Robustness Analysis of a Deep Learning based sCT generation

Auteur : Delporte, Guillaume
Promoteur(s) : Phillips, Christophe
Faculté : Faculté des Sciences appliquées
Diplôme : Master : ingénieur civil en science des données, à finalité spécialisée
Année académique : 2024-2025
URI/URL : http://hdl.handle.net/2268.2/23235

University of Liège
School of Engineering and Computer Science

Robustness Analysis of a Deep Learning based sCT generation algorithm

Academic supervisor
Prof. Christophe Phillips
Industrial supervisor
Ir. Geoffroy Herbin

Jury members
Prof. Pierre Geurts
Prof. Christophe Debruyne
Ir. Geoffroy Herbin

Master's thesis completed by Guillaume Delporte
in order to obtain the degree of Master of Science in Data Science and Engineering

Academic year 2024 – 2025

Abstract

In the context of adaptive proton therapy, a treatment strategy that dynamically adjusts to anatomical changes throughout the course of radiotherapy, generating accurate synthetic CT (sCT) images from cone-beam CT (CBCT) scans remains a major challenge. This is primarily due to the presence of artefacts in CBCT and its limited accuracy in Hounsfield Units (HU), which compromises its reliability for dose calculations. Recent advances in deep learning have enabled promising approaches for direct CBCT-to-sCT translation. However, these methods typically lack robust mechanisms for quantifying uncertainty, which is essential for clinical decision-making, particularly in proton therapy, where dose distribution is highly sensitive to anatomical and HU variations. This thesis investigates the potential of two complementary state-of-the-art uncertainty quantification techniques to enhance the trustworthiness and interpretability of deep learning-based sCT generation models.

The study focuses on two main types of uncertainty: epistemic, which reflects model ignorance, and aleatoric, which captures data-inherent noise. Monte Carlo Dropout (MCD) and heteroscedastic regression are used to model each, respectively, within a U-Net architecture. The models are evaluated on a simulated dataset from IBA and on real patient data, which presented substantial artefacts and was preprocessed accordingly by IBA. Evaluation relies on multiple metrics, including MAE, RMSE, SSIM, PSNR, Expected Calibration Error (ECE), and the Pearson correlation between uncertainty and error (PCC).

In the simulated setting, MCD alone achieves an MAE of 21.08 ± 2.29 HU, and ECE of 14.63 HU, with a moderate correlation between uncertainty and error (PCC = 0.54). The combined model improves these metrics, reaching an MAE of 20.65 ± 2.00 HU, ECE of 4.41 HU, and PCC of 0.65. On the real dataset, where artefacts and noise are more prevalent, the combined model still improves over MCD alone though performance remains limited.

The discussion highlights that while MCD uncertainty maps aligns with general error structure in both datasets, they tend to be overconfident and insufficiently calibrated. The introduction of aleatoric modeling improves calibration and interpretability, particularly in identifying regions of anatomical ambiguity or noise. However, A failure to disentangle both uncertainty types still limit clinical applicability.

These findings suggest that while current uncertainty quantification methods such as MCD and heteroscedastic regression represent meaningful steps toward more reliable sCT generation, they remain insufficient to fully meet clinical trustworthiness and interpretability demands.

Résumé

Dans le cadre de la protonthérapie adaptative, qui s'ajuste aux changements anatomiques au fil du traitement, la génération de synthetic CT (sCT) à partir de scans cone-beam CT (CBCT) reste un défi majeur. Cela s'explique par les artefacts présents dans les CBCT et leur faible précision en Hounsfield Units (HU), ce qui limite leur fiabilité pour les calculs de dose. Les récents progrès en deep learning ont permis des approches prometteuses pour la traduction directe CBCT-vers-CT. Toutefois, ces méthodes manquent souvent de mécanismes solides de quantification de l'incertitude, pourtant cruciaux en clinique, notamment en protonthérapie où la distribution de dose est très sensible aux variations anatomiques et aux HU.

Ce mémoire étudie le potentiel de deux techniques d'estimation d'incertitude de pointe et complémentaires pour renforcer la fiabilité et l'interprétabilité des modèles de génération de sCT basés sur le deep learning. L'étude se concentre sur deux types d'incertitude : l'incertitude épistémique, liée à l'ignorance du modèle, et l'incertitude aléatorique, qui reflète le bruit intrinsèque aux données. Le Monte Carlo Dropout (MCD) et la régression hétéroscédastique sont utilisés pour modéliser respectivement chacune d'elles, au sein d'une architecture U-Net. Les modèles sont évalués sur un dataset simulées fourni par IBA, ainsi que sur des données réelles de patients, contenant de nombreux artefacts et prétraitées par IBA. L'évaluation repose sur plusieurs métriques : MAE, RMSE, SSIM, PSNR, Expected Calibration Error (ECE) et la corrélation entre incertitude et erreur (PCC).

Sur le dataset simulé, le modèle MCD atteint une MAE de 21.08 ± 2.29 HU et une ECE de 14.63 HU, avec une corrélation modérée entre incertitude et erreur (PCC = 0.54). Le modèle combiné améliore ces résultats, atteignant une MAE de 20.65 ± 2.00 HU, une ECE de 4.41 HU et un PCC de 0.65. Sur les données cliniques, avec plus d'artefacts, le modèle combiné reste supérieu, bien que les performances restent limitées.

La discussion met en évidence que, si les cartes d'incertitude issues du MCD sont généralement alignées avec la structure de l'erreur, elles ont tendance à être trop confiantes et insuffisamment calibrées. L'introduction de la modélisation aléatorique améliore la calibration et l'interprétabilité, notamment en identifiant les zones d'ambiguïté anatomique ou de bruit. Toutefois, l'incapacité à distinguer clairement les deux types d'incertitude limite encore l'applicabilité clinique.

