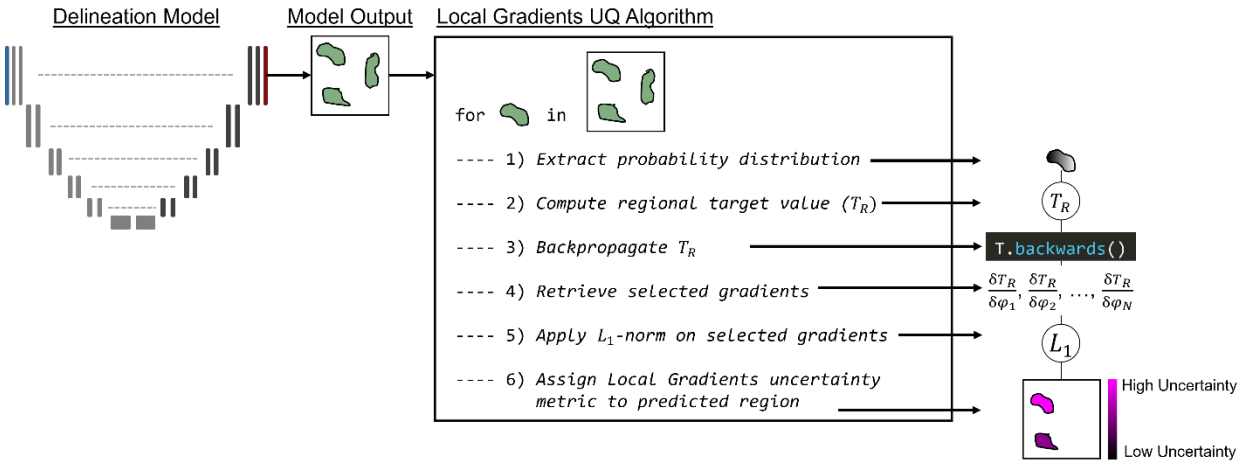



*Figure 25 Schematic description of the Local Gradients UQ algorithm applied to the metastatic disease delineation task. For each delineated (localized) region predicted by the deep learning model, the steps to acquire the uncertainty measure are as follows: 1) extract the softmax probability values within the predicted region, 2) compute the regional target value $T_R$, 3) backpropagate $T_R$ to populate gradient information on model parameters, 4) retrieve gradient information from the selected model parameters, 5) apply the $L_p$-norm to aggregate the gradient information into a single value, 6) assign the Local Gradients UQ measure to the current region to populate the Uncertainty Map.*

### Targeted Model Parameters

When computing the Local Gradients UQ measure, gradient information can be acquired from any learnable parameter in a trained model. In this work, we used a U-Net model which consists of an encoder, a decoder, and equal-resolution skip connections (Ronneberger et al., 2015). Each resolution level of the encoder and decoder contained two convolutional blocks, each consisting of $N$ 3D convolutions using a $3 \times 3 \times 3$ kernel, instance normalization, and ReLU activation operations. Both model training and the Local Gradients UQ method were implemented using the PyTorch deep learning library. The Local Gradients UQ measure was obtained by targeting gradient information from all learnable parameters within a given decoder convolutional block. Targeting a U-Net model's decoder parameters should also capture sensitivities in the mode's encoder side because encoder activations are passed to the decoder side via skip connections. Thus, we only specifically selected to target decoder parameters in this work. As conveyed in Figure 26, we denote the targeted decoder blocks using the

resolution level starting from the bottleneck layer (i.e., levels 0-4) and level pair ordering (i.e., pair 0 or 1). For instance, we use 'Blocks_4-0' to indicate gradients that are acquired from the first (0, zero-based indexing) convolutional block of the fifth (4) resolution layer of the decoder. In all experiments, parameters from decoder Block_4-0 and Block_4-1 were used for the Local Gradients UQ measure. A sensitivity study was performed in Section 4.3.3 to determine the utility of targeting different decoder block configurations.

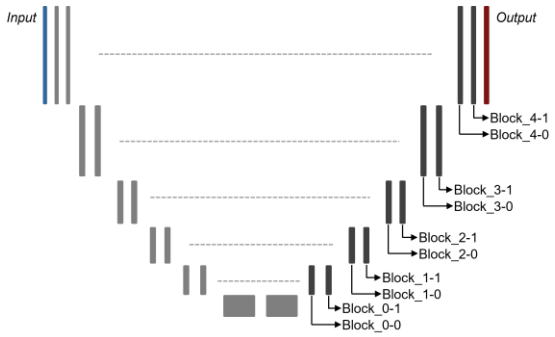



*Figure 26 Schematic showing the architecture of the lesion delineation model with decoder convolutional block labeling.*

## Uncertainty Measures for Comparison

We compared the Local Gradients UQ measure to two measures derived from the *softmax* probability outputs within predicted regions. These measures were selected for comparison because, like the Local Gradients UQ measure, they are post hoc and will not change a trained model's output and some form of model probability value or entropy have been investigated as UQ baselines previously (Diao et al., 2022; Jungo et al., 2020). First, we compared to the mean of the *softmax* predicted class probability within a predicted region as in

$$MP(x)_R = \frac{1}{N}\sum_{i}^{N} p_i \,. \tag{18}$$

Second, we compared the Local Gradients UQ measure to using equation 14 on its own, which is simply the KL divergence without any associated gradient information described by

$$KLD(x)_R = \frac{1}{N}\sum_{i}^{N} p_i \log\left(\frac{p_i}{q_i}\right). \tag{19}$$

The direction of the uncertainty measure scales in equations 18 and 19 are opposite to the Local Gradients UQ measure, where high magnitudes denote high uncertainties. To account for this difference and to facilitate more interpretable comparisons across all uncertainty measures, we normalized $MP(x)_R$ and $KLD(x)_R$ in the same manner as the Local Gradients UQ measure (equation 17) and reversed their scales according to

$$U(x)_{MP,R} = -\frac{MP(x)_R}{P_{95}\{MP_{low}\}} \tag{20}$$

$$U(x)_{KLD,R} = -\frac{KLD(x)_R}{P_{95}\{KLD_{low}\}}. \tag{21}$$

Under this formulation, it is assumed that all uncertainty measures will be small for low uncertainty predictions, and large for high uncertainty predictions. For brevity $U(x)_{MP,R}$, $U(X)_{KLD,R}$, and $U(x)_{LG,R}$ are referred to as, $U_{MP}$, $U_{KLD}$, and $U_{LG}$ in all subsequent experiments, respectively.

### 4.2.2 Datasets

Two datasets were used to evaluate the performance of the Local Gradients UQ measure. We selected datasets of patients with one of two of the most common malignant metastases sites—liver and bone— which together account for almost half of all metastases(Riihimäki et al., 2018). The first dataset

consisted of abdominal CT scans of patients with liver metastases and the second dataset consisted of $^{18}$F-NaF PET/CT scans of patients with bone metastases. The selected datasets are especially relevant for investigating UQ because of their clinical importance. Each dataset also presents unique image interpretation challenges which are likely to induce greater uncertainty. For instance, CT acquisition can be non-standardized between imaging sites, scanners, patient cohorts etc. This non-uniformity is likely to convolute the model fitting process. Similarly, $^{18}$F-NaF PET/CT scans are often difficult to interpret due to the high incidence of benign uptake in areas of high osteoblastic activity (e.g., arthritic joints) (Even-Sapir et al., 2006; Iagaru et al., 2012; Sheikhbahaei et al., 2019). Thus, the model will be forced to learn the subtle differences between metastatic malignant and benign bone lesions, introducing an added layer of complexity and uncertainty.

### Dataset 1 – Liver Metastases

The imaging dataset of liver metastases consisted of both publicly and institutionally available data. The public data was acquired from the LiTS liver tumor segmentation challenge (Bilic et al., 2023). This is a diverse dataset consisting of CT scans from seven institutions acquired across a variety of scanners, scan acquisition protocols, and patient metastatic malignant pathologies. All scans were considered diagnostic level with high image resolution and contrast enhancement. We used the 131 available training scans with corresponding ground-truth data. Across all scans, the liver tumors were contoured and independently verified by trained radiologists. In the 131 scans, a total of 908 tumors were contoured (per patient median: 3, range: [0, 75]).

The institutional dataset consisted of patients with metastatic neuroendocrine tumors (NETs) presenting on the liver and was used only for model testing and UQ assessment. All patients were treated at the University of Wisconsin Hospital and Clinics (UWHC) with peptide receptor radionuclide

therapy using $^{177}$Lu-DOTATATE. Prior to and throughout the course of treatment, patients received diagnostic level CT and $^{68}$Ga-DOTATATE PET/CT imaging for staging and treatment response assessment. We only used the attenuation correction CT scan component of the $^{68}$Ga-DOTATATE PET/CT scans in our analyses. We refer to the attenuation correction CT scans as the *institutional-attenuation correction* (AC-CT) test set and the diagnostic-level CT scans as the *institutional-diagnostic* (CE-CT) test set. The median time between the acquisition of AC-CT and CE-CT scans for each patient was 4 months (range: [0, 30] months).

### Dataset 2 – Bone Metastases

The imaging dataset of bone metastases consisted of institutionally available data only. Patients were acquired from a previously conducted prospective study (NCT01516866) at the UWHC (Lin et al., 2016), using $^{18}$F-NaF PET/CT to assess response in patients with prostate cancer metastasized to bone. For the present work, we retrospectively selected all patient scans acquired before the initiation of a systemic treatment. Scans were acquired across three different sites on either a Discovery VCT (GE Healthcare, Waukesha, WI) PET/CT scanner or a Gemini (Philips Healthcare, Amsterdam, Netherlands) PET/CT scanner. All PET scans were scatter and attenuation corrected, and scans between different scanners were harmonized as in (Jallow, 2006.; Lin et al., 2016).

A total of 37 baseline $^{18}$F-NaF PET/CT images were acquired of which a nuclear medicine physician manually contoured a total of 1,833 bone lesions (per patient median: 42, range: [14, 123]). The same nuclear medicine physician classified all lesions on a five-point scale defined as (1) definitely malignant, (2) likely malignant, (3) equivocal, (4) likely benign, and (5) definitely benign. The distribution of the contoured lesions across the five-classes is summarized in Table 11a. For training the lesion delineation model, these classes were condensed to a 3-point scale, with distributions shown in Table

1b. A maximum intensity projection (MIP) image of an example $^{18}$F-NaF PET scan with overlaid physician contours and disease classifications is shown in Figure 27.

*Table 11 Distribution of physician likelihood lesion classifications for Dataset 2 – Bone Metastases. Lesions were classified on a 5-point scale in the original ground-truth data (a). For training the base lesion delineation model, lesion classes were condensed to a 3-point scale (b).*

| ***(a) Ground-Truth Classes*** | | | | |
|---|---|---|---|---|
| Definitely Malignant (Class 1) | Likely Malignant (Class 2) | Equivocal (Class 3) | Likely Benign (Class 4) | Definitely Benign (Class 5) |
| 863 | 212 | 85 | 213 | 224 |
| ***(b) Training Classes*** | | | | |
| Malignant (Class 1) | | Equivocal (Class 2) | Benign (Class 3) | |
| 1075 | | 85 | 437 | |

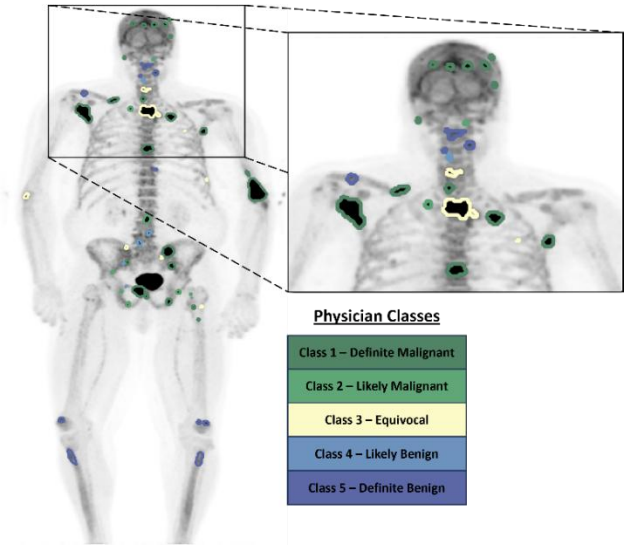



*Figure 27 A maximum intensity projection image of an example patient with bone metastases imaged with $^{18}$F-NaF PET and overlaid physician delineations with disease likelihood classifications.*

### 4.2.3 Lesion Delineation Model

The Local Gradients UQ method was designed to be implemented on a previously trained model. Thus, we first trained a model for metastatic malignancy delineation in each of the imaging datasets. The two models were trained in-house using the *nnUNet* repository (Isensee et al., 2021), a powerful tool for

biomedical image delineation which follows a standard U-Net architecture (Ronneberger et al., 2015), which is routinely used across a wide variety of medical image delineation tasks. When using *nnUNet*, a subset of the training hyperparameters are automatically and heuristically selected given characteristics of the training data and available computer infrastructure. The *nnUNet* model has achieved highly competitive performance on a variety of delineation tasks , making it well-suited to assess any UQ method because uncertainty due to poor model design and training will be minimized.

A single, full-resolution 3D model was trained for each dataset and used across all experiments. The model loss function was set as the sum of Dice Coefficient and Cross Entropy, and the model was trained for 1000 epochs using a batch size of 2 and 250 minibatches per epoch. Data augmentation was applied during training across all experiments using the following techniques: rotation, scaling, gaussian noise, gaussian blur, brightness, contrast, low resolution simulation, gamma augmentation, and mirroring. The image patch size used during training varied across experiments due to the differences in training data size and resolution. All models were trained on an RTX Titan GPU workstation with 25 GB of memory.

### Image Pre- and Output Post-Processing

Prior to training, all training images were normalized to zero-mean-unit-variance and were resampled to a common voxel size using linear interpolations. For *Dataset 2 – Bone Metastases*, we did not consider lesions that were located in the patients' hands because the physician did not classify suspicious lesions in the hands. All predicted regions with a volume smaller than 0.25 $cm^3$ were removed as in post-processing.

### Liver Metastases Model Training

For the experiments on *Dataset 1 – Liver Metastases*, we trained an *nnUNet* model to delineate both the liver and liver lesions on CT. 105 of the 131 (80%) of the LiTS images were used for model training. All images were resampled to a common voxel spacing of 1.5 $mm^3$, and an image patch size of [112, 128, 160] was used during training. Cross validation was not necessary for this model given the large amount of training data available.

**Bone Metastases Model Training**

For the experiments on *Dataset 2 – Bone Metastases*, we trained an *nnUNet* model for metastatic bone lesion delineation and classification. For this purpose, we collapsed all malignant and benign disease classes (including definite and likely categories) from the five-class physician-rated disease likelihood classifications into a three-class scale defined as: (1) malignant, (2) equivocal, and (3) benign and trained a model to delineate and classify lesions along this three-class scale. The distribution of annotated lesions across these three classes is displayed in Table 11b. Due to the relatively small dataset size, five-fold cross validation was used during training with 80%/20% train/test splits to obtain test predictions on each of the 37 baseline $^{18}$F-NaF PET/CT scans. All images were resampled to a common voxel spacing of 2.5 $mm^3$, and an image patch size of [128, 64, 288] voxels was used during training. The lesion delineation performance of the *nnUNet* model in this dataset has previously been reported in (Schott et al., 2023).

### 4.2.4 Experiments

**Response to Artificially Degraded Data**

In this first experiment, we investigated the Local Gradients UQ measure's response to artificially degraded image quality using data from *Dataset 1- Liver Metastases*. Fundamentally, this was

implemented as a proof-of-concept study to observe if the uncertainty measures followed expected behavior. We assumed that a reliable uncertainty measure should increase with a decrease in image quality, making the images harder to visually interpret, and consequently the predictions more uncertain. The trained liver lesion CT model was used for inference on the test data split from the LiTS dataset (N=27). For each test image $I \in \mathbb{R}^{n \times m \times z}$, artificially degraded images were generated using the following degradation methods with varying magnitudes ($\sigma$):

1) Additive Gaussian Noise: $I_{Noised} = I + W$
2) Additive Speckle Noise: $I_{Noised} = I + W * I$
3) Gaussian Smoothing: $I_{Smoothed} = I \otimes K$

where $W \sim \mathcal{N}(\mu = 0, \sigma)$ is an $n \times m \times z$ gaussian noise matrix with 0-mean and $\sigma$-standard deviation, $K \sim \mathcal{N}(\mu = 0, \sigma)$ is a gaussian smoothing kernel with 0-mean and $\sigma$-standard deviation, $*$ is the point-wise multiplication operation, and $\otimes$ is the convolution operation. Degradation magnitudes ranged from $\sigma_{GN} = [0, 70]$ voxels for additive Gaussian noise, $\sigma_{SN} = [0, 1]$ voxels for additive speckle noise, and $\sigma_{GS} = [0, 4]$ voxels for Gaussian smoothing, each in 11 steps. Uncertainty information was then computed for each predicted region in all the degraded test images. An example of one test image per degradation type and magnitude is shown in Figure 28.

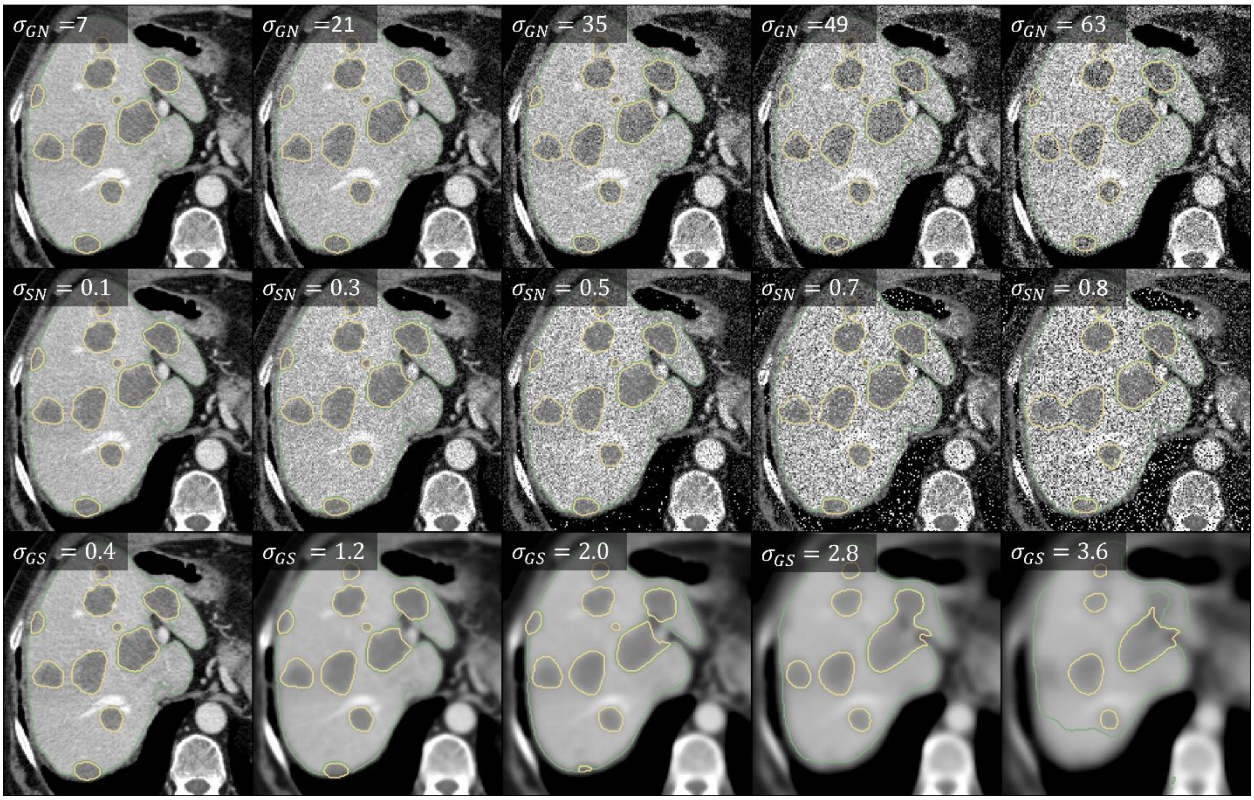



*Figure 28 An example test image axial slice across degradation types. Top Row: Additive Gaussian Noise ($\sigma_{GN}$). Middle Row: Additive Speckle Noise ($\sigma_{SN}$). Bottom Row: Gaussian Smoothing ($\sigma_{GS}$). Green and yellow contours indicate predicted liver organ and liver lesion delineations, respectively.*

Predicted regions were tracked across the increasing degradation magnitudes for each degradation type using a previously developed lesion tracking tool (Santoro Fernandes, et al., 2024). This tool constructs matching correspondence between lesions and establishes lesion tracks across the non-degraded and degraded images across all degradation magnitudes for a single test image while accommodating the splitting and merging of lesions. Predicted region uncertainties were acquired for each region in each region track. In the case of a single region splitting into multiple regions at a given magnitude, a weighted average, weighted by region volumes, of the split regions' uncertainty measures was used. Thus, tracks of uncertainty measures across degradation magnitudes were constructed.