Ces résultats suggèrent que, bien que les méthodes utilisées pour quantifier l'incertitude représentent des avancées vers une génération plus fiable de sCT, elles restent insuffisantes pour répondre pleinement aux exigences cliniques en matière de confiance et d'interprétabilité.

Acknowledgments

I would like to express my sincere gratitude to IBA for providing the infrastructure, resources, and a collaborative environment that made this thesis possible.

I am especially thankful to Geoffroy, whose guidance, expertise, and thoughtful feedback were instrumental throughout this work.

I would also like to thank Prof. Christophe Phillips, my academic supervisor, for his support, encouragement, and valuable insights during writing of this thesis.

A heartfelt thank you goes to Rachel, whose unwavering love, patience, and encouragement were a constant source of strength. Her presence was a quiet but essential anchor throughout this journey and made the day-to-day logistics so much easier, allowing me to fully focus on my research.

I am equally grateful to Fabien, my best friend, for his daily motivation and practical help. His reliability and good humor made even the most stressful situations feel lighthearted, bringing much-needed balance and perspective during this intense period.

I am also deeply thankful to Rayane, who provided steady academic support throughout this journey. Our regular check-ins and discussions helped me stay on track and maintain momentum, and his presence brought both clarity and motivation at key moments in the process. What I will remember most are the words we often shared, quoting Camus: *"One must imagine Sisyphus happy."* A quiet motto that carried us through.

Finally, I want to thank my dad for guiding me toward these magnificent civil engineering studies, which turned out to be a path I truly love. I am also deeply grateful to my mom, who has shown me, through her strength and resilience, how powerful quiet determination can be. I love you.

Contents

Chapter 1

Introduction and Research Context

1.1 Introduction

Modern radiation therapy (RT) aims to achieve an optimal balance between tumor eradication and the preservation of healthy tissues. This balance critically depends on the accuracy of the dose distribution. This dose distribution in turn relies on medical imaging to create a treatment planning. This is typically done using a CT scan, namely called "planning CT" (pCT), which provides a 3D representation of the patient's anatomy . However, the patient's anatomy may differ largely from the pCT at time of the treatment, as the patient could gain or lose weight, or the tumor could shrink. As such, Adaptive Radiation Therapy (ART) aims to leverage daily imaging to account for these changes and ensure that the delivered dose remains consistent with the original treatment plan. ART is made possible by onboard imaging systems (OBI) that allow for daily imaging of the patient and thus for an adaptation of the treatment plan (Ghaznavi et al. 2025; Paganetti et al. 2021).

Proton therapy (PT) is an advanced form of radiation therapy that uses protons instead of photons to deliver dose to tumors. Unlike photons, which gradually deposit energy along their path, protons release most of their energy at a specific depth in tissue, a phenomenon known as the *Bragg peak*. This allows the delivery of high radiation dose to the tumor while sparing healthy tissues located before and beyond the target. As can be seen in Figure 1.1, it allows for a precise dose distribution, unlike conventional radiotherapy, which uses high-energy X-rays (photons) to destroy cancer cells. These photons travel through the body and release energy both before and after reaching the tumor, which can lead to significant radiation exposure to surrounding healthy tissues.

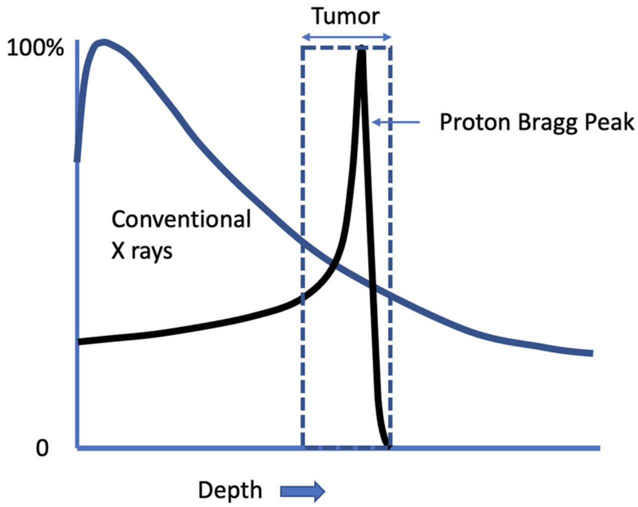


Figure 1.1. Comparison of dose distribution between proton therapy and conventional photon therapy. The Bragg peak allows for a more precise dose delivery to the tumor whereas conventional photon therapy leads to a dose distribution that cannot spare healthy tissues. This figure is taken from Lapen et al. (2023).

While the benefits of proton therapy are well established, the use of the *Bragg peak* also appeals for tighter control of the dose distribution. That is because any change in the patient's anatomy can lead to a shift of the dose distribution, which can be detrimental to the treatment. Hence, Adaptive Proton Therapy (APT) aims to combine the advantages of proton therapy to the daily treatment adaptation capabilities of ART (Ghaznavi et al. 2025).

In clinical proton therapy workflows, Cone-Beam Computed Tomography (CBCT) is a popular imaging modality due to its availability and integration possibilities in treatment rooms (Landry et al. 2018). It also has the advantages of being low-cost and to give a lower radiation dose during acquisition (Ghaznavi et al. 2025). However, the image quality of CBCT remain insufficient for quantitative dose planning, as some artefacts are caused by the geometry of the CBCT modality. To overcome these limitations, a proposed solution is the adaptation of the CBCT using miscellaneous methods such as deformable image registration (DIR) or analytical image-based correction method (AIC) to map the CBCT into CT space (Thummerer, Zaffino, et al. 2020).

With the advent of deep learning, this problem can be tackled in a new way. Instead of using these traditional correction methods, deep learning models can learn how to convert CBCT images directly into CT space in what is called a synthetic CT (sCT). These sCTs are designed to look like real CT scans and to contain accurate Hounsfield Units, making them suitable for dose planning in proton therapy (Thummerer, Zaffino, et al. 2020). For more details on Hounsfield Units, see subsection 2.1.3.

Furthermore, sCT generation from CBCT is inherently a regression task, where the goal is to find a mapping from CBCT to sCT. In this type of task, the model always produces a result, even when the input is very different from the data it was trained on. If there is no way to measure how uncertain the prediction is, it becomes difficult to know whether the output is reliable (Kendall and Gal 2017). In a clinical setting, a failure to identify cases where a model is uncertain can have serious consequences and in the case of PT, can lead to a miscalculation of the dose distribution (McGowan et al. 2013).

This thesis aims to critically evaluate the robustness of deep learning based sCT generation. Specifically, it is investigated whether incorporating uncertainty quantification (UQ), through the form of epistemic and aleatoric uncertainty, can enhance the interpretability and reliability of model outputs. Using a publicly available dataset, tailoring it to the problem as well as some simulated data and a well known uncertainty estimation framework, the predictive confidence of the model and its correlation with reconstruction errors is evaluated. In this context, interpretability refers to the degree to which the uncertainty estimates produced by the model can be meaningfully understood and visually or quantitatively related to relevant anatomical structures or image characteristics. An interpretable uncertainty map should highlight regions that are intuitively expected to be less reliable, such as those affected by image noise, metal artifacts, or anatomical ambiguity, and should allow for the distinction between different sources of uncertainty in order to adapt the clinical process. Trustworthiness concerns the reliability and usefulness of uncertainty estimates in reflecting actual prediction error. A trustworthy method should not only correlate with true error but also be well-calibrated, and able to flag out-of-distribution inputs. Ultimately, this leads to the central research question: *Can state-of-the-art deep learning-based uncertainty quantification methods provide trustworthy and interpretable outputs when applied to synthetic CT generation from CBCT images?*

1.2 State of the Art

Uncertainty plays a major role in intelligent systems and understanding what a model does not know can be a way to mitigate some of the risks associated with the deployment of these systems. In the context of deep learning, these highly parametrized models predictions are often taken as-is, with the implicit assumption that the model is right. However, this may not always be the case (Kendall and Gal 2017).

For example, in 2018, Amazon was accused of having a recruitment tool that was biased towards female candidates (BBC News 2018). Another example is Telsa's autopilot confusing a white truck with the sky, leading to a fatal accident (U.S. Department of Transportation, National Highway Traffic Safety Administration 2017). Another example is Google's image recognition system in 2015, which misclassified African American individuals as gorillas. (BBC News 2015). In all of these situations uncertainty quantification could have flagged these predictions as low confidence, leading to human intervention and the possible prevention of these accidents.

In a clinical context, as it is the case for this thesis and wether one is using deep learning or not, robustness and the quantification of uncertainty plays a major role in the decision making process (Yang et al. 2009; Lambrou et al. 2011). As a matter of fact, not taking into account uncertainty can lead to the incapacity to detect errors and can even lead to the under or overdosage to target volumes and organs at risk (OARs) (McGowan et al. 2013). In the context of synthetic CT generation for radiotherapy planning, an undetected error in predicted Hounsfield Units (HU) may result in inaccurate density estimation. This, in turn, can propagate through dose calculation algorithms and lead to incorrect dose delivery within the target volume or OARs.

Moreover, a human expert who is uncertain is able to talk with his colleagues for another opinion, but a model without uncertainty quantification is not able to do so. (Leibig et al. 2017) Hence, a first question that arises is *what exactly is uncertainty, and how can it be decomposed into its various components?*

1.2.1 Aleatoric and Epistemic Uncertainty

In 2009, Kiureghian et al. (2009) first described the binary classification of uncertainty as being aleatoric or epistemic, with the first being inherent randomness or variability that cannot be reduced through additional information, and the latter being uncertainty due to lack of knowledge, which can, in principle, be reduced with more data or better models.

This classification has been extended by Kendall and Gal (2017), who formalized the definition of Kiureghian et al. (2009) in the context of Deep Learning for the use in Neural Networks. He extended the definition of aleatoric uncertainty, separating it as heteroscedastic aleatoric uncertainty, uncertainty that could lead to different level of noise in the output in regards to different inputs and homoscedastic aleatoric uncertainty, which is uncertainty that is constant with the input data.

More recently, these way of defining uncertainty have been further compiled and analyzed by Hüllermeier et al. (2021). They exemplified the difference between aleatoric and epistemic uncertainty with the example of a coin toss and language translation. On the one hand and in the case of a coin toss, the aleatoric uncertainty is the fact that the coin can land on either side, and that this cannot be reduced by any means. Therefore, even the hypothetical perfect model would not be able to predict the outcome of the coin toss. On the other hand, epistemic uncertainty that represents the lack of knowledge, can be exemplified with language translation. As stated by Hüllermeier et al. (2021), if one is translating a text and the translation of a specific word is not clear between two options, the probability for each option is said to be the same as in the coin flip case, except that it would be easy to get rid of this uncertainty by adding more text with a representation of this word in the training data.

Figure 1.2 illustrates the difference between aleatoric and epistemic uncertainty in a regression setting. Aleatoric uncertainty is prevalent in the right region where observations are dense but inherently noisy, while epistemic uncertainty is high in the central region where no training data is available.