To assess the response of each uncertainty measure as a function of degradation magnitude, the percent difference of uncertainty measures of matched predicted regions across non-degraded and degraded images were computed. This was performed for each degradation type and magnitude on a given test image $x$ as

$$Percent\ Diff_{x,R,\sigma=i} = \frac{\tilde{x}_{R,\sigma=i} - \tilde{x}_{R,\sigma=0}}{\tilde{x}_{R,\sigma=0}} \times 100, \quad (22)$$

where $\tilde{x}_{\sigma=0}$ denotes the undegraded test image, and $\tilde{x}_{\sigma=i}$ denotes the degraded test image at degradation magnitude $i$, and $R$ denotes the predicted region where they uncertainty is being measured. This relative metric facilitates comparisons of uncertainties at differing degradation magnitude scales. When inferring on test images with increasing degradation magnitude, the delineation model may begin to miss lesions that it previously delineated on less degraded data. Thus, we restricted this analysis to persisting lesions (i.e., lesions that were delineated by the model across all degradation magnitudes).

### Comparison on Low- and High-Quality Clinical Data

In this second experiment, we again investigated the Local Gradients UQ measure's response to poor image quality and utilize data from *Dataset 1 – Liver Metastases*. Instead of artificially degrading image quality as in the first experiment, we observed the uncertainty measure's response to lower quality clinical image acquisitions. Again, we assumed that a reliable uncertainty measure should increase with a decrease in image quality. For this experiment, we compared uncertainty measures between matched lesions in the *institutional-diagnostic* (CE-CT) and the *institutional-attenuation correction* (AC-CT) datasets. First, inference was performed on both test datasets and uncertainty measures were computed for each predicted lesion. Subsequently, liver tumors were manually matched

across CE-CT and AC-CT scans for each patient and their uncertainties were compared. Anatomical variations in matching tumors across scan types are expected to be minimal since NETs progress very slowly (Dromain et al., 2019). Differences in the uncertainty measure between matched CE-CT and AC-CT scans were assessed using statistical significance (Wilcoxon signed rank test).

### False Positive Filtering

In this third experiment, we utilized *Dataset 2 – Bone Metastases to* investigate the Local Gradients UQ measure's ability to filter false positive (FP) from true positive (TP) predicted regions. Following the assumption in (Nair et al., 2020), we assumed that FP predicted regions are induced by high predictive uncertainty inherent to the deep learning model. Thus, it is expected that FP regions will have higher uncertainty than TP regions. In our evaluation, a FP region was defined as a predicted region which does not overlap with any corresponding lesion in the ground-truth data. The difference in uncertainty measure magnitude between FP and TP regions was investigated. Statistical significance between uncertainty measures from FP and TP predicted regions was established using the Wilcoxon rank-sum test. Additionally, the area under the receiver operator curve (AUC) and the false positive rate at the 95% sensitivity threshold (FPR95) for distinguishing FP from TP regions using the uncertainty measure were computed.

Within the FP filtering experiment, we performed a sensitivity study to determine the optimal gradient information to leverage when using the Local Gradients UQ method. We investigated FP filtering performance of the Local Gradients UQ method when targeting different configurations of convolutional blocks in the model's decoder. We performed this sensitivity study using the KL divergence and the mean probability of a predicted region as regional target functions to determine if the selection of the regional target function is critical for the methodology. Using gradient information from the KL divergence

calculation, as in Figure 25, is equivalent to the Local Gradients UQ measure ($U_{LG}$). To use gradient information from the mean probability value within a predicted region, we set $T_R = MP(x)_R$ in equation $14$ and followed the subsequent steps of the Local Gradients UQ algorithm Figure 25. We refer to this modification of the Local Gradients UQ measure as $U_{MP,Gradients}$. Finally, both regional target functions were assessed without gradient information ($KLD(x)_R$ for KL divergence, as in equation $19$, and $MP(x)_R$ for mean probability, as in equation$18$) to assess the added utility of the gradient space for each regional target function.

### Correspondence With Physician-rated Disease Likelihood

In the fourth experiment, we investigated the Local Gradients UQ measure's correspondence with physician-rated disease likelihood. We utilized the data from *Dataset 2 – Bone Metastases*, and we expected the uncertainty measures to follow that of a human observer. For instance, when a human observer is very certain a given region belongs to a given class, the corresponding uncertainty measures of that region should be low. This model used in this experiment was trained to delineate and classify lesions on the three-point scale: (1) malignant disease, (2) equivocal disease, and (3) benign disease. Predicted regions were then grouped according to the corresponding ground-truth classification along the five-point scale of physician-rated likelihood of disease classification (Table 11). In the case of multiple classes being matched to a single predicted region, the class with the majority of voxels within a region was selected. Under this setup, it is expected that predicted lesions should yield uncertainty measures in increasing magnitude across physician-rated *definitely*, *likely*, and *equivocal* disease classes (for both malignant and benign disease groups). Quantitative analysis is performed separately for malignant and benign groups, where the equivocal class is included in both. Differences between each physician-rated class were established using the Wilcoxon rank-sum test, AUC scores, and median percent differences defined as

$$Percent\ Diff_{A,B} = \frac{U_A - U_B}{U_B} \times 100, \quad (23)$$

where $U_A$ and $U_B$ denote the uncertainty measures for two arbitrary groups for comparison $A$ and $B$ (e.g., definitely malignant lesions ($A$) and likely malignant lesions ($B$)).

### 4.2.5 Uncertainty Measure Normalization

Each uncertainty measure was normalized according to equations 17, 20, and 21. The normalization constant for each uncertainty measure ($P_{95}\{LG_{low}\}$, $P_{95}\{KLD_{low}\}$, and $P_{95}\{MP_{low}\}$ ) was acquired from the distribution of uncertainty measures for predicted regions with expected low uncertainty. Normalization constants were specific to each model trained. For the liver metastases model, the normalization constant was defined using the uncertainty measures of the predicted lesions in the non-degraded LiTS validation data. For the bone metastases model, we used the uncertainty measures of the TP predicted regions to define the normalization constant.

## 4.3 Results

### 4.3.1 Response to Artificially Degraded Data

The liver metastases model predicted 94 lesions on the 27 non-degraded LiTS validation images. The lesion tracking analysis generated 62, 54, and 29 persistent lesion tracks across all degradation magnitudes for additive Gaussian noise, additive speckle noise, and Gaussian smoothing degradations, respectively.

Figure 29 shows the median percent difference between uncertainty measures of matching lesions on non-degraded and degraded images at increasing degradation magnitudes. For all three

degradation types, the $U_{KLD}$ measure responded stronger to the degraded image data than the $U_{MP}$ measure, where the $U_{KLD}$ measure median percent difference between the non-degraded and most degraded images was 2.16%, 1.78%, and 3.48% for the additive Gaussian noise, additive speckle noise, and Gaussian smoothing degradations, respectively. The $U_{LG}$ measure's response to artificial image degradation was much stronger than both the $U_{MP}$ and $U_{KLD}$ measures. The median percent differences of the $U_{LG}$ measure between the non-degraded images and most degraded images were 33.41%, 27.73%, and 62.35% for additive Gaussian noise, additive speckle noise, and Gaussian smoothing, respectively.

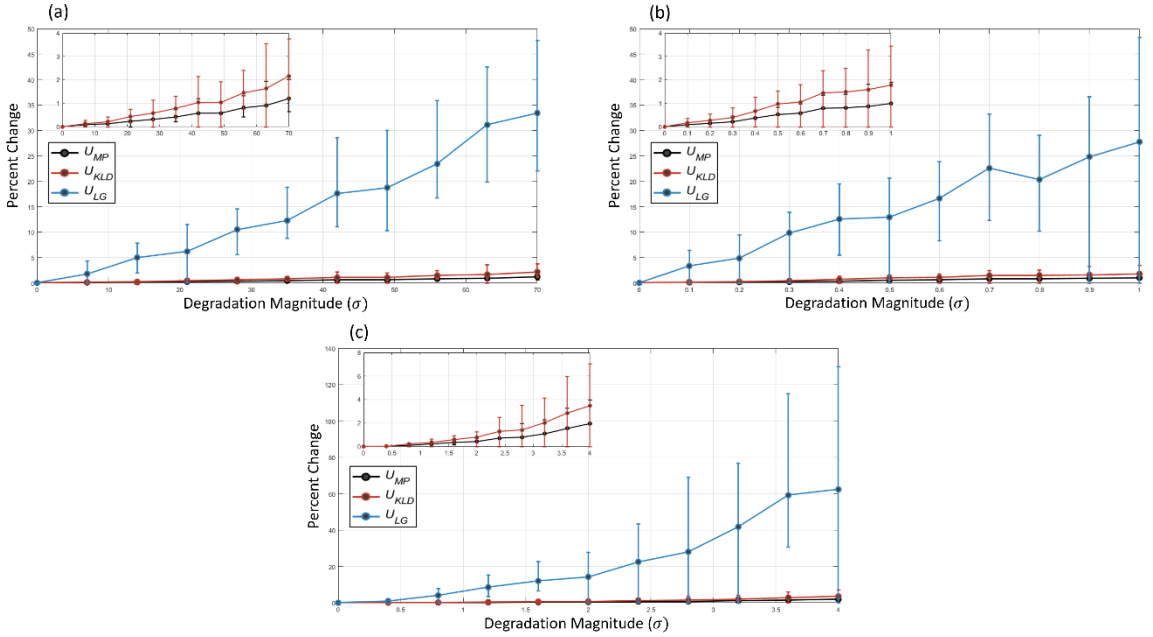



*Figure 29 The median percent difference between the uncertainties of predicted regions on non-degraded and degraded images as a function of image degradation magnitude for (a) additive Gaussian noise, (b) additive speckle noise, and (c) gaussian smoothing degradations. Inset figures show the uncertainty response of the UMP and UKLD measures on a smaller scale. Error bars indicate the interquartile range at each degradation magnitude.*

### 4.3.2 Comparison on Low- and High-quality Data

In the institutional liver metastases test set, a total of 158 and 78 lesions were delineated by the base CT liver organ and lesion model in the *institutional-diagnostic* (CE-CT) and *institutional-attenuation correction* (AC-CT) test scans, respectively. Of these lesions, 43 pairs were manually matched between CE- and AC-CT image pairs. The Local Gradients UQ measure was calculated for each lesion pair.

The distributions of paired lesions on CE-CT and AC-CT for each uncertainty measures are shown in Figure 30. No statistical evidence was found for a difference between predicted lesion uncertainty measures on CE-CT and AC-CT images for the $U_{MP}$ and $U_{KLD}$ measures. Conversely, the $U_{LG}$ uncertainty measure exhibited statistical significance between CE-CT and AC-CT groups ($p<0.05$), where the AC-CT lesions generally yielded larger uncertainty magnitudes than CE-CT lesions.

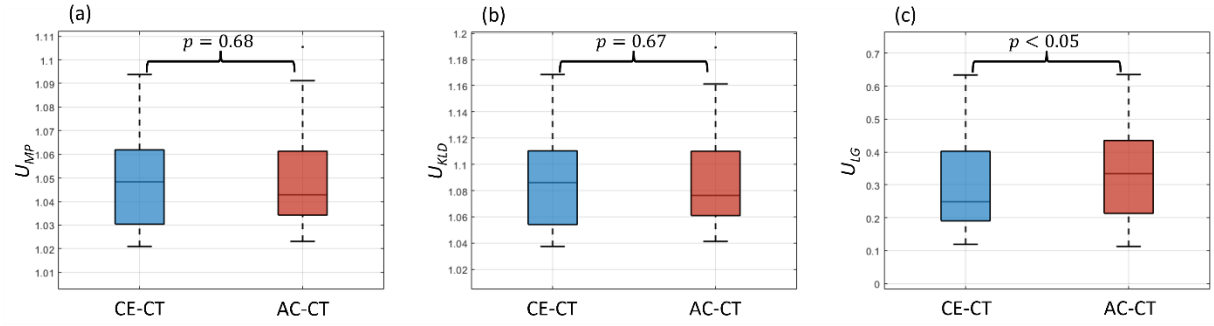



*Figure 30 Uncertainty measures of matched predicted liver tumor delineations in high-quality (Contrast Enhanced CT) and low-quality (Attenuation Correction CT) medical images. Comparisons between CE-CT and AC-CT are drawn between three uncertainty measures: (a) Mean Probability (UMP), (b) KL Divergence (UKLD), and (c) Local Gradients UQ (ULG).*

In Figure 31 we show two scenarios of matched CE-CT and AC-CT lesions with their corresponding $U_{LG}$ measures overlaid. Figure 31a shows an example with the expected pattern, where the lesions are much more visible on CE-CT than on AC-CT, and the $U_{LG}$ measures were higher on CE-CT than on AC-CT. In contrast, Figure 31b shows an example test scan where the $U_{LG}$ measure was higher on

CE-CT than on AC-CT for a matched lesion. However, the lesion on CE-CT appears more differentiated than on AC-CT, where it appears larger and more uniform.

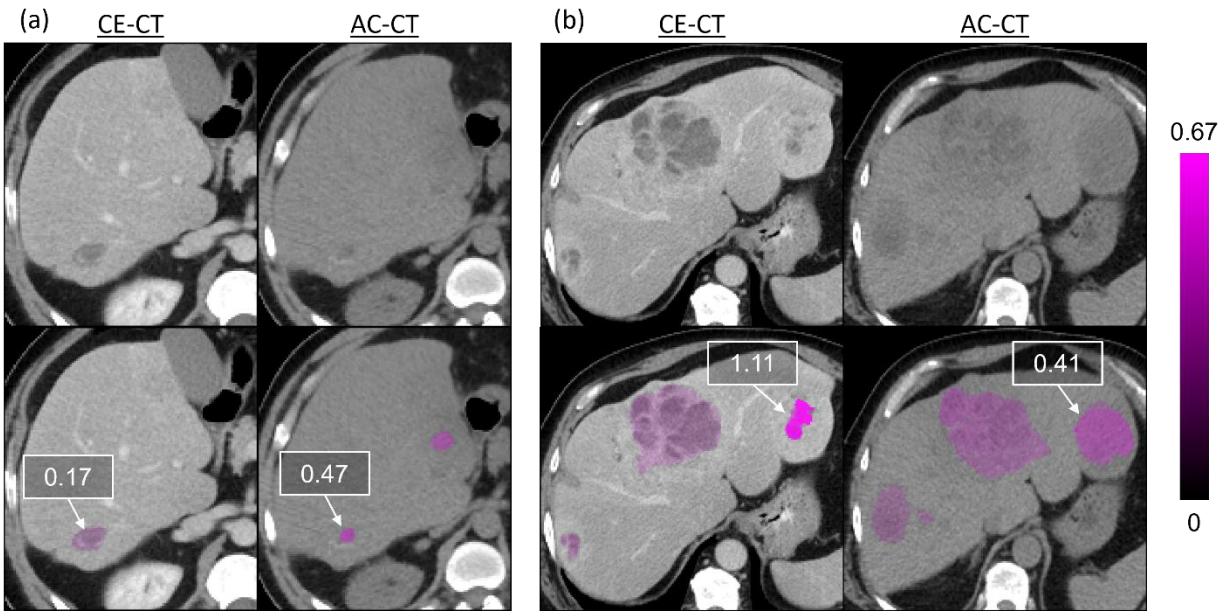



*Figure 31 A qualitative evaluation of two example test scans with Local Gradients UQ measures ($U_{LG}$) overlaid on predicted liver tumor regions. (a) shows an example where the liver tumor is more visible in the CE-CT and has a lower uncertainty measure than the corresponding tumor and uncertainty measure on the AC-CT. (b) shows an example where the uncertainty measure of a predicted tumor on the CE-CT has a higher uncertainty measure than the same tumor on the AC-CT. This discrepency could be explained by the more hetergeneous presentation of the tumor on the CE-CT than on the AC-CT. $U_{LG}$ uncertainty measure is listed in the box for each lesion. The color bar on the right indicates increasing $U_{LG}$.*

### 4.3.3 False Positive Filtering

The FP filtering performance of the two tested regional target functions ($T_R$) with and without gradient information is summarized in Table 2. Performance is shown for the different decoder blocks from which gradient information is acquired. We report the performance of each block individually and the combination of each block pair at each level of the decoder (e.g., Block 4-0, Block 4-1).

*Table 12 FP filtering performance of the Local Gradients UQ method when targeting gradient information from different decoder convolutional blocks (bottom). Two predicted regional target functions were tested: (1, right) using the KL Divergence of probability values and (2, left) mean probability value. The performance of using the target function without gradient information is also reported (top).*

| | Regional Target Function ($T_R$) | | | | | |
|---|---|---|---|---|---|---|
| | Mean Probability ($MP(x)_R$) | | | KL Divergence ($KLD(x)_R$) | | |
| Targeted Decoder Conv Block(s) | P-Value ↓ | AUC ↑ | FPR95 ↓ | P-Value ↓ | AUC ↑ | FPR95 ↓ |
| | *Without Gradient Information* | | | | | |
| N/a | <0.001 | 0.72 | 0.83 | <0.001 | 0.72 | 0.82 |
| | *With Gradient Information* | | | | | |
| | $U_{MP,Gradients}$ | | | $U_{LG}$ | | |
| Block 4-0 | <0.001 | 0.85 | 0.69 | <0.001 | **0.87** | 0.67 |
| Block 4-1 | <0.001 | **0.87** | 0.65 | <0.001 | **0.87** | 0.64 |
| Block 4-0, Block 4-1 | <0.001 | 0.86 | 0.65 | <0.001 | **0.87** | **0.61** |
| Block 3-0 | <0.001 | **0.87** | 0.68 | <0.001 | **0.87** | 0.67 |
| Block 3-1 | <0.001 | 0.85 | 0.74 | <0.001 | 0.86 | 0.70 |
| Block 3-0, Block 3-1 | <0.001 | **0.87** | 0.67 | <0.001 | **0.87** | 0.66 |
| Block 2-0 | <0.001 | **0.87** | 0.70 | <0.001 | **0.87** | 0.68 |
| Block 2-1 | <0.001 | **0.87** | 0.65 | <0.001 | **0.87** | 0.66 |
| Block 2-0, Block 2-1 | <0.001 | **0.87** | 0.65 | <0.001 | **0.87** | 0.66 |
| Block 1-0 | <0.001 | 0.85 | 0.77 | <0.001 | 0.85 | 0.77 |
| Block 1-1 | <0.001 | 0.86 | 0.76 | <0.001 | **0.87** | 0.73 |
| Block 1-0, Block 1-1 | <0.001 | 0.86 | 0.76 | <0.001 | 0.86 | 0.75 |
| Block 0-0 | <0.001 | 0.82 | 0.79 | <0.001 | 0.83 | 0.79 |
| Block 0-1 | <0.001 | 0.84 | 0.80 | <0.001 | 0.84 | 0.79 |
| Block 0-0, Block 0-1 | <0.001 | 0.83 | 0.70 | <0.001 | 0.83 | 0.79 |

FP filtering was enhanced with the inclusion of gradient information for both regional target functions ($MP(x)_R$ and $KLD(x)_R$) across all convolutional block configurations. When including gradient information, both $U_{MP,Gradients}$ and $U_{LG}$ achieved comparable FP filtering performance in terms of AUC, where the best AUC was 0.87 for multiple block configurations. $U_{LG}$, however, achieved the best FPR95 performance of 0.61 when using gradient information from decoder Block_4-0 and Block_4-1.

Figure 32 shows the distribution of TP and FP predicted regions for the $U_{MP}$, $U_{KLD}$, and $U_{LG}$ uncertainty measures and the ROC curves of each measure for filtering FPs from TPs, where the $U_{LG}$ uses gradient information from the optimal block configuration (decoder Block_4-0 and Block_4-1). All three uncertainty measures yielded statistical significance between TP and FP groups. The $U_{LG}$ measure,

however, achieved superior FP filtering performance over $U_{MP}$, and $U_{KLD}$, increasing AUC by 21% and decreasing FPR95 by 26%.