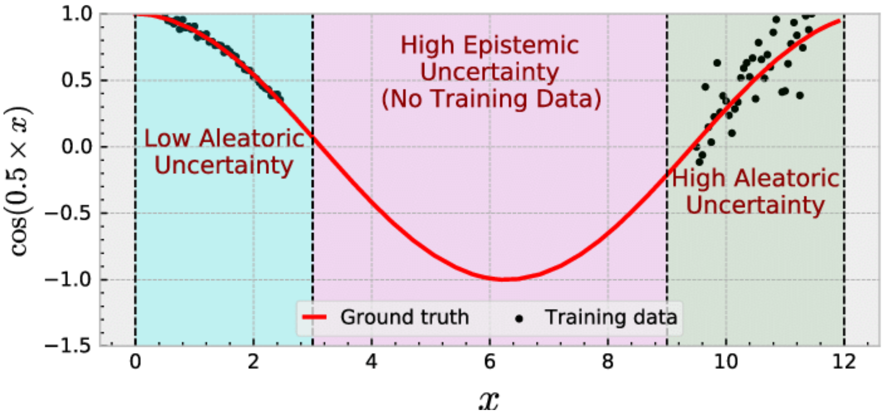


Figure 1.2. Illustration of uncertainty types in a regression setting. Aleatoric uncertainty reflects input-dependent noise, being low in well-observed, low-variance regions (left) and high where the data is inherently noisy (right), while epistemic uncertainty dominates in the central region where no training data is available. Figure taken from Tuna et al. (2022).

Another way of understanding uncertainty is through the concepts of dissonance and vacuity, as described by Gao et al. (2024). Vacuity captures uncertainty coming from a lack of evidence or knowledge and is analogous to epistemic uncertainty. Vacuity is most prominent when the model is exposed to unfamiliar or poorly represented inputs. In contrast, dissonance is present when different sources of information within a single input offer conflicting evidence, leading the model to make inconsistent predictions. Although the model may have sufficient evidence from each part individually, the contradictions among them elevate the overall uncertainty. Dissonance is a more nuanced form of uncertainty that does not align with epistemic or aleatoric categories, but instead incorporates aspects of both (Gao et al. 2024).

However, these two concepts will not be explored further in this thesis, but they are worth mentioning as they provide a different perspective on uncertainty quantification and can be useful in approfounding the comprehension on the matter.

1.2.2 Methods for Uncertainty Quantification

While the distinction between each type of uncertainty provides a mean for understanding uncertainty in a broader context and an important question is: *how can we quantify it in practice when using neural networks?*

This section tries to answer this question and presents the main strategies for quantifying uncertainty in deep learning models, categorized by methodological families. Mainly, the methods are divided into *Approximate Bayesian*, *Ensemble*, *Deterministic*, *Data Centric* and *Post-hoc* methods as can be seen in Figure 1.3. Ultimately, Each method will be analyzed regarding which type of uncertainty it estimates, and its practical advantages and limitations.

To structure this section, the general organization of methods proposed by Abdar et al. (2021) was used as an initial inspiration, as their work provides a broad overview of uncertainty quantification methods across classification, regression, and reinforcement learning tasks.

However, the present work focuses exclusively on regression problems, in regards with the objectives of this thesis being sCT generation as stated in Section 1.1. The selection of methods, the research process, and the analysis were conducted independently in order to stay in line with the relevant context. The reader interested in other contexts, such as classification or reinforcement learning, is encouraged to refer to Abdar et al. (2021).

Approximate Bayesian Methods

One of the first ways to capture epistemic uncertainty is through the use of Bayesian Neural Networks (BNNs) (Neal 2012; MacKay 1992). In these networks the ignorance about the model, meaning the uncertainty over which set of weights best represents the true data-generating process, is represented by defining a prior distribution $p(\omega)$ over the weights ω of the network. This prior is then updated using the training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ via Bayes theorem, resulting in a posterior distribution:

$$p(\omega \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \omega)\, p(\omega)}{p(\mathcal{D})} = \frac{\prod_{i=1}^{N} p(y_i \mid x_i, \omega) p(\omega)}{\int p(\mathcal{D} \mid \omega)\, p(\omega)\, d\omega} \tag{1.1}$$

Here, $p(\mathcal{D} \mid \omega)$ is the likelihood, and $p(\mathcal{D})$ is the evidence, a normalizing constant given by the integral of the likelihood function evaluated over all possible weight configurations.

However, this posterior distribution is often intractable in practice, because, as said, computing the evidence $p(\mathcal{D})$ requires integrating over all possible configurations of the network weights which is generally impossible in deep learning. Therefore, the use of approximate inference methods is required to evaluate or approximate the posterior distribution.

This can be done using Markov Chain Monte Carlo (MCMC) techniques that can generate samples from the desired posterior $p(\omega \mid \mathcal{D})$ by constructing a chain that converges to the target distribution. Neal et al. (2011) applied MCMC to Bayesian Neural Networks and demonstrated how Hamiltonian Monte Carlo (HMC) can be used to efficiently sample from this distribution over neural network weights. However, MCMC methods can be computationally expensive and slow to converge especially in the context of deep learning. This is because in real-world applications where the data can be numerous, each iteration to update the chain can become very costly (Neal et al. 2011).

To overcome this, Stochastic Gradient MCMC (SG-MCMC) methods were introduced, starting with Stochastic Gradient Langevin Dynamics (SGLD) by Welling et al. (2011). These methods combine stochastic optimization with MCMC sampling by using mini batches of data to estimate gradients and by adding noise to the updates (Chen et al. 2014). This transforms standard gradient descent into a procedure that can approximate sampling from the posterior, using ideas from Langevin dynamics. This approach allowed the use of large quantities of data within the framework and was further extended in methods like Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) (Chen et al. 2014).

Another approach to approximate the posterior distribution is through Variational Inference (VI). In this approach, the idea is to build a family of distributions $q(\omega; \nu)$ that approaches the true posterior $p(\omega \mid \mathcal{D})$. The goal is then to find the parameters ν of the variational distribution that minimize $\mathrm{KL}(q(\omega;\nu)\,\|\,p(\omega \mid \mathcal{D}))$, the Kullback-Leibler (KL) divergence between the true posterior and the variational distribution. The KL divergence measures the difference between two probability distributions. Graves (2011) showed that minimizing this quantity is equivalent to maximizing the evidence lower bound (ELBO) defined as:

$$\mathrm{ELBO}(\nu) = \mathbb{E}_{q(\omega;\nu)}\left[\log p(\mathcal{D} \mid \omega)\right] - \mathrm{KL}(q(\omega;\nu)\,\|\,p(\omega)) \tag{1.2}$$

This formulation transforms Bayesian inference into an optimization problem that is compatible with stochastic gradient descent and mini-batch training.

Building on this idea, Blundell et al. (2015) introduced Bayes by Backprop (BBB), a practical method that brings variational inference to Bayesian Neural Networks by placing a Gaussian distribution over each weight in the network. Their main innovation was to use the reparameterization trick, which allows the model to compute gradients even when weights are sampled from a distribution. This trick is necessary because directly sampling weights from a distribution breaks the computational graph, preventing gradient-based optimization. This made it possible to train the networks using standard backpropagation (Rumelhart et al. 1986).

Another way of estimating epistemic uncertainty is based on Dropout, as shown by Srivastava et al. (2014). Dropout is a technique that randomly drops units from the network at each training step and was first developed to prevent overfitting (Krizhevsky et al. 2012). It is said to perform model averaging, as it can be seen as training an ensemble of models with shared weights (Srivastava et al. 2014). However, Gal et al. (2016) showed that Dropout can be interpreted as a form of variational inference in BNN's. As confirmed in the Deep Learning course by Louppe (2024), Dropout corresponds to a specific variational family.

Specifically, one can think of the weights ω as being organized layer by layer, and further divided into individual units within each layer. The variational parameters ν correspond to trainable weight values, denoted by $\mathbf{m}_{i,k}$. Under this variational distribution, each weight is either zeroed out or takes its corresponding trainable value $\mathbf{m}_{i,k}$. Mathematically, this is described as a mixture of two Dirac distributions where one is centered at zero with probability p, and one is centered at $\mathbf{m}_{i,k}$ with probability $1 - p$. This means that each weight has a probability p of being "turned off" during training, and a probability$1 - p$ of keeping its learned value. To sample from this distribution during training, a binary random variable is drawn for each unit to decide whether it is kept or dropped. This process is mathematically equivalent to the standard dropout mechanism, where units are randomly deactivated during training with a fixed probability p.

These methods all learn to estimate uncertainty at training time. In contrast, Ritter et al. (2018), extending the work of MacKay (1992), proposed a method based on the Laplace approximation and that does not need re-training. The Laplace approximation is obtained by taking a second-order Taylor expansion of the log-posterior around its mode (Ritter et al. 2018). This results in a Gaussian approximation of the posterior in Equation 1.1, where the mean is the Maximum a Posteriori estimate and the covariance is an approximation of the Hessian of the log-posterior at that point. The covariance of the Gaussian posterior provides an estimate of epistemic uncertainty in the model parameters.

Ensemble Methods

There are other ways to estimate epistemic uncertainty that does not rely on Bayesian principles. One of them is the use of ensembling (Hansen et al. 1990; Barber et al. 1997) with Deep Ensembles proposed by Lakshminarayanan et al. (2017). Deep Ensembles relies on training multiples networks with different random initializations as well as using adversarial training (Goodfellow et al. 2015). With this ensemble of networks, the epistemic uncertainty is estimated by computing the variance of the predictions across the different networks. This method has been showed to have comparable performance with method relying on BNN's and VI (Lakshminarayanan et al. 2017). However, this approach involves training several networks and handling multiple models at inference, which may be less practical in some scenarios.

Beyond vanilla Deep Ensembles, several frameworks combining ensemble learning with approximate Bayesian inference. One of them is Pearce et al. (2020) anchored ensembling. Another approach from B. He et al. (2020), Bayesian Deep Ensembles via Neural Tangent Kernels, draws a link between deep ensembles and Bayesian inference by using insights from wide neural networks and Gaussian processes. These methods aim to combine the diversity benefits of ensembles with the formal structure of Bayesian inference (B. He et al. 2020).

Deterministic Methods

Instead of relying on sampling or ensembling, some methods aim to estimate epistemic uncertainty in a single network pass. These methods are often refered to as deterministic uncertainty methods (Van Amersfoort et al. 2020) and are therefore more efficient than Bayesian and Ensemble methods (Gao et al. 2024). Determinstic Deep Learning presents three main types of approaches. The first being distance-aware methods, the second being Prior Networks (PN) and the third being Evidential Deep Learning (EDL). While Ulmer et al. (2023) present PN and EDL as a joint family due to their shared use of the Dirichlet distribution and subjective logic, Gao et al. (2024) choose to separate them. They argue that PN model a prior distribution, while EDL directly parametrizes the posterior distribution. Following this distinction, and in line with Gao et al. (2024), this thesis also treats them as two separate categories.

Distant-aware methods are based on the idea that a model that can capture the distance between it's training examples and the test example can provide a measure of uncertainty (Liu et al. 2020). Van Amersfoort et al. (2020) proposed a method called Deterministic Uncertainty Quantification (DUQ). It is built upon RBF network (LeCun et al. 1998). They use a novel loss function and obtain a model that they claim match Deep Ensembles in terms of performance (Van Amersfoort et al. 2020).

Another approach to have epistemic uncertainty quantification while using a distance aware formulation, formulated by Liu et al. (2020), where they first formalize this task as a min-max problem. They then leverage a method they called Spectral-normalized Neural Gaussian Process (SNGP), which is obtained by adding spectral normalization (Miyato et al. 2018) to the weights of the network and relying on the Gaussian process (GP) framework (Rasmussen 2003). They evaluate SNGP on different vision and language tasks and concluded that the results were on par with Deep Ensembles. Still, the authors also critique the precision of the current spectral normalization approach noting that a specific condition may be insufficient in practice, particularly for convolutional layers, noting poor interaction with particular layers.

Though it is useful to have a grasp of these methods to deepen the understanding of the field, these are not developed for regression tasks but for classification and the authors do not specify how to adapt them to regression problems. Moreover, DUQ and SNGP cannot differentiate aleatoric and epistemic uncertainty (Mukhoti et al. 2021).