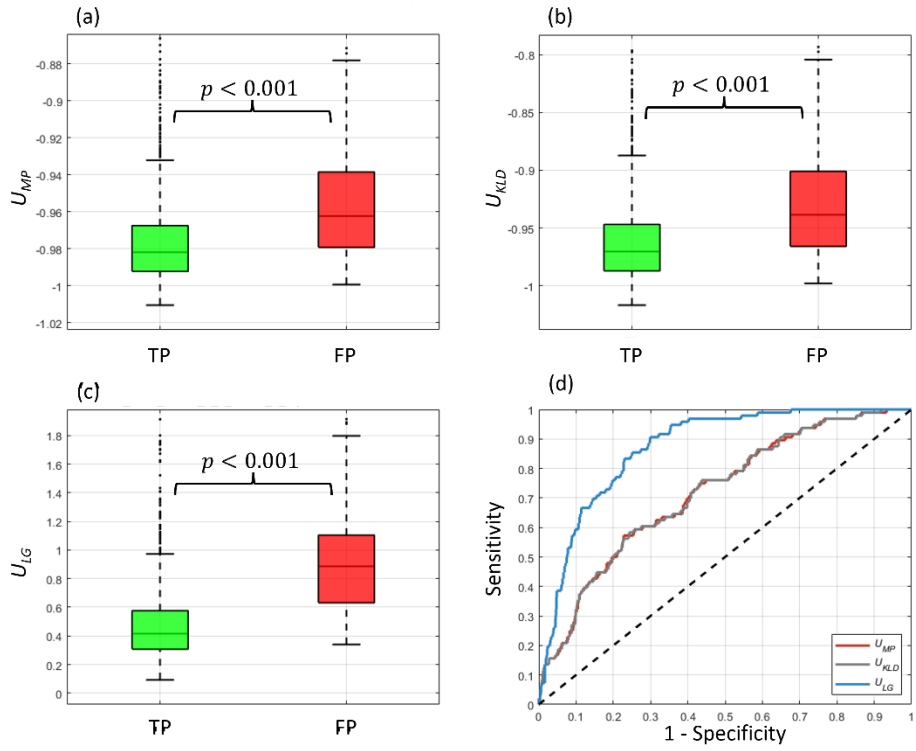



*Figure 32 TP and FP distributions of the three tested predicted region uncertainty measures: Mean Probability (a), KL Divergence (b), and Local Gradients UQ (c). (d) The ROC for filtering-out FP predicted regions using each uncertainty measure.*

### 4.3.4 Correspondence With Physician-rated Disease Likelihood

The correspondence between the tested uncertainty measures and physician-rated disease likelihood is shown in Figure 33. Statistical separation between classes in the malignant (i.e., classes 1, 2, and 3) and benign groups (i.e., classes 3, 4, and 5) for each tested uncertainty measure is reported in Table 13. According to the AUC scores, the median percent differences, and p-values between paired groups, the $U_{LG}$ measure achieved the best separation between disease classes and correspondence to physician-rated disease likelihood for the malignant classes. This is not the case, however, for the benign

classes where both the $U_{MP}$ and $U_{KLD}$ achieved better AUC scores and p-values. The median percent differences across benign groups were still better for the $U_{LG}$ measure.

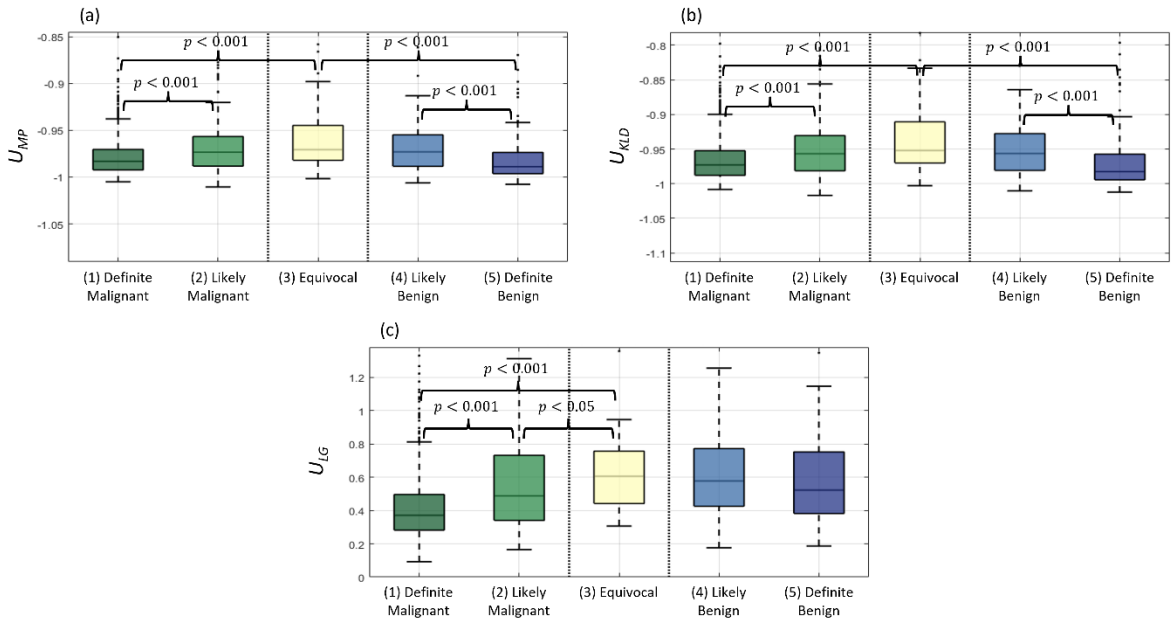



*Figure 33 Correspondence between uncertainty measures and physician-rated disease likelihood. Predicted regions were matched to the 5-class ground-truth data of physician likelihood classifications. Results are shown for the (a) Mean Probability ($U_{MP}$), (b) KL Divergence ($U_{KLD}$), and (c) Local Gradients ($U_{LG}$) UQ measures.*

*Table 13 Separation between physician-rated clases of disease likelihood for each pair of classes across malignant (left) and benign (right) groups. Results are shown across the three tested uncertainty measures, $U_{MP}$, $U_{KLD}$, and $U_{LG}$. percent differnces are shown as absolute values of median percent differences.*

| | Malignant Classes | | | Benign Classes | | |
|---|---|---|---|---|---|---|
| | 1-2 | 1-3 | 2-3 | 3-4 | 3-5 | 4-5 |
| | | | *$U_{MP}$ (Mean Probability)* | | | |
| AUC ↑ | 0.62 | 0.67 | 0.56 | 0.55 | **0.72** | **0.68** |
| Percent Difference ↑ | 0.009 | 0.01 | 0.003 | 0.002 | 0.02 | 0.02 |
| P-Value ↓ | **<0.001** | **<0.001** | 0.27 | 0.33 | **<0.001** | **<0.001** |
| | | | *$U_{KLD}$ (KL Divergence)* | | | |
| AUC ↑ | 0.62 | 0.68 | 0.56 | **0.56** | **0.72** | **0.68** |
| Percent Difference ↑ | 0.02 | 0.02 | 0.004 | 0.004 | 0.03 | 0.02 |
| P-Value ↓ | **<0.001** | **<0.001** | 0.26 | **0.32** | **<0.001** | **<0.001** |
| | | | *$U_{LG}$ (Local Gradients UQ)* | | | |
| AUC ↑ | **0.67** | **0.79** | **0.62** | 0.54 | 0.59 | 0.56 |
| Percent Difference ↑ | **0.27** | **0.50** | **0.22** | **0.05** | **0.15** | **0.10** |

| | | | | | | |
|---|---|---|---|---|---|---|
| P-Value ↓ | **<0.001** | **<0.001** | **<0.05** | 0.53 | 0.11 | 0.16 |

## 4.4 Discussion

In this work, we introduced a novel, gradient-based method to quantify the uncertainty of regional outputs for deep learning-based medical image assessments. We demonstrated the utility of this UQ method using the application of metastatic disease delineation. We used four uncertainty validation assessments: (1) response to artificially degraded image quality, (2) comparison between matched high- and low-quality clinical images, (3) false positive (FP) filtering, and (4) correspondence with physician-rated malignant disease likelihood. In all experiments, comparisons were drawn between the novel UQ measure ($U_{LG}$), and measures based on the mean probability of a predicted region ($U_{MP}$) and the KL divergence of a predicted region without the use of gradients ($U_{KLD}$). Overall, our results indicate that leveraging the gradient space of a trained model is useful for UQ. This agrees with the work of (Huang et al., 2021), which leveraged gradients for out-of-distribution detection (OOD) for 1-dimensional natural image classification models. In their work, however, a small gradient score was associated with high uncertainty (or high OOD likelihood), opposite to our formulation. This difference may be due to our localization of regional outputs for gradient computation, which may more appropriately model sensitivities of specific regional outputs to learned model parameters (equation 13).

In the image degradation experiment, all three uncertainty measures increased with increasing image degradation magnitude across degradation types. Thus, all three measures followed the expected behavior and displayed some level of reliability. When compared to the non-gradient measures, the $U_{LG}$ measure exhibited a stronger response to image degradation, indicating its potentially higher sensitivity to abnormal image information. This may be advantageous in a clinical setting where a trained model

may encounter a wide array of abnormal image information such as artifacts which would merit human intervention. At the same time, a highly sensitive uncertainty measure may excessively flag regions as abnormal, potentially slowing down the image analysis process. This potential advantage of the $U_{LG}$ measure should be considered within the context of the image analysis task. The work of (González et al., 2022) similarly demonstrated the utility of an uncertainty measure using artificial image degradations, however, this study specifically targeted OOD uncertainty (a form of epistemic uncertainty). In contrast, we designed a method to target uncertainty more generally. While the image degradation experiments are indicative of the $U_{LG}$ measure's utility in capturing OOD uncertainty, more work should be dedicated specifically to evaluating this type of uncertainty using the Local Gradients UQ method.

The $U_{LG}$ measure achieved only slight statistical significance in distinguishing between matched lesions in low-quality (AC-CT) and high-quality (CE-CT) CT images ($p < 0.05$). Meanwhile, the $U_{MP}$ and $U_{KLD}$ measures did not achieve statistical significance for this task ($p > 0.05$). One explanation for why this separation was not stronger is that it may not be consistently true that liver tumors are more detectable on CE-CT than on AC-CT from a PET/CT scan. For instance, in Figure 31 we showed two examples of matched CE-CT and AC-CT liver tumors with overlaid $U_{LG}$ measures. In the second example (Figure 31b), our hypothesis that predicted tumor delineations in the lower quality (AC-CT) dataset will yield higher uncertainties was challenged. Disease presentation factors such as differences in tumor differentiation between scan types may explain this discrepancy. Thus, comparing uncertainty values between these high- and low-quality images from a clinical setting may be limited.

The $U_{LG}$ measure outperformed the non-gradient measures in the FP filtering assessment. The benefit of the $U_{LG}$ measure over the reference, non-gradient measures was perhaps the most evident in this assessment compared to other assessments. These results support the assumption that incorrectly

predicted regions contain high levels of uncertainty. This indicates the $U_{LG}$ measure's utility in removing erroneously predicted regions. The work of (Nair et al., 2020) similarly showed that FP regions contain higher levels of uncertainty than TP regions when using MC dropout-based uncertainty measures. A potential drawback of these measures, however, is their voxel-wise uncertainty output which is aggregated into a region-wise measure for detection assessments. In this process, a region may falsely be designated as a FP due to a small portion of its regions containing high voxel-wise uncertainty (e.g., a non-obvious boundary). The $U_{LG}$ measure, which directly generates a region-wise uncertainty output, mitigates this limitation. The sensitivity study performed within the FP filtering assessment showed that leveraging gradient information from the last two convolutional blocks of the model (decoder blocks 4-0 and 4-1) yielded sufficiently effective performance, indicating the selection of layers from which to acquire gradient information is important. These results are consistent with (Huang et al., 2021a) which implemented a similar methodology for UQ within the natural imaging domain and found that gradient information closer to the model's output yielded superior UQ. Using gradient information close to the model output lends the added advantage of faster gradient computation times. FP filtering results were comparable for both gradient-based uncertainty measures ($U_{MP,Gradients}$ and $U_{LG}$), indicating the inclusion of gradient information from different regional target functions lends useful uncertainty quantification.

Lastly, the $U_{LG}$ measure showed superior correspondence with physician-rated disease likelihood score among malignant disease classes. However, the correspondence across benign disease classes was stronger for the non-gradient based measures. This discrepancy may be due to the imbalance in the number of lesions in the malignant and benign classes (Table 11), where the model saw more malignant lesions during training, making its output more stable for this disease type. Assessments involving the association between deep learning-based uncertainty and reader-based uncertainty are not common,

likely due to the burdensome challenge of acquiring reader-based uncertainty. In the work of (Klanecek et al., 2023), an uncertainty measure was similarly related to qualitative reader assessments, however, these assessments consisted of the "acceptability" of deep learning-based delineations. Both of these approaches are subject to reader errors and biases. This is especially true in our assessment where the distinction between malignant and benign disease may be very subtle and where classes were defined by a single reader. A previously performed inter-physician reader study on a subset of 14 patients in *Dataset 2 – Bone Metastases* showed only moderate agreement in lesion classification across four physicians (Perk, et al., 2018). Lesion classifications from the single physician used in this work were correlated with the consensus of the multi-physicians in this subset of the data. Thus, we can expect a moderate amount of noise within the lesion classifications from the single physician used for training and assessment which may explain the discrepancy we observed in $U_{LG}$ measure performance between malignant and benign disease classes. Performing this assessment using the concordance of the multiple readers would be more stable, however, acquiring the multi-reader classifications for the whole dataset was not possible.

The main challenge of this work is the assessment and validation of uncertainty measures. It is common practice to implement validation assessments where model-uncertainty is approached in the same manner as reader-uncertainty. Thus, we assume there to be a strong correlation between model- and reader-uncertainty. While this assumption lends interpretability, it may not entirely hold. More work should be dedicated to understanding the nuances of model-uncertainty and to developing standardized assessments. While the assessments in our work also base model-uncertainty on reader-uncertainty, we included several assessments in this work to test the robustness of the developed $U_{LG}$ measure. Another challenge of this work is the unbounded nature of the Local Gradients UQ method. To account for this and to enhance the interpretability across uncertainty measures, we normalized the Local Gradients UQ

measures to the 95$^{th}$ percentile of a set of scores of predicted regions with assumed low uncertainty. This normalization should be performed separately for each trained model, where it is required to identify predicted regions with assumed low uncertainty. For example, assumed low uncertainty predictions may come from regions with multi-reader confirmation or from regions with high conspicuity. However, doing so may not be straightforward or possible in certain datasets. Additionally, given the constraint of designing a post hoc method that does not change a trained model's output, our method only acts on the model's true and false positive output and does not allow for the recovery of false negative predicted regions. Meanwhile, other methods such as MC dropout have the potential to recover false negative regions if these regions contain high prediction uncertainty. Thus, the Local Gradients UQ method may not be optimal for certain applications. Lastly, the Local Gradients UQ method calculates gradients via backpropagation for each predicted region. As a result, the computation time is not consistent across test images and scales with the number of predicted regions within an image. Additional memory may also be required for this method due to running backpropagation upon inference.

The Local Gradients UQ method could readily be applied to other deep learning medical image tasks, especially those with ways to localize model outputs. For instance, the Local Gradients UQ method could be useful for UQ in other structure delineation or detection tasks, especially for challenging structures such as the duodenum (Gibson et al., 2018). The Local Gradients UQ method could also be applied to deep learning-based image registration (e.g., Dalca et al., 2019), where the propagated structures from the learned displacement fields can be used to localize model outputs over which to quantify uncertainty. As a final example, the Local Gradients UQ method could be useful for deep learning-based MRI-to-CT image synthesis tasks (e.g., T. Wang et al., 2021), where UQ on structures that are difficult to synthesize can be acquired given pre-existing contours for model output localization. It is

most natural to apply the Local Gradients UQ method to deep learning tasks with spatially representative outputs such as in the examples described above. For computer vision tasks with 1D outputs, such as image classification, other methods, such as the one proposed by (Huang et al., 2021a), might be more appropriate.

Careful protocols should be established for how any uncertainty measure is to be used in deployed clinical settings. For instance, some level of human monitoring and intervention would likely be necessarily embedded into automated region delineation-based workflows. Here, displaying an uncertainty measure with each predicted region would be helpful so that the user can more quickly review predictions that derive from abnormal image information, remove false positive regions, and review potentially misclassified regions. Different intervention levels can be established by monitoring uncertainty levels on a priori assumed low uncertainty regions, ideally on a validation dataset. Lastly, different intervention approaches should be established for each image analysis task given the clinical use of each predictive model. More work should be dedicated to further understand how uncertainty measures can be appropriately leveraged in deployed clinical settings.

## 4.5 Conclusion

In summary, this work introduced the Local Gradients uncertainty quantification method. We found that the localized gradient information, inherent to this method, enhanced region-based uncertainty quantification across four validation assessments. These results indicate that users do not have to deviate far from their model to gather uncertainty information. The model's own gradient space can effectively be leveraged for uncertainty quantification.

# 5 General Discussion

## 5.1 Summary

In this thesis, we introduced UQ for deep learning-based medical image analysis, emphasized its importance in ensuring the safety of clinically deployed models, and presented several quantitative investigations that provided new critical insights into various methods of UQ for medical image analysis applications. This thesis was motivated by the hypothesis that **UQ methods for medical image analysis applications will enhance the accuracy, reliability, and utility of deep learning model outputs intended to inform clinical decision-making**. Using metastatic lesion segmentation as the primary application, this hypothesis was explored through the primary goal of this thesis: **to investigate UQ methods for the metastatic lesion segmentation task, address their limitations, and develop novel UQ approaches** to overcome these limitations and **improve UQ utility**. In pursuing this goal and investigating the associated aims, this thesis introduced several specific and novel innovations supporting the clinical need for comprehensive UQ.

**Specific Aim 1** of this thesis investigated OOD UQ. To this end, work dedicated to **SA1a** (presented in Chapter 2) explored several established embedded feature-based OOD detection methods, which were implemented and quantitatively evaluated under the application of metastatic lesion segmentation on abdominal CT images. Importantly, a dedicated study of embedded feature-based OOD methods for a medical image analysis task has not previously been performed. Our findings suggested that **none of the established embedded feature-based OOD measures were sufficient to detect simulated OOD medical images**.

Under the same assessment conditions, a novel approach to OOD detection based on information theory, termed InfoOOD, was investigated and quantitatively compared to the established methods in **SA1b** (presented in Chapter 2). This work was the first to demonstrate the utility of a post hoc information bottleneck optimization process for OOD detection. In addition to its unexplored novelty, this approach offers greater statistical interpretability than established methods. To assess the performance of the InfoOOD measure, we uniquely defined OOD test data using simulated artifacts on CT images, overcoming a distinct challenge for clinically relevant OOD method evaluations. Prior work utilized much less clinically relevant OOD data types, such as different imaging modalities, anatomical locations, applied spatial shifts, or image corruptions. **The introduced simulated artifact approach in this work much more closely resembled expected clinical OOD data, offered a better understanding of the sensitivities of different OOD approaches, and holds potential for use as a critical benchmark evaluation for clinical OOD measures**.

Based on this realistic quantitative evaluation approach, the introduced InfoOOD measure outperformed established embedded feature-based OOD measures. Not only did it achieve higher OOD detection sensitivity (e.g., $AUC = 0.90$ vs. $AUC = 0.49$ for detecting images with simulated strong magnitude low dose acquisition), but it also yielded better correlations between the OOD measure and segmentation model performance (e.g., $\rho = -0.52$ vs. $\rho = -0.06$ for the correlation between OOD measure and predicted lesion sensitivity). A disadvantage of the proposed InfoOOD measure is the additional inference-time computational constraint compared to established embedded feature-based methods. However, this time demand was marginal and would only be a concern in real-time image analysis scenarios. Consequently, **the novel information bottleneck approach for OOD detection offers an alternative, superiorly sensitive, and more interpretable OOD measure and appears promising to be used clinically to detect OOD data and to provide users with a measure of model prediction reliability.**

**Specific Aim 2** of this thesis investigated ID UQ. In **SA2a** , established methods were investigated for a particularly critical and challenging clinical task (presented in Chapter 3). Despite the expanding body of research dedicated to UQ for medical imaging applications (see Section 1.4), no previous studies have investigated UQ for the whole body, metastatic lesion segmentation task. Doing so is critical because image analytics pipelines are being proposed for a variety of clinical tasks built from segmentation model outputs (e.g., see Section 5.3.1). The uncertainty from the segmentation stage in these pipelines will inevitably propagate to these downstream clinical tasks, potentially inducing clinical errors. Lesions on whole body images are also challenging to segment due to factors such as low contrast and variable radiotracer uptake patterns in the case of PET/CT imaging.