For that reason, Mukhoti et al. (2021), building upon the work of Van Amersfoort et al. (2020) and Liu et al. (2020), proposed a method called Deep Deterministic Uncertainty (DDU). Their method require minimal architectural change and can disentangle epistemic and aleatoric uncertainty. Additionally, DDU demonstrates competitive performance in out-of-distribution detection and does not rely on fine-tuning with OOD data.

Beyond Distance-Aware techniques, the next class of formulation focuses on explicitly modeling uncertainty through Prior Networks, first formalized by Malinin et al. (2018). Prior Networks are designed to explicitly model distributional uncertainty by parameterizing a distribution over predictive distributions, allowing the model to distinguish between aleatoric uncertainty and uncertainty caused by data shift (Quiñonero-Candela 2009). They are able to achieve better performance than Monte Carlo Dropout for out of distribution (OOD) detection. However, they require careful training with synthetic OOD data and are primarily designed for classification tasks, which complicates their extension to regression (Malinin et al. 2018).

Evidential Deep Learning (EDL) was introduced by Sensoy et al. (2018) for classification and later extended to regression by Amini et al. (2020). Rather than placing priors on network weights as in 1.2.2, EDL places priors directly over the likelihood function by modeling the evidence for each output through higher-order distributions (Amini et al. 2020). Using subjective logic theory (Jsang 2018), it trains the network to predict the parameters of a Dirichlet distribution. It was extended by Amini et al. (2020) to handle regression tasks, where the model predicts the parameters of an Normal Inverse-Gamma (NIG) distribution as Deep Evidential Regression (DER). DER achieves similar or better-calibrated uncertainty estimates and faster inference than Monte Carlo Dropout and Deep Ensembles, while keeping pace in predictive performance. However, its success heavily relies on proper regularization. The distinction made between aleatoric and epistemic uncertainty is not as clear as in other methods. This is explained further in (Gao et al. 2024), a survey on EDL and its applications.

Data Centric Methods

One of the first ways described to capture aleatoric uncertainty is by the use of Mixture Density Networks (MDNs), introduced by Bishop (1994). Rather than providing a single point estimate, an MDN predicts the parameters of a mixture of Gaussians conditioned on the input, thereby modeling a full predictive distribution. This allows the network to account for heteroscedastic uncertainty, with the predicted variance naturally varying across different regions of the input space.

However, practical challenges have limited their adoption. Training instability and difficulties in optimizing mixture models were already noted by Bishop (1994), while mode collapse and the scalability issues of MDNs for high-dimensional outputs have been later (Rupprecht et al. 2017).

Another way is modeling this aleatoric uncertainty directly through the model's likelihood function. In this setting, assuming homoscedasticity, aleatoric uncertainty can be modeled by introducing a fixed or learnable variance parameter as formalized by Kendall and Gal (2017). Although this approach is simple and effective, it assumes that the noise level is constant across the input space.

Another more intricate way proposed by Nix et al. (1994), is to treat the variance of the output no longer as a fixed global parameter but rather as a function of the input. This approach allows the model to learn a distinct data-dependent variance $\sigma^2(x)$ for each input. In practice, Kendall and Gal (2017) achieved this by adding an additional output to the network, that predicts the log-variance $\log(\sigma^2(x))$ of the output. The log-variance is chosen for numerical stability as well as to ensure that the predicted variance is positive.

The loss function of the model is then determined by using the principle of maximum likelihood estimation (MLE), which is analogous to minimizing the KL divergence between the predicted distribution and the true distribution of the data. Instead of maximizing the likelihood of the data given the model, the model is trained to minimize the negative log-likelihood (NLL) of the data.

$$\mathcal{L}(\theta) = -\log p(y \mid x; \theta) = \frac{1}{2\sigma^2(x)}(y - \hat{y}(x))^2 + \frac{1}{2}\log \sigma^2(x) + \frac{1}{2}\log(2\pi) \tag{1.3}$$

The loss in Equation 1.3 consists of two terms as we can ignore the constant during optimization: the first term is the squared error scaled by the predicted variance, and the second term is a regularization term that penalizes large predicted variances. This approach is widely used and has been applied in conjunction with Deep Ensembles (Lakshminarayanan et al. 2017) and Monte Carlo Dropout (Kendall and Gal 2017) as it will be seen in subsection 1.2.3.

However, even if this method is seen as the standard way of estimating aleatoric uncertainty, it is known to be prone to overconfidence in the variance estimates (Skafte et al. 2019). Methods have been proposed to mitigate this issue. Skafte et al. (2019) suggest several strategies. Among many others, ways to provide better variance gradients, splitting the mean and variance optimization steps to stabilize training and fitting an inverse-Gamma distribution instead of a point estimate. Similarly, Stirn and Knowles (2020) propose to treat the variance prediction variationally in order to try to reduce vanishing or exploding variance.

Even if these additions try to mitigate the overconfidence in variance, Seitzer et al. (2022) argue that the mean estimates can be subpar and they propose a new loss, the β-NLL loss. They claim that this loss, if nicely tuned, can largely improve the estimates.

Post-hoc Methods

Another simple yet effective strategy for capturing aleatoric uncertainty is Test-Time Data Augmentation (TTA), as introduced by Ayhan et al. (2018). In this approach, data augmentation is leveraged and multiple test inputs are generated using basic transformations such as random crops, flips, rotations, and color perturbations. The trained model is then evaluated on these perturbed inputs without retraining, and the variability in the predictions is used to estimate the uncertainty. This method best feature is that it does not require any modification to the model architecture or retraining, making it easy to implement and apply to existing models. Hence it's name as a post-hoc method. Ayhan et al. (2018) demonstrated that TTA could effectively capture heteroscedastic aleatoric uncertainty in medical imaging tasks, and that the resulting uncertainty estimates were useful for decision making processes. However, TTA is criticized as it can degrade performance by turning accurate predictions into erroneous ones (Shanmugam et al. 2021).