We implemented and investigated four established UQ methods for this task: model probability entropy, Monte Carlo dropout, deep ensembles, and test-time augmentation. Across four quantitative evaluations, the **test-time augmentation** method was **quantified as the superior UQ method for the challenging clinical task of whole body malignant metastases segmentation imaged on PET/CT**. While it is difficult to claim a given UQ method explicitly targets one source of uncertainty, each UQ method tends to weight towards a specific source. In this case, test-time augmentation tends towards capturing aleatoric uncertainty. This suggests the importance of curating more accurately delineated lesions in ground-truth training data. Additional work should be dedicated to exploring advanced data augmentation strategies, such as additive image noise, to further enhance the UQ performance of the test-time augmentation method for metastatic lesion segmentation and other clinical tasks. Overall, **our comprehensive and quantitative results favoring the superiority of the test-time augmentation method are critical to establishing clinical UQ for this task and for those in the research community who seek to incorporate metastatic lesion segmentation uncertainty into their work.**

A novel ID UQ approach, termed *Local Gradients*, was investigated and evaluated in pursuit of **SA2b** (presented in Chapter 4). This method aimed to address the limitations of established ID UQ methods (see Section 1.4.2). Namely, **the Local Gradients method was designed to be post hoc and to not change a trained model's output space, meaning it could readily be applied to deployed models without the need for model re-training or re-evaluation**. This criterion is essential for promoting the integration of UQ methods into clinical settings, as it allows uncertainty information to be added to previously validated and deployed models. The method assessed model output sensitivities, localized to predicted regions, to trained model parameters via gradient information computation and aggregation. Since the Local Gradients UQ method requires a single forward pass through a single, deterministic model, it is most appropriately categorized as a single-network, deterministic UQ method (see Section 1.4.2). However, a single backward pass is also required to acquire model gradient information, making this method slightly more computationally involved than other single-network, deterministic methods.

The quality of the Local Gradients UQ measure was quantitatively assessed across four evaluations and compared to output-based UQ measures, which adhered to the same constraints (i.e., post-hoc and do not alter a trained model's output space). Across these evaluations, the Local Gradients UQ measure achieved better performance for detecting perturbed images (e.g., a median uncertainty value percent difference between the most degraded and non-degraded image of 62.4% vs. 3.45), for detecting low-quality clinical images (e.g., $p < 0.05$ vs $p = 0.67$), for detecting false positive predicted segmented regions (e.g., $AUC = 0.87$ vs. $AUC = 0.72$), and for correspondence with physician-rate likelihood of malignant disease classes (e.g., $AUC = 0.79$ vs. $AUC = 0.68$ for distinguishing between definitely malignant and equivocal lesion prediction). A possible disadvantage of the Local Gradients measure is the additional computation demand required for gradient calculation. However, this computational time constraint would likely only be prohibitive for real-time image analysis tasks. In

addition, the computational demands of the Local Gradients measure are significantly lower than those of previously established methods (summarized in Section 3.3.1). **The Local Gradients method fundamentally offers a new ID UQ method to the field, demonstrating novel deep learning model capabilities by uniquely leveraging test-time gradient information to characterize prediction uncertainty.**

Several aspects of this work support the main hypothesis which formed the foundation of this thesis. For example, it was shown that the investigated uncertainty measures can be used to detect false positive predicted regions in Chapters 3 and 4. This reduction in false positive regions inherently **enhanced the accuracy of the deep learning models**. The **enhancement of model reliability** using UQ methods was also demonstrated throughout this work. For instance, all investigated uncertainty measures were shown to detect poor-quality image data, including data with simulated artifacts (Chapter 2) or added image noise (Chapters 3 and 4). Strong correlations between uncertainty measures and model performance metrics, as reported in Chapters 2 and 3, further highlighted how UQ can enhance model reliability. At a high level, the implementation of each UQ method in this work demonstrated and **enhanced the utility of deep learning models**, where the models themselves could effectively be leveraged to acquire uncertainty information. Specific and novel enhancements to model utility were illustrated in Chapter 2, where an information theory-based OOD measure was obtained with only a slight modification to model architecture, and in Chapter 4, where it was shown how a model's gradient space—typically used only during training—was leveraged for UQ.

In a broader context, the technical innovations in this thesis **address the urgent demand for advanced uncertainty quantification methods to support the safe and reliable use of deep learning-based image analysis within clinical practice**. Each contribution **advances specific components of a broader framework for robust UQ** (Figure 34), where ID uncertainty must be quantified separately and

contingently upon successful OOD uncertainty assessment (see Section 1.4). Through systematic evaluations of both existing and novel UQ methods, this work provides critical insights into their utility and limitations, informing strategies for optimal clinical integration of UQ. Collectively, **these technical innovations form a foundation upon which a vast array of uncertainty-aware applications can be implemented and ultimately bring us closer to the widespread clinical implementation of UQ to ensure the safety of clinically deployed deep learning-based predictive models**. To get to this point, however, careful consideration must be made regarding the logistical steps of UQ implementation within clinical settings.

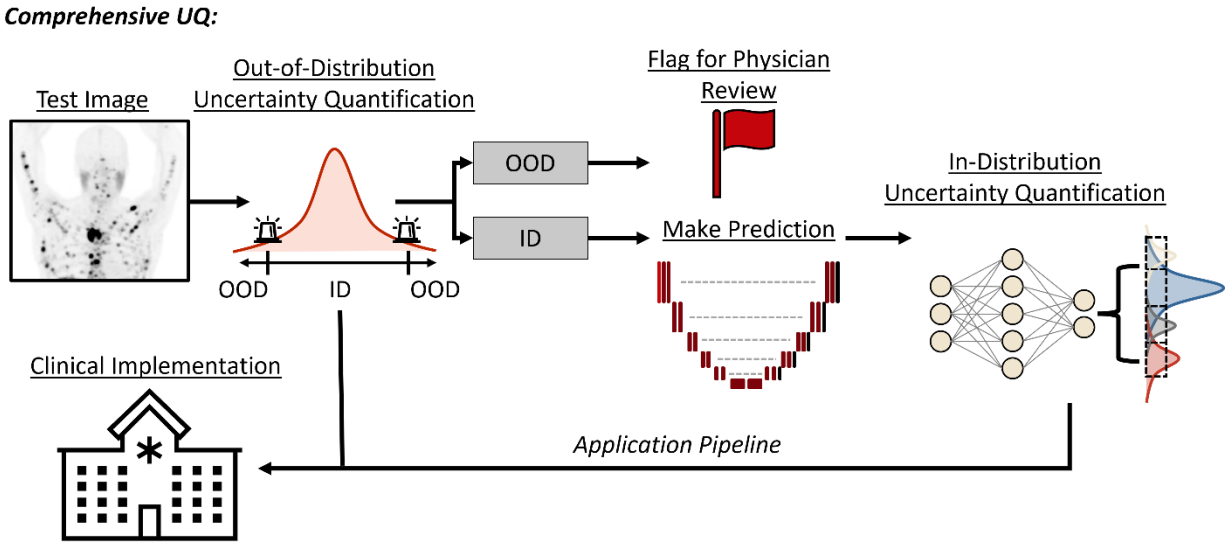



*Figure 34 Schematic describing a comprehensive approach to uncertainty quantification. In alignment with recent research, OOD and ID UQ are performed separately because single UQ methods do not sufficiently capture both types of uncertainty. Model prediction and ID UQ are invoked based on successful OOD uncertainty assessment. Images with high OOD uncertainty are abstained from model prediction and flagged for physician review. The OOD and ID uncertainty associated with an image and model prediction can also be used in a variety of applications and ultimately, used to gauge the trustworthiness of model predictions within clinical settings.*

## 5.2 Future Work – Clinical Implementation

The intent of deep learning-based medical image analysis research is to deploy predictive models into clinical settings to aid patient care. Doing so with embedded UQ will be critical to maximize model utility and minimize potential risks of patient harm. The U.S. Food and Drug Administration already requires medical devices that rely upon first principles-based (e.g., mechanistic) computational models to have robust UQ mechanisms (Center for Devices and Radiological Health, 2023). UQ requirements for data-driven computational models are likely imminent. Much of the technical groundwork in UQ measure development has been established in the literature (see Section 1.4), with added value from the original work described in this thesis. Thus, now is an opportune time to begin considering and defining the logistics regarding the use of uncertainty measures within clinical settings. Three key areas to address include the commissioning and continual monitoring of UQ measures, the human (i.e., clinician) interaction with UQ measures, and the potential knowledge gap in the medical physics field that must be filled.

### 5.2.1 UQ Commissioning and Continual Monitoring

Standards regarding the clinical deployment of predictive models have already been formulated. For example, the AAPM Task Group Report 273 outlines standards and best practices for the implementation of artificial intelligence (AI) for computer-aided diagnosis (Hadjiiski et al., 2023). While the report raises the importance of UQ, it does not provide any suggestions about how UQ may support clinical implementation. Figure 35 summarizes the key steps for clinical implementation from the Task Group report and offers areas where UQ can benefit each step. In the data collection step, UQ methods can be used to engineer robust and representative datasets via approaches such as active learning (further discussed in Section 5.3.3). During the model training and validation step, assessing the quality

of uncertainty measures in addition to model accuracy will be crucial. On-site model commissioning should involve further validation of the model accuracy and uncertainty measure reliability using locally acquired data. Finally, routine quality assurance (QA) will be required to ensure the ongoing quality of the predictive model and uncertainty measures to mitigate concerns such as data drift, where image quality changes due to routine scanner use (Merkow et al., 2024). While routine QA schedules should be established, intermittent QA may be performed based on UQ measure trends, where, for example, an increasing trend in overall uncertainty magnitude may merit interventional QA. Once such regulatory workflows are set, observational studies assessing clinician decision-making and patient outcomes after UQ implementation should be performed to understand the clinical impact.

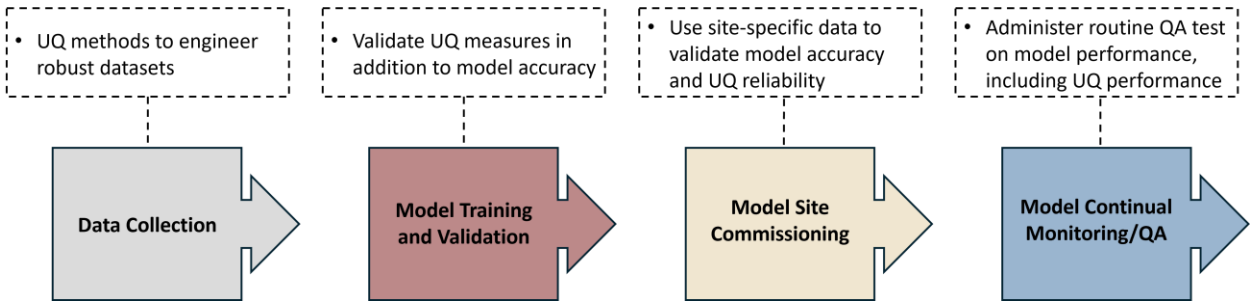



*Figure 35 Essential steps for predictive model clinical deployment with specific areas in which UQ can and should be included.*

Several innovations from this thesis work hold the potential to support the steps necessary for the commissioning and monitoring of predictive models and UQ measures within clinical settings (Figure 35). For example, the introduced UQ measures, especially the InfoOOD measure presented in Chapter 2, could readily be used in the *data collection* step to ensure representative data is included to ensure model robustness across broad feature distributions and expected aleatoric uncertainties (further discussed in Section 5.3.3). In the *model training and validation* and *model site commissioning* steps, the quantitative UQ assessments introduced throughout this thesis, such as the running model inference using images with simulated noise or artifacts, will aid in understanding model robustness and UQ

measure dependability. Similarly, these quantitative evaluations will be essential to implement for the *model continual monitoring/QA* step to track and document model and UQ measure performance. The UQ measures introduced and implemented in this work could further aid in *model continual monitoring* at an online level, for example, by using the InfoOOD measure to monitor the deviation of embedded features across time. This will likely capture model drift due to changes in image acquisition performance, indicating the need for manual interventions such as scanner calibration.

### 5.2.2 Human-UQ Measure Interaction

Additional work should also be dedicated towards understanding how to best represent UQ measures for optimal clinical use. Reporting raw uncertainty measures will be insufficient as these are most often unbounded and lack interpretability. Reporting normalized uncertainty measures, as proposed for the Local Gradients UQ measure reported in Chapter 4, may lend added interpretability. Simply rescaling UQ measures to a bounded range (e.g., [0, 1]) may also aid interpretability. However, care must be taken to ensure uncertainties are not mistaken for predicted risk levels. Reporting predictions along with error bars or the predictive distribution representing uncertainty will help mitigate this risk. Similar to what was done for reporting the predicted risk of unhealthy lymph node classification (Dihge et al., 2023), it may be useful to report an uncertainty within the distribution of reported uncertainties for a given patient cohort so the clinician can draw connections between past and current cases. Where appropriate, displaying interpretability maps along with uncertainty may further enhance clinician trust. These maps indicate important image regions relevant to the model prediction (Huff et al., 2021) and may be used to investigate high uncertainty outputs, offering clinicians insights into the model decision. Overall, more work needs to be done in human (i.e., clinician) computer

interaction to maximally leverage predictive models and associated uncertainty measures within clinical settings.

### 5.2.3 Medical Physics Knowledge Gap

Lastly, a major hurdle to the implementation of UQ in clinical settings is a potential knowledge gap in medical physicists' expertise (Figure 36). Medical physicists hold strong expertise in both image acquisition (e.g., CT physics) and therapeutic interventions (e.g., external beam radiation therapy). Consequently, in a future with routine clinical and image-based predictive modeling, it is clear that medical physicists should thoroughly understand the model inputs (e.g., image acquisition) and the repercussions of clinical decisions (e.g., therapeutic planning) made based on model outputs. However, knowledge related to modeling itself is also essential. It will be critical for medical physicists to be adequately trained on AI concepts to complete this critical chain of knowledge related to predictive model workflows. This will guarantee appropriate personnel are staffed to ensure the predictive models are implemented correctly and safely.

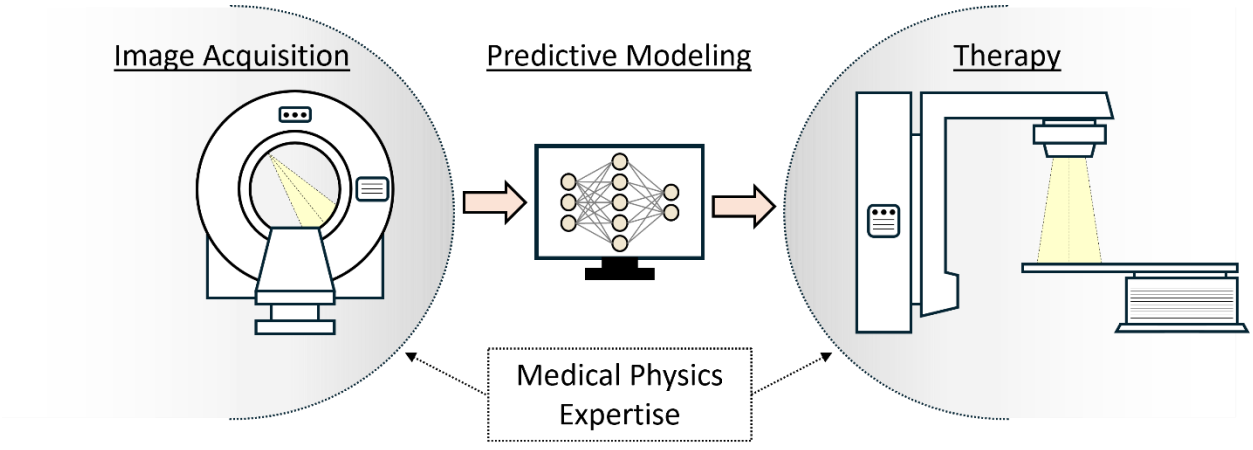



*Figure 36 Schematic demonstrating the potential predictive modeling knowledge gap within the medical physics field.*

Initiatives are being made to augment medical physics training with concepts related to AI. For example, several researchers have proposed the need to augment the current medical physics graduate-level curriculum with coursework in AI (Ng & Wong, 2022; Zanca et al., 2021), including different educational tracks based on prospective expertise in different clinical areas such as imaging modalities. Others are advocating to train currently practicing medical physicists in AI so that the field can more readily embrace an AI future (Wu et al., 2024). Across all petitions, there is a sense of importance and a push for urgency to not fall behind on the quickly advancing AI methods. It is also clear that medical physicists are best suited for such training and duties due to a strong mathematical foundation and strong familiarity with clinical practices and regulation policies. However, curriculum related to the UQ of model predictions is missing from these appeals. Including UQ concepts in medical physics AI training will be critical to fostering a foundational understanding of the behavior and limitations of clinically deployed deep learning models.

Hopefully, the work in this thesis will help address the current UQ knowledge gap surrounding UQ in the medical physics field. The introduction of predictive uncertainty, its sources, and its implications in Chapter 0 is intended to support greater familiarity with UQ for medical image applications. This work also emphasizes the importance of assessing UQ measures from several quantitative perspectives and the importance of evaluating a variety of UQ methods for specific tasks. Lastly, this work expands the understanding of deep learning model utility, for example, by the unique leveraging of model components, such as the localized model gradient space method presented in Chapter 4. Not only will this advance the exploration of novel UQ methods, but it will also promote the greater familiarity and understanding of the often-overlooked inner mechanisms and components of deep learning models, which will be increasingly important for medical physicists and other clinical professionals tasked with the governance over these models in clinical settings.

## 5.3 Future Work – Technical Developments

While UQ is already well-positioned to begin formulating its clinical implementation, ongoing technical advancements are essential to deepen our understanding of predictive uncertainty and to expand the clinical utility of UQ methods. Several of these developments can be built directly upon the work presented in this thesis. Novel work will be necessary to keep pace with the rapid onset of new and evolving AI technologies. Nevertheless, the findings in this thesis provide a foundation for applying UQ methods to emerging tools and models. In the near future, key research directions include the propagation of UQ measures into downstream clinical tasks, acquiring voxel-wise OOD measures, the integration of UQ methods into Active Learning frameworks, greater theoretical understandings of uncertainty sources within medical settings, and UQ for next-generation AI technologies.

### 5.3.1 Uncertainty Propagation into Downstream Clinical Tasks

The UQ innovations presented in this thesis were developed and evaluated using the metastatic lesion segmentation task. While important, the resulting segmentations are not necessarily clinically useful on their own. However, they enable a host of downstream clinical tasks designed to inform patient care. For example, our group recently developed a predictive model of treatment response for patients with metastatic neuroendocrine tumors undergoing systemic therapy, which was built utilizing segmentation-based extracted image features (Santoro-Fernandes et al., 2024b). Propagating uncertainty information from the deep learning-based segmentations to the final patient response prediction is a critical next step in introducing a measure of reliability for more robust and safe use of the predictive model output.

We have acquired preliminary results for uncertainty propagation into this previously developed predictive model (Santoro-Fernandes et al., 2024b) using the same dataset as in Section 3.2.1. This dataset consisted of patients receiving systemic therapy with pre- and post-therapy PET/CT imaging, along with long-term patient outcome data. In this predictive model study, deep learning-based lesion segmentations were used to extract lesion-wise biomarkers across multi-timepoint imaging, which were subsequently aggregated into patient-wise features. Feature selection was performed to determine the three predictive features (denoted as F1, F2, and F3 and described in Figure 37), which were used to train a multivariate linear regression model to predict patient-wise progression-free (PFS) survival. Additional information about the predictive model and features can be found in (Santoro-Fernandes et al., 2024b)

Uncertainty was propagated into this predictive model via sampling of the segmentation probability maps to acquire preliminary prediction uncertainty. Based on our findings in Chapter 3, the test-time augmentation UQ method was used in this preliminary work due to its superior UQ performance for the metastatic lesion segmentation task. For each patient, test-time augmentation was used to acquire ensembled probability outputs. Bernoulli sampling over each probability map was performed $N = 50$ times to acquire 50 segmentation masks per patient. For each mask, a set of the three predictive features was extracted, resulting in 50 unique feature sets to train 50 unique predictive models. Consequently, a predictive distribution was defined for each test patient using the 50 predictive models from which the mean and prediction variation were extracted to define the prediction and prediction uncertainty, respectively.

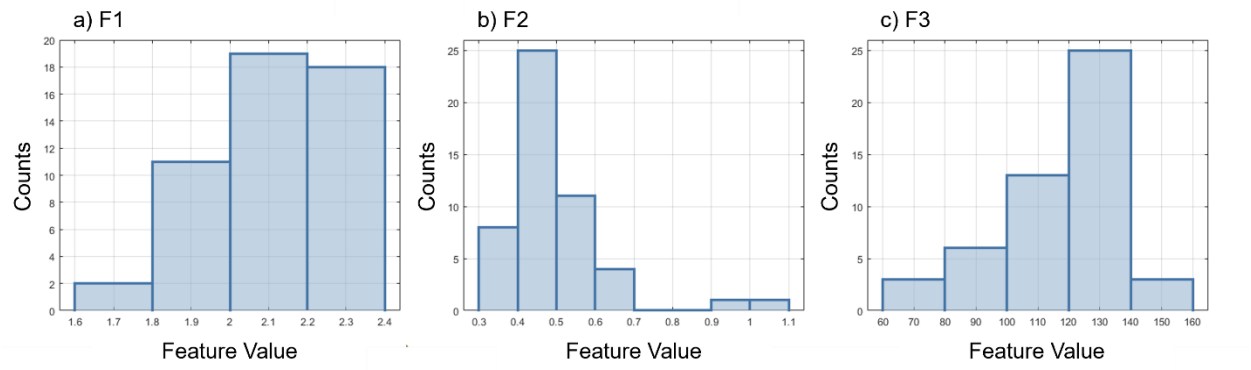



*Figure 37 Sampled predictive feature values from n=50 Bernoulli segmentation samples for one example test patient. F1 consisted of computing the average of each new lesion found on the post-therapy scans, normalized to the liver uptake, and the minimum of these lesion-wise features was used as the patient-wise feature. F2 consisted of computing the average of all lesions found on pre-therapy scans, normalized to the spleen update, and the minimum of these lesion-wise features was likewise used as the patient-wise feature. F3 consisted of computing the variance of each persisting lesion found on the post-therapy scans in the head region without uptake normalization, and the summation of these lesion-wise features was used as the patient-wise feature.*

An example of the feature distributions derived from the Bernoulli sampling for a single test patient is shown in Figure 37. Figure 38 shows the PFS prediction for each patient. The error bars represent the prediction 95% confidence intervals derived using the predictive output distribution (as opposed to using some statistical sampling technique such as bootstrapping). Thus, the error bars help clinicians determine how much to rely on model prediction for each patient. The average prediction uncertainty range was 7.3 months, the smallest range was 1.7 months, and the largest range was 43.8 months. For patients with high prediction uncertainty ranges, the clinician may decide to rely less on the predictive model outputs and more on traditional decision-making criteria.

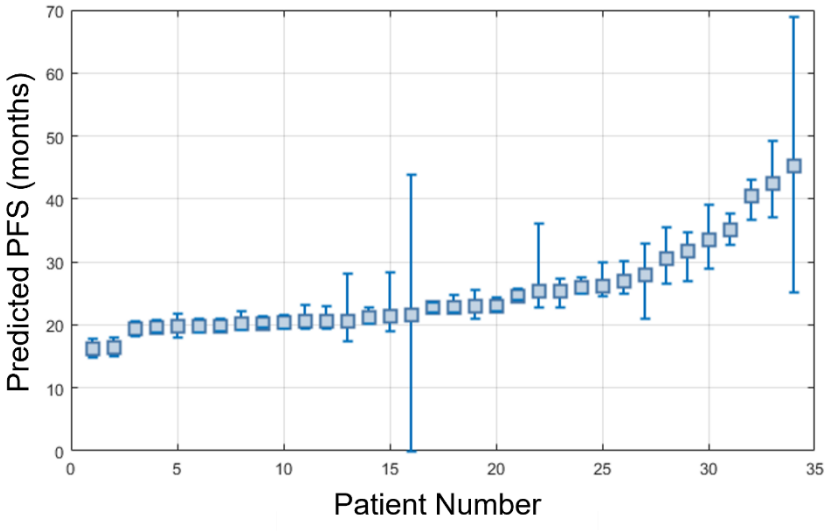



*Figure 38 Predicted progression free survival (PFS) for each patient. Error bars indicate 95% confidence intervals derived from the predictive model output distribution.*

These preliminary results indicate the feasibility and utility of propagating segmentation uncertainty into downstream clinical tasks, such as patient outcome prediction. Additional work is needed to further develop and evaluate the resulting prediction uncertainty. For example, building feature distributions from test-time augmentation samples, as opposed to the Bernoulli sampling approach, would more accurately reflect segmentation uncertainty. However, this would have required running the necessary lesion-matching algorithm (Santoro-Fernandes et al., 2024a) for each augmentation step, which was computationally prohibitive for this preliminary study. The uncertainty of this lesion-matching algorithm should also be quantified and propagated into the outcome prediction modeling. Finally, additional quantitative assessments, such as evaluating the associations of prediction uncertainty with prediction performance, should be performed.

### 5.3.2 Voxel-wise OOD Uncertainty Quantification

Most OOD detection algorithms quantify OOD uncertainty at the image level, like the proposed InfoOOD measure described in Chapter 2. However, targeting image-level OOD uncertainty may be incomplete for several reasons. For instance, while image-wise approaches effectively detect global shifts in image data, such as the presence of image artifacts or shifts in the image noise properties, they may be less sensitive at detecting more localized OOD regions within an image, such as obscure anatomy or localized image artifacts (e.g., internal motion). Consequently, image-level OOD UQ methods are unable to localize abnormalities within an image. Furthermore, image-wise OOD UQ methods lack the capacity to assess the OOD uncertainty of the voxel-wise predictions made by semantic segmentation models. Lastly, most ID UQ algorithms acquire voxel-wise uncertainty measures, like the ones implemented in Chapter 3. Image-wise OOD UQ impairs more advanced methods which may seek to combine OOD and ID uncertainty measures. Therefore, it may be critical to quantify the voxel-wise OOD uncertainty of a semantic segmentation model.

A proposed framework for acquiring voxel-wise OOD measures is shown in Figure 39. Based on insights gained from implementing OOD measures in Chapter 2, this framework relies on an embedded feature-based OOD approach. An input image is inferred upon by a trained, U-Net segmentation model. During the forward pass through the model, the feature maps that share the same spatial dimensionality as the input image are extracted. These features are fed to a 1-dimensional convolutional layer, followed by sigmoid activation and binarization to acquire predicted class labels. The features corresponding to the voxel in the image are extracted according to the prediction labels. If the original input image came from the model's train dataset, then the extracted features are used to update the class-wise feature distribution using Welford's online algorithm for distribution mean and variance estimation (Welford,

1962). If the original input image is a test image, then the voxel-wise distance between the test image features and the feature distribution defined by the train data is measured and serves as the voxel OOD measurement.

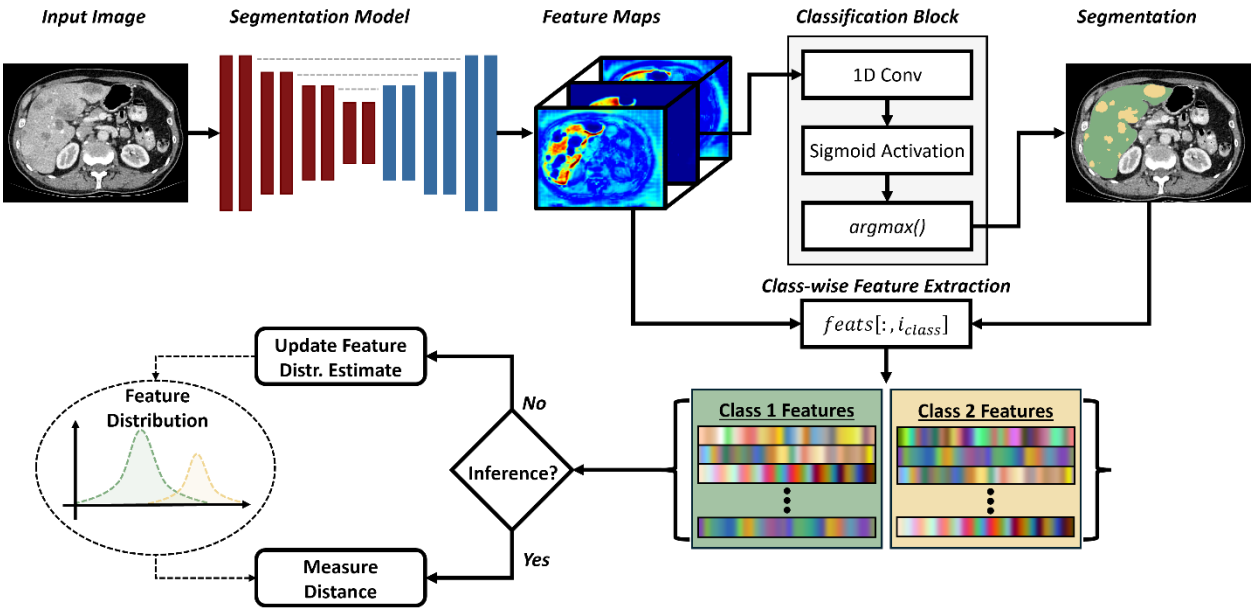



*Figure 39 Schematic describing the voxelOOD methodology. For each image, embedded features of channel depth 32 are extracted prior to the final 1D convolution, sigmoid activation, and binarization (via the argmax operation) steps. The predicted voxel-wise classes were used to extract class-wise features. If features were extracted from an image in the train data, the features are used to update the estimated training, class-wise feature distribution. If features were extracted from a test image, the distance between the test image features and the train feature distribution is computed as a voxel-wise distance metric.*

Preliminary experimentation was performed using this proposed voxel-wise OOD framework. A U-Net model was trained using the nnUNet repository (Isensee et al., 2021) to segment liver organ and metastatic lesions on abdominal CT images from the Liver Tumor Segmentation Benchmark (LiTS) dataset (Bilic et al., 2023). $N = 104$ and $N = 27$ images were used for model training and testing, respectively. Using the trained model, the voxel-wise and class-specific feature distribution was estimated using the train images. For each test image, voxel-wise OOD measures were acquired by

measuring the Mahalanobis distance (Mahalanobis, 1936) between its extracted features and the train feature distribution.

The performance of the proposed voxel-wise OOD measure was first assessed qualitatively in Figure 40. In the top row, an example is shown where regions in the liver associated with pronounced portal vein enhancement have high OOD measures. In the middle row, elevated OOD measures are associated with a false negative detected region. In the bottom row, abnormal regions with very high intensity, possibly due to liver calcification, similarly have large OOD values. These examples indicate that the voxel OOD measurement can localize regions with localized abnormalities or regions associated with prediction errors.

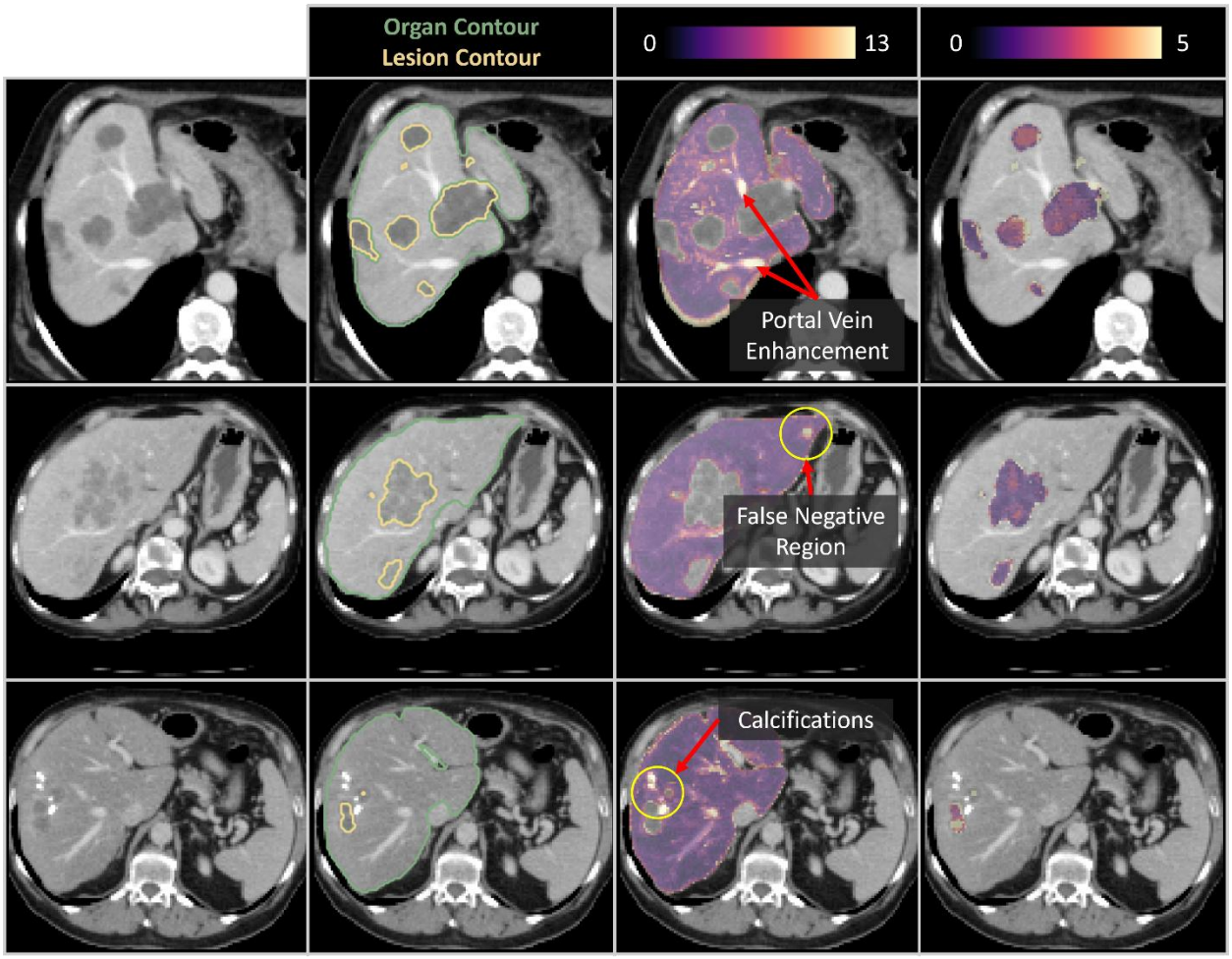



*Figure 40 Three qualitative examples of the voxelOOD results, including the abdominal CT slice (first column), the overlaid predicted segmentations (second column), the overlaid Class 1 voxelOOD results (third column), and the overlaid Class2 voxelOOD results (fourth column).* ***Top row:*** *An image that exhibited high Class 1 voxelOOD measurements related to portal vein contrast enhancement.* ***Middle row:*** *An image that exhibited high Class 1 voxelOOD measurements in a false negative predicted region.* ***Bottom row:*** *An image that exhibited high Class 1 voxelOOD measurements related to calcifications in the liver organ tissue.*

The voxel-wise OOD measure was then quantitatively evaluated using a coverage-based assessment. Specifically, the average test data Dice coefficient between the predicted and ground-truth segmentations was computed as a function of decreasing voxel-wise OOD thresholds, where each threshold corresponded to a percentile of the class-wise OOD measurements within each image. For example, at the 80$^{th}$ percentile, only voxels with OOD measurements below the 80$^{th}$ percentile were

included in the Dice coefficient calculation. As such, the Dice coefficient was expected to increase as the threshold decreased since voxels below lower thresholds correspond to lower OOD uncertainties.

Preliminary results for the Dice coefficient coverage evaluation are shown in Figure 41. The average Dice coefficient for each prediction class increased as the voxel OOD threshold decreased. The average liver organ Dice coefficient increased from 0.82 to 0.97 from the first to the final threshold. Similarly, the average liver lesion Dice coefficient increased from 0.70 to 0.88. These results indicate that the voxel-wise OOD measure corresponds with model segmentation performance, where higher voxel-wise OOD measures are associated with predicted segmentation errors and degrade overall performance.

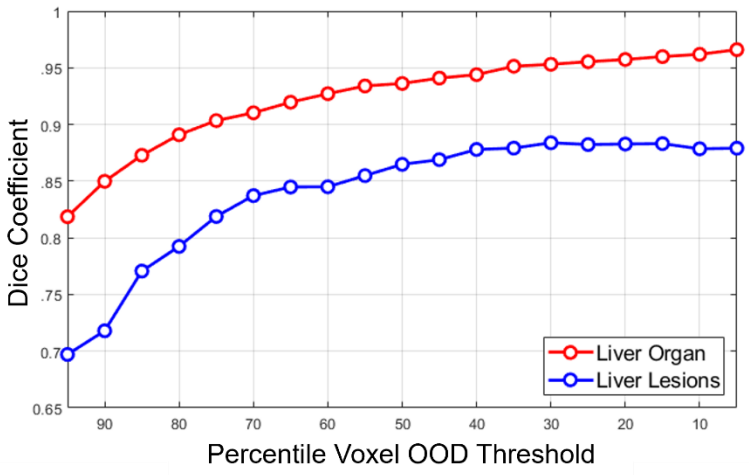



*Figure 41 Dice coefficient for predicted liver organ and lesions as a function or voxel OOD percentile threshold.*

These qualitative and quantitative preliminary results indicate utility in the voxel-wise approach for OOD UQ. Additional work is needed to fine-tune the approach and to further assess the measure's performance. For example, other distance measurements, such as the k-nearest neighbors algorithm (Sun et al., 2022), should be evaluated. Comparison to other output-based OOD measures (see Section

1.4.1), which are inherently voxel-wise, should be made. Finally, the potential of acquiring a single, comprehensive UQ measure that effectively combines voxel-wise OOD and ID measurements should be explored.

### 5.3.3 Active Learning for Metastatic Lesion Segmentation

A technical application of the uncertainty work developed in this thesis is active learning. In standard model training approaches, researchers use large, annotated datasets to train deep learning models. Large datasets are deemed necessary to train reliable models because it is assumed that maximizing the number of training samples maximizes the coverage of the train image feature space (i.e., all real images relevant to the deep learning task). Consequently, the model becomes more familiar with all types of data during training. However, large, annotated datasets are difficult to curate for medical image analysis tasks due to restrictive data-sharing policies and laborious expert-level image labeling. As a solution, active learning seeks to intelligently minimize training set sizes while maintaining good model performance (Budd et al., 2021).

An overview of the iterative active learning process is shown in Figure 42. An initial model is first trained on a small, annotated subset of the available data. The trained model is then deployed on an initially large pool of unlabeled data. The amount of model *informativeness* of each unlabeled image is then assessed, either via UQ and/or via some representativeness measure. When UQ is used, it is assumed that the unlabeled images with the highest uncertainty offer the most information gain to the model. When a representativeness measure is used, it is assumed that the unlabeled images with the most distinct image representation offer the most information gain to the model. A natural approach to measure representativeness is to use an OOD measure. Some works either only use uncertainty or representativeness to measure image informativeness, while others use some combination of the two

(Budd et al., 2021). The unannotated images with the highest informativeness are subsequently annotated and augmented to the training data. The augmented train set is then used for model re-training or continued training. This iterative process is continued until acceptable model performance is achieved.

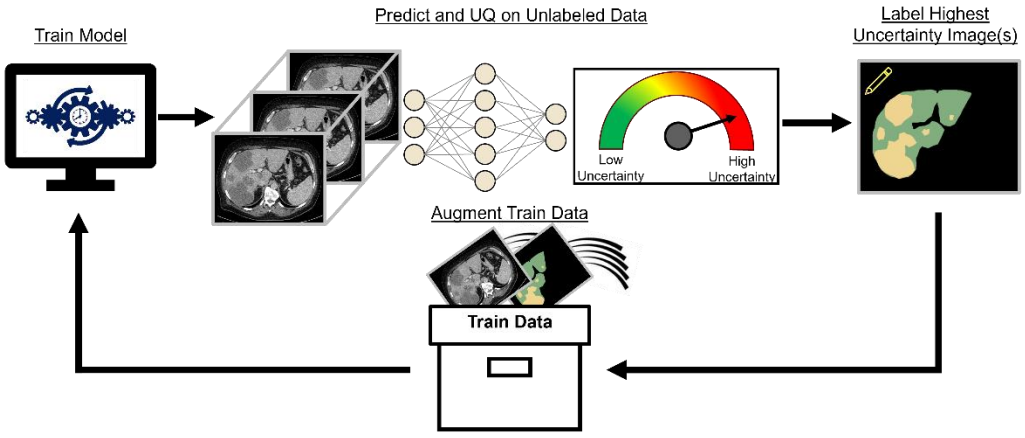



*Figure 42 Schematic of an active learning framework, where images are added to the training pool based on their level of informativeness to the training process. Here, we show uncertainty as an informativeness measure, where images with high uncertainty are selected such that, the model learns how to account for uncertainty during training appropriately.*

Active learning has been investigated across a variety of medical image tasks, including brain segmentation on MRI and cell segmentation on pathology images (Budd et al., 2021). This framework, however, has not been applied to the whole body metastatic lesion segmentation task. Doing so for this task is critical for two primary reasons. First, data labeling for this task is very time consuming and engineering an optimal and minimally sized dataset via active learning will save labeling time, allowing for more research on a wider variety of malignancies and image modalities. Second, active learning may enhance the accuracy and robustness of metastatic lesion segmentation models, ultimately positively enhancing patient care.