Conformal Prediction (CP) is another post-hoc method that can be used to estimate uncertainty. Although this approach differs substantially from any other method, as it is a purely statistical framework, it offers formal guarantees on prediction reliability under minimal assumptions. It was first introduced by Shafer et al. (2008). CP assesses how well a new example relies to the training data by computing a nonconformity score and a p-value.

Some work has been built upon this idea, mainly by Papadopoulos (2008), who introduced Inductive Conformal Prediction (ICP) to improve computational efficiency and allows use with neural networks. Another important contribution, by Romano et al. (2019), introduced Conformalized Quantile Regression (CQR), which combines the statistical guarantees of conformal prediction with the flexibility of quantile regression.

1.2.3 Combination of Methods for Uncertainty Quantification

While there are a handful of methods that can be used to estimate both types of uncertainty, except for some of the deterministic methods, most of them are designed to capture either epistemic or aleatoric uncertainty. However, some methods have been proposed to combine different approaches in order to capture both types of uncertainty.

One of the most well-known combination is the one proposed by Kendall and Gal (2017). The authors leverage Monte Carlo Dropout to capture epistemic uncertainty and they parametrize the aleatoric uncertainty through the model's likelihood function. This proposal is simple yet effective as it allows uncertainty estimation in any architecture given the fact that MCDO is implementable in the said architecture. Nonetheless, multiple forward passes are therefore required, which can be computationally expensive. The difficulty also rises with training the heteroscedastic model, as it can be sensitive and lead to very poor estimates (Seitzer et al. 2022).

Other approaches based on Bayesian Inference have also been proposed to capture both types of uncertainty. For example, Depeweg et al. (2018) uses BNNs alongside Latent Variables to model both types of uncertainty. Another notable recent approach by Immer et al. (2023) derives the first efficient Laplace approximation that can be used with heteroscedastic modeling.

Another approach by Lakshminarayanan et al. (2017), is through the use of Deep Ensembles alongside the use of a heteroscedastic likelihood. This method is similar to the one proposed by Kendall and Gal (2017) but instead of using MCDO, they use Deep Ensembles to capture epistemic uncertainty. This method is also simple and effective, but carries the same drawbacks as Deep Ensembles, such as the need to train multiple models and the increased computational cost.

1.3 Research Objectives and Aims

The main goal of this thesis is to investigate the role and relevance of uncertainty quantification in the process of Synthetic Computed Tomography (sCT) generation. More specifically, it aims to assess whether state-of-the-art deep learning-based uncertainty estimation techniques can provide outputs that are both trustworthy and interpretable when applied to sCT generation from Cone Beam Computed Tomography (CBCT) images. The focus is placed on understanding the added value of uncertainty information in this context, particularly in relation to the robustness and potential clinical applicability of such methods.

To achieve this goal, the thesis is structured as follows:

1. Understand the crossing between medical imaging for proton therapy and deep learning as well as the challenges associated with the generation of sCT. This involves understanding the principles of sCT generation and the role of deep learning in this process.

2. Understand the data that will be used in this thesis, as well as it's main limitations and assess some carefully thought pre-processing steps in the hope of improving the data quality and therefore the analysis of the results.

3. Obtain a baseline for the sCT generation process by reproducing the results of previous work. Namely, the work of Thummerer, Zaffino, et al. (2020) and the work of the SynthRad2023 challenge winners (Huijben et al. 2024). Primarily by comparing well known and established model architectures, such as the U-Net (Ronneberger et al. 2015) and the Unet++ (Zhou, Siddiquee, et al. 2019) and deciding, based on results as well as auxiliary information, which model is the most adapted for the task at hand.

4. Implement a state of the art model uncertainty estimation method for regression tasks, specifically the method proposed by Gal et al. (2016), tune it on the chosen baseline and assess its performance.

5. Complement this method with a state of the art aleatoric uncertainty estimation method for regression tasks, specifically the method proposed by Kendall and Gal (2017) and assess its performances.

6. Conduct a thorough analysis of the results obtained with the addition of these uncertainty measures not only in terms of raw performance but also the quality of the uncertainty measure itself. This includes both a qualitative evaluation, in the sense of usability in the medical setting and general behavior and a quantitative evaluation through the use of metrics such as the Pearson correlation coefficient, expected calibration error and calibration plots. The main goal being to assess whether the uncertainty measures are reliable enough to support clinical decision-making.