The UQ work developed in this thesis directly supports an active learning framework for metastatic lesion segmentation. The novel OOD detection approach outlined in Chapter 2 offers a highly sensitive measure of data representativeness to construct a heterogeneous yet representative training data distribution. The introduced InfoOOD measure in Chapter 2 would be preferred over alternative OOD measures because it can be used to directly quantify the amount of information gain a new image provides to the model (Schulz et al., 2020). Consequently, integrating the InfoOOD measure within an Active Learning framework will provide a greater interpretation of the added value of selected images to the training process and possibly result in more accurate and robust models. Furthermore, the results of Chapter 3 suggest that using test time augmentation is the optimal UQ to measure informativeness associated with uncertainty for this task. Work should be done to investigate the utility of OOD and UQ measures to acquire maximally informed datasets for robust and optimally sized metastatic lesion segmentation datasets.

### 5.3.4 Theoretical Understanding of Uncertainty Sources

Most UQ studies within the medical image domain consist of applying methods and insights formed in the natural image domain. While this has generated beneficial, validated, and relevant uncertainty measures, doing so risks introducing potential limitations that arise from translating the work into the medical domain. The computational medical image analysis field will benefit greatly by directly investigating key theoretical areas of UQ within the medical image domain.

A key idea to challenge is the notion that uncertainties sourced from OOD and ID data should be separately targeted. As described in Section 1.6, UQ methods that target ID uncertainty may fail to capture OOD uncertainty due to concerns over whether the learned decision boundaries, often co-aligned with ID uncertainty, extrapolate well to OOD test data. Thus, it is generally accepted that

comprehensive UQ involves independent OOD and ID UQ methods. However, this assumes that OOD data is very different from the train data (i.e., far OOD). While this is often the case for natural image tasks, where OOD data is defined using semantic shifts, it may not be true for medical image tasks, where OOD data is defined using more subtle covariate-shifted data.

Direct evidence for challenging the separation of OOD and ID UQ was presented in this thesis. In Chapter 4, for example, we proposed and validated a gradient-based method for ID UQ. The authors of (Huang et al., 2021a) similarly proposed leveraging model gradient information for UQ; however, this was aimed at capturing OOD uncertainty within the natural image domain. Furthermore, many of the UQ evaluations presented in this work involved assessing uncertainty as a function of OOD distance. This was accomplished by perturbing test data at increasing and iterative magnitudes by means of artifact simulation (Chapter 2) or inserting image noise (Chapters 3 and 4). Many of the UQ measures were able to capture and detect this perturbation process. Clearly, a single UQ method holds utility in both OOD and ID UQ. Nevertheless, the iterative perturbation approaches proposed in this thesis will play an important role in clarifying the impact of OOD and ID uncertainty on UQ measures. More work should be dedicated to understanding the extent to which ID UQ methods can be used for medical image tasks and at what point dedicated OOD UQ methods should be used.

A deeper understanding of the different uncertainty sources in medical settings is the need for uncertainty disentanglement, which seeks to separate the aleatoric and epistemic components of uncertainty. A variety of works exist for quantitative uncertainty disentanglement within the natural image domain (Mucsányi et al., 2024). However, such rigorous efforts have not been made in the medical image domain, where special modifications may need to be made. For example, since OOD data is often much closer to ID data in medical image tasks, the distributional aspect of epistemic uncertainty may be more entangled with its other components (e.g., poor model fitting). Thus, it may be beneficial

to disentangle uncertainty into three components for medical image tasks—aleatoric, epistemic (excluding distributional), and distributional. Doing so would greatly benefit deep learning-based medical image analysis. For example, such uncertainty disentanglement will enable the identification of specific uncertainty sources and consequently will allow researchers to take the necessary steps to reduce uncertainty (e.g., by means of correcting data noise, adjusting the model, or adding more data).

While disentangling uncertainty sources holds the potential to impact theoretical work greatly, it may hinder clinical implementation. Instead, a single, intuitive, and dependable uncertainty measure should be constructed to facilitate seamless clinical integration. The current separation of OOD and ID UQ measures introduces complexity in the clinical decisions associated with model predictions. For example, how should a prediction be interpreted with conflicting magnitudes for OOD and ID UQ measures? What if both UQ measures are moderately elevated? UQ would be much better adopted in the clinic if the total uncertainty was combined into a single measure. Normalizing each uncertainty measure by a separate calibration test set may aid this need for combination, as was done in Chapter 4 for the Local Gradients uncertainty measure. However, the optimal combination strategy is not straightforward because the relationship between ID and OOD uncertainty with model prediction error is not well-founded. Due to this unknown, data-driven approaches, such as a learnt weighting mechanism that maximizes correlation with prediction accuracy, should be explored.

### 5.3.5 Uncertainty Quantification for Next-Generation AI

AI technologies are rapidly advancing. The work in this thesis was developed using single data modality deep learning models trained using clinically acquired data for narrow-focused and deterministic tasks. For the next generation of AI technologies, models will rely heavily on much larger, synthetically derived data for training, accommodate different types of data modalities (e.g., images and

text), and be capable of a wide variety of prediction tasks (Acosta et al., 2022; Hoopes et al., 2024; Khosravi et al., 2024). These technological innovations hold great potential to revolutionize clinical care. As model capabilities scale, however, so do the associated risks. Thus, it is imperative for researchers to proactively design safety mechanisms, including UQ approaches, to mitigate large-scale risks of next-generation AI.

Medical image training datasets have historically been much smaller than natural image datasets due to concerns over patient privacy and lack of image-sharing infrastructure. Researchers have proposed addressing this problem using synthetically derived data from generative models. Consequently, the data available scales infinitely. As datasets scale in this way, model performance is expected to enhance. For example, augmenting a natural training dataset with synthetic images enhanced the performance of a chest x-ray classification model (Khosravi et al., 2024). Although untested, the uncertainty of models trained using synthetically augmented datasets may similarly become better (i.e., reduce), potentially removing the concern of high uncertainty outputs. However, this necessitates the synthetic data is high quality and representative of patient populations. UQ methods could be implemented on the data generation models to ensure poorly generated data does not negatively impact downstream model performance.

Models are also becoming increasingly multi-modal, enabling the simultaneous processing of different types of data, such as images and text. The utility of multi-modal models has been demonstrated in the medical image domain, for example, by generating radiology reports, enhancing predictive image models with laboratory data, and more (Cui et al., 2023). However, UQ methods have not been developed for medical image multi-modal models. An important area of research will be to ensure the concordance of uncertainty language descriptors (such as “likely”, “ambiguous”, “definitely”, etc.) and model prediction uncertainties. Moving beyond multi-modal models are foundational models

capable of being directed to perform any number of tasks given a text prompt. For example, a single radiology foundation model demonstrated incredible performance for target segmentation, image interpretation, image processing such as cropping about an anatomical area, and more (Hoopes et al., 2024). UQ can help ensure model outputs are reliable, as was done in this thesis work. In addition, UQ may be extended to different foundation model applications, such as indicating when a requested task lies outside of the model's ability.

However, the risks associated with state-of-the-art foundation models extend well beyond UQ. Seemingly human-like capabilities are being engineered into multi-modal models, such as logical reasoning (F. Xu et al., 2023), allowing models to behave in a way that resembles a level of consciousness. The boundary between deep learning as merely being software to it being something with personal agency is becoming blurred. The capabilities of these models extend well beyond what the developers are aware of (Wei et al., 2022). Concerningly some studies have documented unintended and unethical model behavior, such as intentional lying (Azaria & Mitchell, 2023). It will be essential to ensure the goals of these models are aligned with those of the human users, for example, by careful model alignment research (Y. Wang et al., 2023). As increasingly capable models are being introduced into medical setting (Keshavarz et al., 2024), it will be imperative to conduct similar studies within clinical contexts.

Even so, as these models become increasingly capable, potentially surpassing expert-level human intelligence, it may not be apparent exactly how a model might behave maliciously. It may not be possible to foresee all possible negative scenarios conceived by something that is assuredly smarter than us. There is a non-zero percent risk that highly intelligent and agentic models may behave in uncontrollable and destructive manners (Grace et al., 2024). The ramifications of such a scenario extend well beyond clinical settings. Proactive, ongoing, and rigorous principles for the use of AI within and

outside of clinical settings must be established and strictly adhered. Until then, the future remains entirely uncertain.

# 6 Conclusion

Uncertainty quantification is essential to ensure the safe implementation of clinical deep learning models. In this thesis, we presented important advances to uncertainty quantification for deep learning-based medical image analysis. We first implemented and assessed existing embedded feature-based approaches for out-of-distribution detection for the application task of metastatic lesion segmentation application. Using the same data, trained model, and assessment criteria, we next introduced a novel approach to out-of-distribution detection, which made use of information bottleneck theory to detect abnormal test images. Next, we assessed several UQ methods for the overlooked whole body metastatic lesion segmentation task. Lastly, we introduced a novel UQ method that leverages localized gradient information to assess model output sensitivities to trained model parameters, offering a low-cost and post hoc UQ option that preserves model performance accuracy. Each of these innovations holds great potential to positively impact patient care, whether as reliability measures to better inform clinicians on how to incorporate predictive models in their work or as a foundation for additional research, such as incorporating UQ measures into downstream image analytics tasks.

Deep learning technologies are rapidly evolving, and tools with profound capabilities are increasingly being introduced. As the capabilities of these technologies continue to expand, so do the associated risks. Research and development in AI safety, such as uncertainty quantification, must keep pace with the rate of deep learning innovations. This is especially true in the medical domain, where the integration of AI with patient care will surely be ubiquitous. Only when we introduce these technologies rightly, with appropriate risk mitigation such as uncertainty quantification, will their utility in bringing about a healthier future be fully realized.

# References

*Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008*

*Abdi, L., Valiuddin, M. M. A., Viviers, C. G. A., de With, P. H. N., & van der Sommen, F. (2025). Typicality Excels Likelihood for Unsupervised Out-of-Distribution Detection in Medical Imaging. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, 149–159. https://doi.org/10.1007/978-3-031-73158-7_14*

*Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. In Nature Medicine (Vol. 28, Issue 9, pp. 1773–1784). Nature Research. https://doi.org/10.1038/s41591-022-01981-2*

*Anthony, H., & Kamnitsas, K. (2023). On the Use of Mahalanobis Distance for Out-of-distribution Detection with Neural Networks for Medical Imaging. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, 136–146. https://doi.org/10.1007/978-3-031-44336-7_14*

*Araújo, T., Aresta, G., Schmidt-Erfurth, U., & Bogunović, H. (2023). Few-shot out-of-distribution detection for automated screening in retinal OCT images using deep learning. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-43018-9*

*Arega, T. W., Bricq, S., & Meriaudeau, F. (2025). Post-hoc out-of-distribution detection for cardiac MRI segmentation. Computerized Medical Imaging and Graphics, 119. https://doi.org/10.1016/j.compmedimag.2024.102476*

*Asgharnezhad, H., Shamsi, A., Alizadehsani, R., Khosravi, A., Nahavandi, S., Sani, Z. A., Srinivasan, D., & Islam, S. M. S. (2022). Objective evaluation of deep uncertainty predictions for COVID-19 detection. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-05052-x*

*Assouline, J., Cannella, R., Porrello, G., de Mestier, L., Burgio, M. D., Raynaud, L., Hentic, O., Cros, J., Tselikas, L., Ruszniewski, P., Vullierme, M. P., Vilgrain, V., Duran, R., & Ronot, M. (2023). Volumetric Enhancing Tumor Burden at CT to Predict Survival Outcomes in Patients with Neuroendocrine Liver Metastases after Intra-arterial Treatment. Radiology: Imaging Cancer, 5(1). https://doi.org/10.1148/rycan.220051*

*Azaria, A., & Mitchell, T. (2023, April 25). The Internal State of an LLM Knows When It's Lying. The Conference on Empirical Methods in Natural Language Processing. http://arxiv.org/abs/2304.13734*

*Balagopal, A., Nguyen, D., Morgan, H., Weng, Y., Dohopolski, M., Lin, M. H., Barkousaraie, A. S., Gonzalez, Y., Garant, A., Desai, N., Hannan, R., & Jiang, S. (2021). A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy. Medical Image Analysis, 72. https://doi.org/10.1016/j.media.2021.102101*

*Ballestar, L. M., & Vilaplana, V. (2021). MRI Brain Tumor Segmentation and Uncertainty Estimation Using 3D-UNet Architectures. Lecture Notes in Computer Science (Including Subseries Lecture Notes in*

*Artificial Intelligence and Lecture Notes in Bioinformatics), 12658 LNCS, 376–390. https://doi.org/10.1007/978-3-030-72084-1_34*

*Berger, C., Paschali, M., Glocker, B., & Kamnitsas, K. (2021). Confidence-Based Out-of-Distribution Detection: A Comparative Study and Analysis. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12959 LNCS, 122–132. https://doi.org/10.1007/978-3-030-87735-4_12*

*Bhat, I., Kuijf, H. J., Cheplygina, V., & Pluim, J. P. W. (2021). Using uncertainty estimation to reduce false positives in liver lesion detection. Proceedings - International Symposium on Biomedical Imaging, 2021-April, 663–667. https://doi.org/10.1109/ISBI48211.2021.9434119*

*Bhat, I., Pluim, J. P. W., Viergever, M. A., & Kuijf, H. J. (2022). Influence of uncertainty estimation techniques on false-positive reduction in liver lesion detection. Journal of Machine Learning for Biomedical Imaging, 1, 1–33. https://doi.org/https://doi.org/10.59275/j.melba.2022-5937*

*Bilic, P., Christ, P., Li, H. B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G. E. H., Chartrand, G., Lohöfer, F., Holch, J. W., Sommer, W., Hofmann, F., Hostettler, A., Lev-Cohain, N., Drozdzal, M., Amitai, M. M., Vivanti, R., … Menze, B. (2023). The Liver Tumor Segmentation Benchmark (LiTS). Medical Image Analysis, 84. https://doi.org/10.1016/j.media.2022.102680*

*Blanc-Durand, P., Jégou, S., Kanoun, S., Berriolo-Riedinger, A., Bodet-Milin, C., Kraeber-Bodéré, F., Carlier, T., Le Gouill, S., Casasnovas, R. O., Meignan, M., & Itti, E. (2021). Fully automatic segmentation of diffuse large B cell lymphoma lesions on 3D FDG-PET/CT for total metabolic tumour volume prediction using a convolutional neural network. European Journal of Nuclear Medicine and Molecular Imaging, 48(5), 1362–1370. https://doi.org/10.1007/s00259-020-05080-7*

*Boldrini, L., Bibault, J. E., Masciocchi, C., Shen, Y., & Bittner, M. I. (2019). Deep Learning: A Review for the Radiation Oncologist. In Frontiers in Oncology (Vol. 9). Frontiers Media S.A. https://doi.org/10.3389/fonc.2019.00977*

*Budd, S., Robinson, E. C., & Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. In Medical Image Analysis (Vol. 71). Elsevier B.V. https://doi.org/10.1016/j.media.2021.102062*

*Camarasa, R., Bos, D., Hendrikse, J., Nederkoorn, P., Kooi, E., van der Lugt, A., & de Bruijne, M. (2020). Quantitative Comparison of Monte-Carlo Dropout Uncertainty Measures for Multi-class Segmentation. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12443 LNCS, 32–41. https://doi.org/10.1007/978-3-030-60365-6_4*

*Carlsen, E. A., Lindholm, K., Hindsholm, A., Gæde, M., Ladefoged, C. N., Loft, M., Johnbeck, C. B., Langer, S. W., Oturai, P., Knigge, U., Kjaer, A., & Andersen, F. L. (2022). A convolutional neural network for total tumor segmentation in [64Cu]Cu-DOTATATE PET/CT of patients with neuroendocrine neoplasms. EJNMMI Research, 12(1). https://doi.org/10.1186/s13550-022-00901-2*

*Center for Devices and Radiological Health. (2023). Assessing the Credibility of Computational Modeling and Simulation in Medical Device Submissions - Guidance for Industry and Food and Drug Administration Staff. https://www.fda.gov/media/154985/download*

*Chaffer, C. L., & Weinberg, R. A. (2011). A Perspective on Cancer Cell Metastasis. https://www.science.org*

*Chan, T. F., & Vese, L. A. (2001). Active Contours Without Edges. In IEEE TRANSACTIONS ON IMAGE PROCESSING (Vol. 10, Issue 2).*

*Chen, X., Wang, X., Zhang, K., Fung, K.-M., Thai, T. C., Moore, K., Mannel, R. S., Liu, H., Zheng, B., & Qiu, Y. (2022). Recent advances and clinical applications of deep learning in medical image analysis. Medical Image Analysis, 79, 4. https://doi.org/10.1016/j.media.2022.1024*

*Chow, L. S., & Paramesran, R. (2016). Review of medical image quality assessment. In Biomedical Signal Processing and Control (Vol. 27, pp. 145–154). Elsevier Ltd. https://doi.org/10.1016/j.bspc.2016.02.006*

*Chu, M., Zinchenko, Y., Henderson, S. G., & Sharpe, M. B. (2005). Robust optimization for intensity modulated radiation therapy treatment planning under uncertainty. Physics in Medicine and Biology, 50(23), 5463–5477. https://doi.org/10.1088/0031-9155/50/23/003*

*Clark, A. M., Ma, B., Taylor, D. L., Griffith, L., & Wells, A. (2016). Liver metastases: Microenvironments and ex-vivo models. Experimental Biology and Medicine, 241(15), 1639–1652. https://doi.org/10.1177/1535370216658144*

*Cremers, D., Rousson, M., & Deriche, R. (2007). A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. International Journal of Computer Vision, 72(2), 195–215. https://doi.org/10.1007/s11263-006-8711-1*

*Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L. A., Wilson, K. T., Landman, B. A., & Huo, Y. (2023). Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. Progress in Biomedical Engineering, 5(2). https://doi.org/10.1088/2516-1091/acc2fe*

*Dalca, A. V., Balakrishnan, G., Guttag, J., & Sabuncu, M. R. (2019). Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Medical Image Analysis, 57, 226–236. https://doi.org/10.1016/j.media.2019.07.006*

*Delbeke, D., Martin, W. H., Sandler, M. P., Chapman, W. C., Wright, ; J Kelly, & Pinson, ; C Wright. (1998). Evaluation of Benign vs Malignant Hepatic Lesions With Positron Emission Tomography. Archives of Surgery, 133(5), 510–516.*

*Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. The Annals of Mathematical Statistics, 38(2), 325–339.*

*Denouden, T., Salay, R., Czarnecki, K., Abdelzad, V., Phan, B., & Vernekar, S. (2018). Improving Reconstruction Autoencoder Out-of-distribution Detection with Mahalanobis Distance. ArXiv, arXiv:1812.02765. http://arxiv.org/abs/1812.02765*

*DeVries, T., & Taylor, G. W. (2018). Leveraging Uncertainty Estimates for Predicting Segmentation Quality. ArXiv. http://arxiv.org/abs/1807.00502*

*Diao, Z., Jiang, H., & Shi, T. (2022). A unified uncertainty network for tumor segmentation using uncertainty cross entropy loss and prototype similarity. Knowledge-Based Systems, 246. https://doi.org/10.1016/j.knosys.2022.108739*

*Dice, L. R. (1945). MEASURES OF THE AMOUNT OF ECOLOGIC ASSOCIATION BETWEEN SPECIES. Ecology, 26(3), 297–302. https://doi.org/10.2307/1932409*

*Dihge, L., Bendahl, P. O., Skarping, I., Hjärtström, M., Ohlsson, M., & Rydén, L. (2023). The implementation of NILS: A web-based artificial neural network decision support tool for noninvasive*

*lymph node staging in breast cancer. Frontiers in Oncology, 13. https://doi.org/10.3389/fonc.2023.1102254*

*Dromain, C., Pavel, M. E., Ruszniewski, P., Langley, A., Massien, C., Baudin, E., Caplin, M. E., Raderer, A. M., Borbath, B. I., Ysebaert, D., Sedláčková, E., Vítek, P., Grønbæk, D. H., Adenis, F. A., Buscail, L., Cadiot, G., Dominguez, S., Ducreux, M., Lombard-Bohas, C., … Wolin, E. M. (2019). Tumor growth rate as a metric of progression, response, and prognosis in pancreatic and intestinal neuroendocrine tumors. BMC Cancer, 19(1). https://doi.org/10.1186/s12885-018-5257-x*

*Even-Sapir, E., Metser, U., Mishani, E., Lievshitz, G., Lerman, H., & Leibovitch, I. (2006). The Detection of Bone Metastases in Patients with High-Risk Prostate Cancer: 99m Tc-MDP Planar Bone Scintigraphy, Single-and Multi-Field-of-View SPECT, 18 F-Fluoride PET, and 18 F-Fluoride PET/CT. Journal of Nuclear Medicine, 47(2), 287–297.*

*Freitas, P. S., Janicas, C., Veiga, J., Matos, A. P., Heredia, V., & Ramalho, M. (2021). Imaging evaluation of the liver in oncology patients: A comparison of techniques. World Journal of Hepatology, 13(12), 1936–1955. https://www.wjgnet.com*

*Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., & Yang, X. (2020). Deep learning in medical image registration: A review. In Physics in Medicine and Biology (Vol. 65, Issue 20). IOP Publishing Ltd. https://doi.org/10.1088/1361-6560/ab843e*

*Fusai, G., & Davidson, B. R. (2003). Management of colorectal liver metastases. In Colorectal Disease (Vol. 5, Issue 1, pp. 2–23). https://doi.org/10.1046/j.1463-1318.2003.00410.x*

*Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning Zoubin Ghahramani. International Conference on Machine Learning. http://yarin.co.*

*Ghesu, F. C., Georgescu, B., Mansoor, A., Yoo, Y., Gibson, E., Vishwanath, R. S., Balachandran, A., Balter, J. M., Cao, Y., Singh, R., Digumarthy, S. R., Kalra, M. K., Grbic, S., & Comaniciu, D. (2021). Quantifying and leveraging predictive uncertainty for medical image assessment. Medical Image Analysis, 68. https://doi.org/10.1016/j.media.2020.101855*

*Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S. P., Clarkson, M. J., & Barratt, D. C. (2018). Automatic Multi-Organ Segmentation on Abdominal CT with Dense V-Networks. IEEE Transactions on Medical Imaging, 37(8), 1822–1834. https://doi.org/10.1109/TMI.2018.2806309*

*González, C., Gotkowski, K., Fuchs, M., Bucher, A., Dadras, A., Fischbach, R., Kaltenborn, I. J., & Mukhopadhyay, A. (2022). Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation. Medical Image Analysis, 82. https://doi.org/10.1016/j.media.2022.102596*

*Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems. http://www.github.com/goodfeli/adversarial*

*Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). Thousands of AI Authors on the Future of AI. ArXiv. http://arxiv.org/abs/2401.02843*

*Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, June 14). On Calibration of Modern Neural Networks. International Conference on Machine Learning. http://arxiv.org/abs/1706.04599*

*Hadjiiski, L., Cha, K., Chan, H. P., Drukker, K., Morra, L., Näppi, J. J., Sahiner, B., Yoshida, H., Chen, Q., Deserno, T. M., Greenspan, H., Huisman, H., Huo, Z., Mazurchuk, R., Petrick, N., Regge, D., Samala, R., Summers, R. M., Suzuki, K., … Armato, S. G. (2023). AAPM task group report 273: Recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging. Medical Physics, 50(2), e1–e24. https://doi.org/10.1002/mp.16188*

*Harmon, S. A., Perk, T., Lin, C., Eickhoff, J., Choyke, P. L., Dahut, W. L., Apolo, A. B., Humm, J. L., Larson, S. M., Morris, M. J., Liu, G., & Jeraj, R. (2017). Assessment of Early [ 18 F]Sodium Fluoride Positron Emission Tomography/Computed Tomography Response to Treatment in Men With Metastatic Prostate Cancer to Bone. J Clin Oncol, 35, 2829–2837. https://doi.org/10.1200/JCO*

*Hartrampf, P. E., Hüttmann, T., Seitz, A. K., Kübler, H., Serfling, S. E., Schlötelburg, W., Michalski, K., Rowe, S. P., Pomper, M. G., Buck, A. K., Eberlein, U., & Werner, R. A. (2023). SUVmean on baseline [18F]PSMA-1007 PET and clinical parameters are associated with survival in prostate cancer patients scheduled for [177Lu]Lu-PSMA I&T. European Journal of Nuclear Medicine and Molecular Imaging, 50(11), 3465–3474. https://doi.org/10.1007/s00259-023-06281-6*

*Hendrycks, D., & Gimpel, K. (2017, October 7). A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. International Conference on Learning Representations. http://arxiv.org/abs/1610.02136*

*Hoopes, A., Butoi, V. I., Guttag, J. V., & Dalca, A. V. (2024). VoxelPrompt: A Vision-Language Agent for Grounded Medical Image Analysis. ArXiv. http://arxiv.org/abs/2410.08397*

*Hu, C., Xia, T., Cui, Y., Zou, Q., Wang, Y., Xiao, W., Ju, S., & Li, X. (2024). Trustworthy multi-phase liver tumor segmentation via evidence-based uncertainty. Engineering Applications of Artificial Intelligence, 133. https://doi.org/10.1016/j.engappai.2024.108289*

*Huang, B., Yang, Q., Li, X., Wu, Y., Liu, Z., Pan, Z., Zhong, S., Song, S., & Zuo, C. (2024). Deep learning–based whole-body characterization of prostate cancer lesions on [68Ga]Ga-PSMA-11 PET/CT in patients with post-prostatectomy recurrence. European Journal of Nuclear Medicine and Molecular Imaging, 51(4), 1173–1184. https://doi.org/10.1007/s00259-023-06551-3*

*Huang, R., Geng, A., & Li, Y. (2021, October 1). On the Importance of Gradients for Detecting Distributional Shifts in the Wild. Neural Information Processing Systems. http://arxiv.org/abs/2110.00218*

*Huff, D. T., Weisman, A. J., & Jeraj, R. (2021). Interpretation and visualization techniques for deep learning models in medical imaging. In Physics in Medicine and Biology (Vol. 66, Issue 4). IOP Publishing Ltd. https://doi.org/10.1088/1361-6560/abcd17*

*Huynh, B. N., Groendahl, A. R., Tomic, O., Liland, K. H., Knudtsen, I. S., Hoebers, F., van Elmpt, W., Dale, E., Malinen, E., & Futsaether, C. M. (2024). Deep learning with uncertainty estimation for automatic tumor segmentation in PET/CT of head and neck cancers: impact of model complexity, image processing and augmentation. Biomedical Physics and Engineering Express, 10(5). https://doi.org/10.1088/2057-1976/ad6dcd*

*Iagaru, A., Mittra, E., Dick, D. W., & Gambhir, S. S. (2012). Prospective evaluation of 99mTc MDP scintigraphy, 18F NaF PET/CT, and 18F FDG PET/CT for detection of skeletal metastases. Molecular Imaging and Biology, 14(2), 252–259. https://doi.org/10.1007/s11307-011-0486-2*

*Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z*

*Jadon, S. (2020). A survey of loss functions for semantic segmentation. 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 1–7.*

*Jallow, N. (2014). Comprehensive assessment of uncertainties and errors in [ 18 F]NaF PET imaging. University of Wisconsin - Madison.*

*Jungo, A., Balsiger, F., & Reyes, M. (2020). Analyzing the Quality and Challenges of Uncertainty Estimations for Brain Tumor Segmentation. Frontiers in Neuroscience, 14. https://doi.org/10.3389/fnins.2020.00282*

*Kalinic, H. (2009). Atlas-based image segmentation: A Survey. Scientific Bibliography, 1–7.*

*Kang, S., Kang, Y., & Tan, S. (2024). Exploring and Exploiting Multi-modality Uncertainty for Tumor Segmentation on PET/CT. IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.1109/JBHI.2024.3397332*

*Karimi, D., & Gholipour, A. (2020). Improving Calibration and Out-of-Distribution Detection in Medical Image Segmentation with Convolutional Neural Networks. http://arxiv.org/abs/2004.06569*

*Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1). https://doi.org/10.1186/s12916-019-1426-2*

*Kendall, A., Badrinarayanan, V., & Cipolla, R. (2015). Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. ArXiv. http://arxiv.org/abs/1511.02680*

*Kendall, A., & Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? Advances in Neural Information Processing Systems.*

*Keshavarz, P., Bagherieh, S., Nabipoorashrafi, S. A., Chalian, H., Rahsepar, A. A., Kim, G. H. J., Hassani, C., Raman, S. S., & Bedayat, A. (2024). ChatGPT in radiology: A systematic review of performance, pitfalls, and future perspectives. In Diagnostic and Interventional Imaging (Vol. 105, Issues 7–8, pp. 251–265). Elsevier Masson s.r.l. https://doi.org/10.1016/j.diii.2024.04.003*

*Khosravi, B., Li, F., Dapamede, T., Rouzrokh, P., Gamble, C. U., Trivedi, H. M., Wyles, C. C., Sellergren, A. B., Purkayastha, S., Erickson, B. J., & Gichoya, J. W. (2024). Synthetically enhanced: unveiling synthetic data's potential in medical imaging research. EBioMedicine, 104. https://doi.org/10.1016/j.ebiom.2024.105174*

*Kirichenko, P., Izmailov, P., & Wilson, A. G. (2020). Why normalizing flows fail to detect out-of-distribution data. Advances in Neural Information Processing Systems, 2020-Decem, 1–27.*

*Klanecek, Z., Wagner, T., Wang, Y. K., Cockmartin, L., Marshall, N., Schott, B., Deatsch, A., Studen, A., Hertl, K., Jarm, K., Krajc, M., Vrhovec, M., Bosmans, H., & Jeraj, R. (2023). Uncertainty estimation for deep learning-based pectoral muscle segmentation via Monte Carlo dropout. Physics in Medicine and Biology, 68(11). https://doi.org/10.1088/1361-6560/acd221*

*Kolarik, M., Burget, R., & Riha, kamil. (2020). Comparing Normalization Methods for Limited Batch Size Segmentation Neural Networks. International Conference on Telecommunications and Signal Processing, 677–680.*

*Kumbhar, S. S., Baheti, A. D., Itani, M., & Nikam, R. (2021). Ambiguous Findings on Radiographs. In Current Problems in Diagnostic Radiology (Vol. 50, Issue 1, pp. 4–10). Mosby Inc. https://doi.org/10.1067/j.cpradiol.2019.10.003*

*Kushibar, K., Campello, V., Garrucho, L., Linardos, A., Radeva, P., & Lekadir, K. (2022). Layer Ensembles: A Single-Pass Uncertainty Estimation in Deep Learning for Segmentation. In L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, & S. Li (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2022 (Vol. 13438). Springer, Cham. https://doi.org/10.1007/978-3-031-16452-1*

*Kuyumcu, S., Özkan, Z. G., Sanli, Y., Yilmaz, E., Mudun, A., Adalet, I., & Unal, S. (2013). Physiological and tumoral uptake of 68Ga-DOTATATE: Standardized uptake values and challenges in interpretation. Annals of Nuclear Medicine, 27(6), 538–545. https://doi.org/10.1007/s12149-013-0718-4*

*Kyriakopoulos, C. E., Heath, E. I., Ferrari, A., Sperger, J. M., Singh, A., Perlman, S. B., Roth, A. R., Perk, T. G., Modelska, K., Porcari, A., Duggan, W., Lang, J. M., Jeraj, R., & Liu, G. (2020). Exploring Spatial-Temporal Changes in 18 F-Sodium Fluoride PET/CT and Circulating Tumor Cells in Metastatic Castration-Resistant Prostate Cancer Treated With Enzalutamide. Journal of Clinical Oncology, 38(31), 3662–3671. https://doi.org/10.1200/JCO.20*

*Lakshminarayanan, B., Pritzel, A., & Deepmind, C. B. (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. Advances in Neural Information Processing Systems.*

*Lambert, B., Forbes, F., Doyle, S., Dehaene, H., & Dojat, M. (2024). Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis. In Artificial Intelligence in Medicine (Vol. 150). Elsevier B.V. https://doi.org/10.1016/j.artmed.2024.102830*

*Lambert, B., Forbes, F., Doyle, S., Tucholka, A., & Dojat, M. (2022). Improving Uncertainty-based Out-of-Distribution Detection for Medical Image Segmentation. http://arxiv.org/abs/2211.05421*

*Lastrucci, A., Wandael, Y., Ricci, R., Maccioni, G., & Giansanti, D. (2024). The Integration of Deep Learning in Radiotherapy: Exploring Challenges, Opportunities, and Future Directions through an Umbrella Review. In Diagnostics (Vol. 14, Issue 9). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/diagnostics14090939*

*Ledda, E., Fumera, G., & Roli, F. (2023). Dropout injection at test time for post hoc uncertainty quantification in neural networks. Information Sciences, 645. https://doi.org/10.1016/j.ins.2023.119356*

*Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. Journal of Big Data, 5(1). https://doi.org/10.1186/s40537-018-0151-6*

*Leung, K. H., Rowe, S. P., Sadaghiani, M. S., Leal, J. P., Mena, E., Choyke, P. L., Du, Y., & Pomper, M. G. (2024). Deep Semisupervised Transfer Learning for Fully Automated Whole-Body Tumor Quantification and Prognosis of Cancer on PET/CT. Journal of Nuclear Medicine, 65(4), 643–650. https://doi.org/10.2967/jnumed.123.267048*

*Li, H., Jiang, H., Li, S., Wang, M., Wang, Z., Lu, G., Guo, J., & Wang, Y. (2020). DenseX-Net: An End-to-End Model for Lymphoma Segmentation in Whole-Body PET/CT Images. IEEE Access, 8, 8004–8018. https://doi.org/10.1109/ACCESS.2019.2963254*

*Li, H., Li, J., Guan, X., Liang, B., Lai, Y., & Luo, X. (2019). Research on Overfitting of Deep Learning. Proceedings - 2019 15th International Conference on Computational Intelligence and Security, CIS 2019, 78–81. https://doi.org/10.1109/CIS.2019.00025*

*Liang, S., Li, Y., & Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, 1–15.*

*Lin, C., Bradshaw, T., Perk, T., Harmon, S., Eickhoff, J., Jallow, N., Choyke, P. L., Dahut, W. L., Larson, S., Humm, J. L., Perlman, S., Apolo, A. B., Morris, M. J., Liu, G., & Jeraj, R. (2016). Repeatability of quantitative 18F-NaF PET: A multicenter study. Journal of Nuclear Medicine, 57(12), 1872–1879. https://doi.org/10.2967/jnumed.116.177295*

*Liu, W., Wang, X., Owens, J. D., & Li, Y. (2020). Energy-based out-of-distribution detection. Advances in Neural Information Processing Systems, 2020-Decem(NeurIPS).*

*Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health, 1(6), e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2*

*Liu, X., Han, C., Cui, Y., Xie, T., Zhang, X., & Wang, X. (2021). Detection and Segmentation of Pelvic Bones Metastases in MRI Images for Patients With Prostate Cancer Based on Deep Learning. Frontiers in Oncology, 11. https://doi.org/10.3389/fonc.2021.773299*

*Liu, Y., Pagliardini, M., Chavdarova, T., & Stich, S. U. (2021, December 9). The Peril of Popular Deep Learning Uncertainty Estimation Methods. Neural Information Processing Systems. http://arxiv.org/abs/2112.05000*

*Lokre, O., Perk, T. G., Weisman, A. J., Govindan, R. M., Chen, S., Chen, M., Eickhoff, J., Liu, G., & Jeraj, R. (2024). Quantitative evaluation of lesion response heterogeneity for superior prognostication of clinical outcome. European Journal of Nuclear Medicine and Molecular Imaging. https://doi.org/10.1007/s00259-024-06764-0*

*Mahalanobis, P. C. (1936). Reprint of: On the Generalized Distance in Statistics. The Indian Journal of Statistics, 80, 1–7.*

*Mai, D. V. C., Drami, I., Pring, E. T., Gould, L. E., Lung, P., Popuri, K., Chow, V., Beg, M. F., Athanasiou, T., & Jenkins, J. T. (2023). A systematic review of automated segmentation of 3D computed-tomography scans for volumetric body composition analysis. In Journal of Cachexia, Sarcopenia and Muscle (Vol. 14, Issue 5, pp. 1973–1986). John Wiley and Sons Inc. https://doi.org/10.1002/jcsm.13310*

*Maruccio, F. C., Eppinga, W., Laves, M. H., Navarro, R. F., Salvi, M., Molinari, F., & Papaconstadopoulos, P. (2024). Clinical assessment of deep learning-based uncertainty maps in lung cancer segmentation. Physics in Medicine and Biology, 69(3). https://doi.org/10.1088/1361-6560/ad1a26*

*McKinley, R., Rebsamen, M., Dätwyler, K., Meier, R., Radojewski, P., & Wiest, R. (2021). Uncertainty-Driven Refinement of Tumor-Core Segmentation Using 3D-to-2D Networks with Label Uncertainty.*

*Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12658 LNCS, 401–411. https://doi.org/10.1007/978-3-030-72084-1_36*

*Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., & Kapur, T. (2020). Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. IEEE Transactions on Medical Imaging, 39(12), 3868–3878. https://doi.org/10.1109/TMI.2020.3006437*

*Mehta, R., Filos, A., Baid, U., Sako, C., McKinley, R., Rebsamen, M., Dätwyler, K., Meier, R., Radojewski, P., Krishnan Murugesan, G., Nalawade, S., Ganesh, C., Wagner, B., Yu, F. F., Fei, B., Madhuranthakam, A. J., Maldjian, J. A., Daza, L., Gómez, C., … Arbel, T. (2022). QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation - Analysis of Ranking Scores and Benchmarking Results. Journal of Machine Learning for Biomedical Imaging, 1, 1–54. https://github.com/RagMeh11/QU-BraTS.*

*Meissen, F., Wiestler, B., Kaissis, G., & Rueckert, D. (2022). On the Pitfalls of Using the Residual Error as Anomaly Score. MIDL, 172. https://github.com/FeliMe/residual-score-pitfalls.*

*Merkow, J., Dorfner, F. J., Yang, X., Ersoy, A., Dasegowda, G., Kalra, M., Lungren, M. P., Bridge, C. P., & Tarapov, I. (2024). Scalable Drift Monitoring in Medical Imaging AI. ArXiv. http://arxiv.org/abs/2410.13174*

*Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2022). Image Segmentation Using Deep Learning: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7), 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968*

*Moreau, N., Rousseau, C., Fourcade, C., Santini, G., Brennan, A., Ferrer, L., Lacombe, M., Guillerminet, C., Colombié, M., Jézéquel, P., Campone, M., Normand, N., & Rubeaux, M. (2022). Automatic segmentation of metastatic breast cancer lesions on18f-fdg pet/ct longitudinal acquisitions for treatment response assessment. Cancers, 14(1). https://doi.org/10.3390/cancers14010101*

*Mucsányi, B., Kirchhof, M., & Joon Oh, S. (2024). Benchmarking Uncertainty Disentanglement: Specialized Uncertainties for Specialized Tasks. Advances in Neural Information Processing Systems. https://github.com/bmucsanyi/untangle.*

*Murase, K., Nakamoto, A., & Tomiyama, N. (2024). Performance of recurrent neural networks with Monte Carlo dropout for predicting pharmacokinetic parameters from dynamic contrast-enhanced magnetic resonance imaging data. Journal of Applied Clinical Medical Physics. https://doi.org/10.1002/acm2.14586*

*Murugesan, B., Liu, B., Galdran, A., Ayed, I. Ben, & Dolz, J. (2023). Calibrating segmentation networks with margin-based label smoothing. Medical Image Analysis, 87. https://doi.org/10.1016/j.media.2023.102826*

*Nair, T., Precup, D., Arnold, D. L., & Arbel, T. (2020). Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation. Medical Image Analysis, 59. https://doi.org/10.1016/j.media.2019.101557*

*Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., & Lakshminarayanan, B. (2019). Do deep generative models know what they don't know? 7th International Conference on Learning Representations, ICLR 2019, 1–19.*

*Nalisnick, E., Matsukawa, A., Teh, Y. W., & Lakshminarayanan, B. (2019). Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality. 4th Workshop on Bayesian Deep Learning, 1–15. http://arxiv.org/abs/1906.02994*

*Ng, K. H., & Wong, J. H. D. (2022). A clarion call to introduce artificial intelligence (AI) in postgraduate medical physics curriculum. In Physical and Engineering Sciences in Medicine (Vol. 45, Issue 1). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s13246-022-01099-2*

*Nomura, Y., Wang, J., Shirato, H., Shimizu, S., & Xing, L. (2020). Fast spot-scanning proton dose calculation method with uncertainty quantification using a three-dimensional convolutional neural network. Physics in Medicine and Biology, 65(21). https://doi.org/10.1088/1361-6560/aba164*

*Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019, June 6). Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. Neural Information Processing Systems. http://arxiv.org/abs/1906.02530*

*Ozaki, K., Higuchi, S., Kimura, H., & Gabata, T. (2022). Liver Metastases: Correlation between Imaging Features and Pathomolecular Environments. Radiographics, 42(7), 1994–2013. https://doi.org/10.1148/rg.220056*

*Pan, X., Sidky, E. Y., & Vannier, M. (2009). Why do commercial CT scanners still employ traditional, filtered back-projection for image reconstruction? In Inverse Problems (Vol. 25, Issue 12). Institute of Physics Publishing. https://doi.org/10.1088/0266-5611/25/12/123009*

*Perk, T., Bradshaw, T., Chen, S., Im, H. J., Cho, S., Perlman, S., Liu, G., & Jeraj, R. (2018). Automated classification of benign and malignant lesions in 18F-NaF PET/CT images using machine learning. Physics in Medicine and Biology, 63(22). https://doi.org/10.1088/1361-6560/aaebd0*

*Perk, T., Chen, S., Harmon, S., Lin, C., Bradshaw, T., Perlman, S., Liu, G., & Jeraj, R. (2018). A statistically optimized regional thresholding method (SORT) for bone lesion detection in 18F-NaF PET/CT imaging. Physics in Medicine and Biology, 63(22). https://doi.org/10.1088/1361-6560/aaebba*

*Rana, M., & Bhushan, M. (2023). Machine learning and deep learning approach for medical image analysis: diagnosis to detection. Multimedia Tools and Applications, 82(17), 26731–26769. https://doi.org/10.1007/s11042-022-14305-w*

*Reinking, M. F., & Osman, M. M. (2009). Prospective evaluation of physiologic uptake detected with true whole-body 18F-FDG PET/CT in healthy subjects. Journal of Nuclear Medicine Technology, 37(1), 31–37. https://doi.org/10.2967/jnmt.108.055004*

*Ren, J., Teuwen, J., Nijkamp, J., Rasmussen, M., Gouw, Z., Grau Eriksen, J., Sonke, J. J., & Korreman, S. (2024). Enhancing the reliability of deep learning-based head and neck tumour segmentation using uncertainty estimation with multi-modal images. Physics in Medicine and Biology, 69(16). https://doi.org/10.1088/1361-6560/ad682d*

*Rezende, D. J., & Mohamed, S. (2016). Variational Inference with Normalizing Flows. ArXiv. http://arxiv.org/abs/1505.05770*

*Riihimäki, M., Thomsen, H., Sundquist, K., Sundquist, J., & Hemminki, K. (2018). Clinical landscape of cancer metastases. Cancer Medicine, 7(11), 5534–5542. https://doi.org/10.1002/cam4.1697*

*Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. ArXiv. http://arxiv.org/abs/1505.04597*

*Rousseau, A. J., Becker, T., Bertels, J., Blaschko, M. B., & Valkenborg, D. (2021). Post training uncertainty calibration of deep networks for medical image segmentation. Proceedings - International Symposium on Biomedical Imaging, 2021-April, 1052–1056. https://doi.org/10.1109/ISBI48211.2021.9434131*

*Sahiner, B., Chen, W., Samala, R. K., & Petrick, N. (2023). Data drift in medical machine learning: implications and potential remedies. In British Journal of Radiology (Vol. 96, Issue 1150). British Institute of Radiology. https://doi.org/10.1259/bjr.20220878*

*Sahlsten, J., Jaskari, J., Wahid, K. A., Ahmed, S., Glerean, E., He, R., Kann, B. H., Mäkitie, A., Fuller, C. D., Naser, M. A., & Kaski, K. (2024). Application of simultaneous uncertainty quantification and segmentation for oropharyngeal cancer use-case with Bayesian deep learning. Communications Medicine, 4(1). https://doi.org/10.1038/s43856-024-00528-5*

*Sahu, S., Schernthaner, R., Ardon, R., Chapiro, J., Zhao, Y., Sohn, J. H., Fleckenstein, F., Lin, M. De, Geschwind, J. F., & Duran, R. (2017). Imaging biomarkers of tumor response in neuroendocrine liver metastases treated with transarterial chemoembolization: Can enhancing tumor burden of the whole liver help predict patient sur vival? Radiology, 283(3), 883–894. https://doi.org/10.1148/radiol.2016160838*

*Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. ArXiv, arXiv:1708.08296. http://arxiv.org/abs/1708.08296*

*Santilli, A., Panyam, P., Autz, A., Wray, R., Philip, J., Elnajjar, P., Swinburne, N., & Mayerhoefer, M. (2023). Automated full body tumor segmentation in DOTATATE PET/CT for neuroendocrine cancer patients. International Journal of Computer Assisted Radiology and Surgery, 18(11), 2083–2090. https://doi.org/10.1007/s11548-023-02968-1*

*Santoro-Fernandes, V., Huff, D. T., Rivetti, L., Deatsch, A., Schott, B., Perlman, S. B., & Jeraj, R. (2024). An automated methodology for whole-body, multimodality tracking of individual cancer lesions. Physics in Medicine and Biology, 69(8). https://doi.org/10.1088/1361-6560/ad31c6*

*Santoro-Fernandes, V., Schott, B., Deatsch, A., Keigley, Q., Francken, T., Iyer, R., Fountzilas, C., Perlman, S., & Jeraj, R. (2024). Models using comprehensive, lesion-level, longitudinal [68Ga]Ga-DOTA-TATE PET-derived features lead to superior outcome prediction in neuroendocrine tumor patients treated with [177Lu]Lu-DOTA-TATE. European Journal of Nuclear Medicine and Molecular Imaging. https://doi.org/10.1007/s00259-024-06767-x*

*Santoro-Fernandes, V., Schott, B., Weisman, A. J., Lokre, O., Cho, S. Y., Perlman, S. B., Perk, T. G., & Jeraj, R. (2025). Full-Body Tumor Response Heterogeneity of Metastatic Neuroendocrine Tumor Patients Undergoing Peptide Receptor Radiopharmaceutical Therapy. Journal of Nuclear Medicine, 66(4), 565–571. https://doi.org/10.2967/jnumed.124.267809*

*Savjani, R. R., Lauria, M., Bose, S., Deng, J., Yuan, Y., & Andrearczyk, V. (2022). Automated Tumor Segmentation in Radiotherapy. In Seminars in Radiation Oncology (Vol. 32, Issue 4, pp. 319–329). W.B. Saunders. https://doi.org/10.1016/j.semradonc.2022.06.002*

*Schott, B., Klaneček, Ž., Deatsch, A., Santoro-Fernandes, V., Francken, T., Perlman, S., & Jeraj, R. (2025). Information Bottleneck-Based Feature Weighting for Enhanced Medical Image Out-of-Distribution Detection. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, 128–137. https://doi.org/10.1007/978-3-031-73158-7_12*

*Schott, B., Pinchuk, D., Santoro-Fernandes, V., Klaneček, Ž., Rivetti, L., Deatsch, A., Perlman, S., Li, Y., & Jeraj, R. (2024). Uncertainty quantification via localized gradients for deep learning-based medical image assessments. Physics in Medicine and Biology, 69(15). https://doi.org/10.1088/1361-6560/ad611d*

*Schott, B., Santoro-Fernandes, V., Klaneček, Ž., Perlman, S., & Jeraj, R. (2025). Uncertainty quantification for deep learning-based metastatic lesion segmentation on whole body PET/CT. Physics in Medicine & Biology, 70(11), 115009. https://doi.org/10.1088/1361-6560/add9df*

*Schott, B., Weisman, A. J., Perk, T. G., Roth, A. R., Liu, G., & Jeraj, R. (2023). Comparison of automated full-body bone metastases delineation methods and their corresponding prognostic power. Physics in Medicine and Biology, 68(3). https://doi.org/10.1088/1361-6560/acaf22*

*Schulz, K., Sixt, L., Tombari, F., & Landgraf, T. (2020). Restricting the Flow: Information Bottlenecks for Attribution. http://arxiv.org/abs/2001.00396*

*Schwaiger, A., Sinhamahapatra, P., Gansloser, J., & Roscher, K. (2020). Is Uncertainty Quantification in Deep Learning Sufficient for Out-of-Distribution Detection? International Joing Conference on Artificial Intelligence.*

*Schwenck, J., Sonanini, D., Cotton, J. M., Rammensee, H. G., la Fougère, C., Zender, L., & Pichler, B. J. (2023). Advances in PET imaging of cancer. In Nature Reviews Cancer (Vol. 23, Issue 7, pp. 474–490). Nature Research. https://doi.org/10.1038/s41568-023-00576-4*

*Seçkin Ayhan, M., & Berens, P. (2018). Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks. Medical Imaging with Deep Learning.*

*Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. Advances in Neural Information Processing Systems, 2018-Decem(NeurIPS), 3179–3189.*

*Shafer, G. (1976). A mathematical theory of evidence (Vol. 42). Princeton university press.*

*Shafer, G., & Vovk, V. (2008). A Tutorial on Conformal Prediction. Journal of Machine Learning Research, 9, 371–421.*

*Sharabi, A., Kim, S. S., Kato, S., Sanders, P. D., Patel, S. P., Sanghvi, P., Weihe, E., & Kurzrock, R. (2017). Exceptional Response to Nivolumab and Stereotactic Body Radiation Therapy (SBRT) in Neuroendocrine Cervical Carcinoma with High Tumor Mutational Burden: Management Considerations from the Center For Personalized Cancer Therapy at UC San Diego Moores Cancer Center. The Oncologist, 22(6), 631–637. https://doi.org/10.1634/theoncologist.2016-0517*

*Sheikhbahaei, S., Jones, K. M., Werner, R. A., Salas-Fragomeni, R. A., Marcus, C. V., Higuchi, T., Rowe, S. P., Solnes, L. B., & Javadi, M. S. (2019). 18 F-NaF-PET/CT for the detection of bone metastasis in prostate cancer: a meta-analysis of diagnostic accuracy studies. Annals of Nuclear Medicine. https://doi.org/10.1007/s12149-019-01343-y*

*Shin, T. Y., Kim, H., Lee, J. H., Choi, J. S., Min, H. S., Cho, H., Kim, K., Kang, G., Kim, J., Yoon, S., Park, H., Hwang, Y. U., Kim, H. J., Han, M., Bae, E., Yoon, J. W., Rha, K. H., & Lee, Y. S. (2020). Expert-level segmentation using deep learning for volumetry of polycystic kidney and liver. Investigative and Clinical Urology, 61(6), 555–564. https://doi.org/10.4111/icu.20200086*

*Sica, G. T., Ji, H., & Ros, P. R. (2000). CT and MR Imaging of Hepatic Metastases. American Journal of Roentgenology , 174(3), 691–698.*

*Simpson, A. L., Peoples, J., Creasy, J. M., Fichtinger, G., Gangai, N., Keshavamurthy, K. N., Lasso, A., Shia, J., D'Angelica, M. I., & Do, R. K. G. (2024). Preoperative CT and survival data for patients undergoing resection of colorectal liver metastases. Scientific Data, 11(1). https://doi.org/10.1038/s41597-024-02981-2*

*Song, Y., Ren, S., Lu, Y., Fu, X., & Wong, K. K. L. (2022). Deep learning-based automatic segmentation of images in cardiac radiography: A promising challenge. In Computer Methods and Programs in Biomedicine (Vol. 220). Elsevier Ireland Ltd. https://doi.org/10.1016/j.cmpb.2022.106821*

*Sun, Y., Ming, Y., Zhu, X., & Li, Y. (2022). Out-of-Distribution Detection with Deep Nearest Neighbors. ICML. https://github.com/*

*Swiha, M., Papa, N., Sabahi, Z., Ayati, N., John, N., Pathmanandavel, S., Crumbaker, M., Li, S., Agrawal, S., Ayers, M., Hickey, A., Sharma, S., Nguyen, A., & Emmett, L. (2024). Development of a Visually Calculated SUVmean (HIT Score) on Screening PSMA PET/CT to Predict Treatment Response to 177Lu-PSMA Therapy: Comparison with Quantitative SUVmean and Patient Outcomes. Journal of Nuclear Medicine : Official Publication, Society of Nuclear Medicine, 65(6), 904–908. https://doi.org/10.2967/jnumed.123.267014*

*Thagaard, J., Hauberg, S., van der Vegt, B., Ebstrup, T., Hansen, J. D., & Dahl, A. B. (2020). Can you trust predictive uncertainty under real dataset shifts in digital pathology? Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12261 LNCS, 824–833. https://doi.org/10.1007/978-3-030-59710-8_80*

*Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. ArXiv, physics/0004057.*

*Tschuchnig, M. E., & Gadermayr, M. (2024). Anomaly Detection in Medical Imaging -- A Mini Review. ArXiv. https://doi.org/10.1007/978-3-658-36295-9_5*

*Valdenegro-Toro, M., & Mori, D. S. (2022). A Deeper Look into Aleatoric and Epistemic Uncertainty Disentanglement. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2022-June, 1508–1516. https://doi.org/10.1109/CVPRW56347.2022.00157*

*Vorontsov, E., Cerny, M., Régnier, P., Di Jorio, L., Pal, C. J., Lapointe, R., Vandenbroucke-Menu, F., Turcotte, S., Kadoury, S., & Tang, A. (2019). Deep learning for automated segmentation of liver lesions at ct in patients with colorectal cancer liver metastases. Radiology: Artificial Intelligence, 1(2). https://doi.org/10.1148/ryai.2019180014*

*Wahl, R. L., Jacene, H., Kasamon, Y., & Lodge, M. A. (2009). From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumors. In Journal of Nuclear Medicine (Vol. 50, Issue SUPPL. 1). https://doi.org/10.2967/jnumed.108.057307*

*Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., & Vercauteren, T. (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing, 338, 34–45. https://doi.org/10.1016/j.neucom.2019.01.103*

*Wang, T., Lei, Y., Fu, Y., Wynne, J. F., Curran, W. J., Liu, T., & Yang, X. (2021). A review on medical imaging synthesis using deep learning and its clinical applications. In Journal of Applied Clinical Medical Physics (Vol. 22, Issue 1, pp. 11–36). John Wiley and Sons Ltd. https://doi.org/10.1002/acm2.13121*

*Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., & Liu, Q. (2023). Aligning Large Language Models with Human: A Survey. ArXiv. http://arxiv.org/abs/2307.12966*

*Wang, Z., Li, L., Yang, X., Teng, H., Wu, X., Chen, Z., Wang, Z., & Chen, G. (2022). Efficacy and safety of stereotactic body radiotherapy for painful bone metastases: Evidence from randomized controlled trials. In Frontiers in Oncology (Vol. 12). Frontiers Media S.A. https://doi.org/10.3389/fonc.2022.979201*

*Webb, J. M., Adusei, S. A., Wang, Y., Samreen, N., Adler, K., Meixner, D. D., Fazzio, R. T., Fatemi, M., & Alizad, A. (2021). Comparing deep learning-based automatic segmentation of breast masses to expert interobserver variability in ultrasound imaging. Computers in Biology and Medicine, 139. https://doi.org/10.1016/j.compbiomed.2021.104966*

*Weber, M., Kersting, D., Umutlu, L., Schäfers, M., Rischpler, C., Fendler, W. P., Buvat, I., Herrmann, K., & Seifert, R. (2021). Just another "Clever Hans"? Neural networks and FDG PET-CT to predict the outcome of patients with breast cancer. European Journal of Nuclear Medicine and Molecular Imaging, 48, 3141–3150. https://doi.org/10.1007/s00259-021-05270-x/Published*

*Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022, June 15). Emergent Abilities of Large Language Models. Transactions on Machine Learning Research. http://arxiv.org/abs/2206.07682*

*Weisman, A. J., Kieler, M. W., Perlman, S. B., Hutchings, M., Jeraj, R., Kostakoglu, L., & Bradshaw, T. J. (2020). Convolutional neural networks for automated pet/ct detection of diseased lymph node burden in patients with lymphoma. Radiology: Artificial Intelligence, 2(5), 1–2. https://doi.org/10.1148/ryai.2020200016*

*Welford, B. P. (1962). Note on a Method for Calculating Corrected Sums of Squares and Products. Technometrics, 4(3), 419–420.*

*Woodland, M., Patel, N., Al Taie, M., Yung, J. P., Netherton, T. J., Patel, A. B., & Brock, K. K. (2023). Dimensionality Reduction for Improving Out-of-Distribution Detection in Medical Image Segmentation. In C. H. Sudre, C. F. Baumgartner, A. Dalca, R. Mehta, C. Qin, & W. M. Wells (Eds.), Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (Vol. 14291, pp. 147–156). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44336-7*

*Wu, D. H., Pen, O., Wang, Y., Stern, R., Bourland, J. D., & Mahesh, M. (2024). Embracing Real AI: A call to action for medical physicists in healthcare. In Journal of Applied Clinical Medical Physics. John Wiley and Sons Ltd. https://doi.org/10.1002/acm2.14456*

*Xu, F., Lin, Q., Han, J., Zhao, T., Liu, J., & Cambria, E. (2023). Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation and Beyond. IEEE Transactions on Knowledge and Data Engineering, 37(4), 1620–1634. https://doi.org/10.1109/TKDE.2025.3536008*

*Xu, Y., Quan, R., Xu, W., Huang, Y., Chen, X., & Liu, F. (2024). Advances in Medical Image Segmentation: A Comprehensive Review of Traditional, Deep Learning and Hybrid Approaches. In Bioengineering (Vol. 11, Issue 10). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/bioengineering11101034*

*Xue, H., Fang, Q., Yao, Y., & Teng, Y. (2023). 3D PET/CT tumor segmentation based on nnU-Net with GCN refinement. Physics in Medicine and Biology, 68(18). https://doi.org/10.1088/1361-6560/acede6*

*Yang, J., Zhou, K., Li, Y., & Liu, Z. (2024). Generalized Out-of-Distribution Detection: A Survey. International Journal of Computer Vision. https://doi.org/10.1007/s11263-024-02117-4*

*Yu, L., Shiung, M., Jondal, D., & McCollough, C. H. (2012). Development and Validation of a Practical Lower-Dose-Simulation Tool for Optimizing Computed Tomography Scan Protocols. J Comput Assist Tomogr &, 36(4), 477–487. www.jcat.org477*

*Zamzmi, G., Venkatesh, K., Nelson, B., Prathapan, S., Yi, P., Sahiner, B., & Delfino, J. G. (2024). Out-of-Distribution Detection and Radiological Data Monitoring Using Statistical Process Control. Journal of Imaging Informatics in Medicine. https://doi.org/10.1007/s10278-024-01212-9*

*Zanca, F., Hernandez-Giron, I., Avanzo, M., Guidi, G., Crijns, W., Diaz, O., Kagadis, G. C., Rampado, O., Lønne, P. I., Ken, S., Colgan, N., Zaidi, H., Zakaria, G. A., & Kortesniemi, M. (2021). Expanding the medical physicist curricular and professional programme to include Artificial Intelligence. Physica Medica, 83, 174–183. https://doi.org/10.1016/j.ejmp.2021.01.069*

*Zhmoginov, A., Fischer, I., & Sandler, M. (2020, July 22). Information-Bottleneck Approach to Salient Region Discovery. ECML PKDD. http://arxiv.org/abs/1907.09578*

*Zimmerer, D., Full, P. M., Isensee, F., Jager, P., Adler, T., Petersen, J., Kohler, G., Ross, T., Reinke, A., Kascenas, A., Jensen, B. S., O'Neil, A. Q., Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B., Shvetsova, N., Fedulova, I., … Maier-Hein, K. (2022). MOOD 2020: A Public Benchmark for Out-of-Distribution Detection and Localization on Medical Images. IEEE Transactions on Medical Imaging, 41(10), 2728–2738. https://doi.org/10.1109/TMI.2022.3170077*