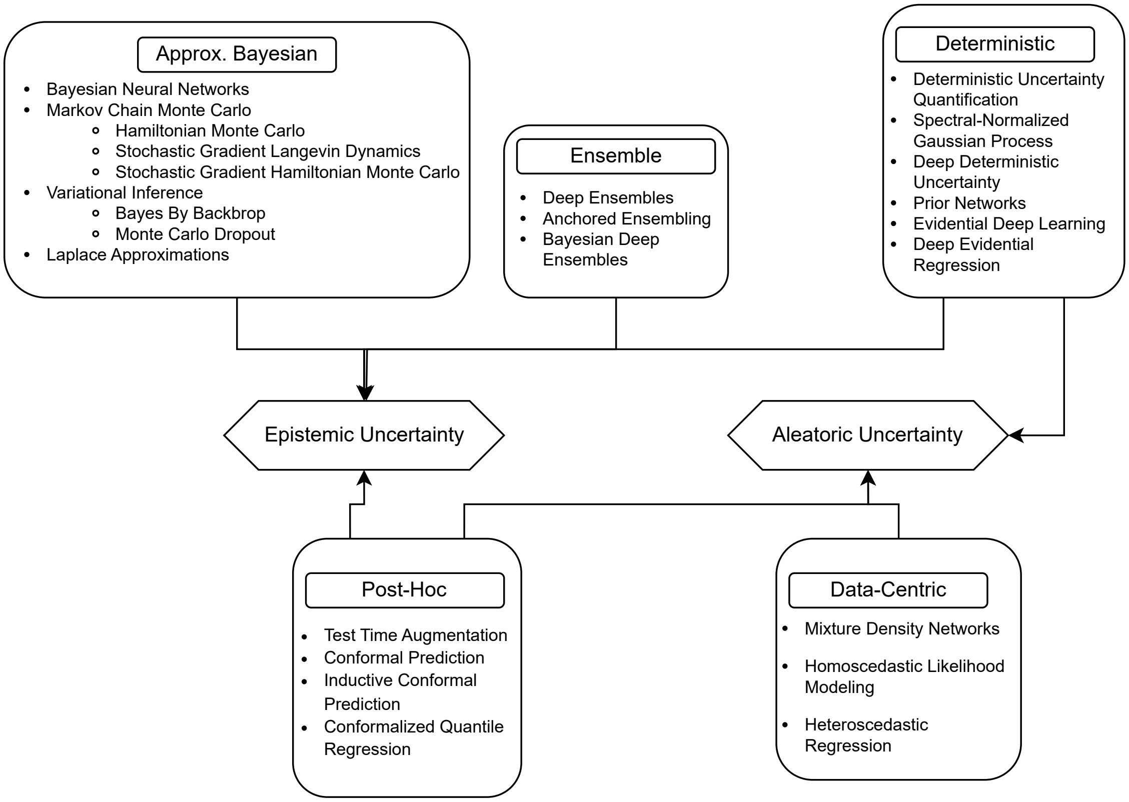


Figure 1.3. This diagram summarizes the principal techniques for uncertainty quantification in deep learning, organized by their methodological nature (Approximate Bayesian, Ensemble, Deterministic, Data-Centric, Post-hoc) and by the type of uncertainty they aim to quantify. Epistemic (model-based uncertainty due to limited knowledge) or aleatoric (data-inherent noise). Approximate Bayesian methods include Bayesian Neural Networks and their tractable variants. Ensemble approaches rely on the diversity across multiple models to estimate uncertainty. Deterministic methods aim to quantify uncertainty in a single forward pass using architectural priors or evidential approaches. Data-centric techniques model noise directly through the likelihood function, while post-hoc methods estimate uncertainty using inference-time techniques. This categorization is adapted from the one proposed by Abdar et al. (2021). The figure is inspired by the one presented by Gao et al. (2024).

Chapter 2

Some Theoretical Background

This chapter introduces the basic theoretical foundations necessary to understand the methods and concepts used throughout this thesis. It first covers essential principles of medical imaging, with a focus on Cone-Beam Computed Tomography (CBCT), its differences with conventional CT, and the artefacts that limit its quantitative use. The second part of the chapter outlines key concepts in deep learning.

2.1 Medical Imaging: Volumetric X-ray Imaging

Among the various medical imaging modalities available today, this thesis focuses on X-ray-based volumetric imaging. Specifically, computed tomography (CT) and cone-beam computed tomography (CBCT) as they are important underlying components of this thesis. The following sections present the physical and mathematical concepts behind CT and CBCT, common artefacts in CBCT, and the motivation for generating synthetic CT images. The main goal is to provide the necessary background to understand the data acquisition process in CT and CBCT imaging, as it is important in data science to know your data and it's provenance.

This section is therefore not intended to be exhaustive, but rather to provide a basic understanding of the principles behind these imaging techniques. The interested reader is invited to refer to the book that this section is based on (Hsieh 2015).

2.1.1 Physics of X-ray Imaging

CT and CBCT are both X-ray based imaging techniques that rely on the physical principle of attenuation to reconstruct a three-dimensional representation of the scanned object using a reconstruction algorithm. This attenuation is governed by the Beer-Lambert law, which describes how the intensity of an X-ray beam decreases as it passes through matter. The general form of the Beer-Lambert law is given in Equation 2.1.

$$I = I_0 \cdot e^{-\int_{ray} \mu(s) ds} \tag{2.1}$$

where I is the transmitted intensity after traversing a material, I_0 is the incident intensity, and $\mu(s)$ is the linear attenuation coefficient at position s along the X-ray path. The variable s represents a point on the ray, and the integral is taken along the path *ray* that the X-ray beam follows as it passes through the object. Both of these imaging techniques therefore share the same physical principle, but they differ in their implementation and application possibilities.

2.1.2 Reconstruction of the Projections

The reconstruction process in both CT and CBCT relies on the attenuation model described by Equation 2.1. Each projection acquired by the scanner represents the integrated attenuation along a straight X-ray path. In other words, the system measures the line integral of the linear attenuation coefficient μ over a given ray path. In fact, dividing Equation 2.1 by I_0 and taking the negative logarithm leads to what is called the projection measurement and can be seen in Equation 2.2.

$$p = -\log\left(\frac{I}{I_0}\right) = \int_{ray} \mu(s)ds \tag{2.2}$$

Equation 2.2 therefore shows that the logarithm of the ratio between the incoming and outgoing X-ray intensities corresponds to the line integral of the attenuation coefficients along the path of the X-ray beam. The reconstruction task in CT thus consists in estimating the internal attenuation distribution of the object based on these measured line integrals. By acquiring a large number of such projections from multiple angles around the object, it becomes possible to mathematically reconstruct the internal structure of the scanned volume.

A simple way to achieve the reconstruction of an image from these line integrals is to use backprojection. In this method, each measured projection is “projected” back across the image volume along the same path it was acquired, distributing its value uniformly along that ray. When projections from multiple angles are combined, their overlaps begin to highlight the actual structures inside the object. However, basic backprojection is fundamentally inadequate for image reconstruction, as it fails to correctly recover the attenuation map μ.

Instead, an exact analytical solution exists based on the Projection Slice Theorem, which establishes a formal link between the 1D Fourier transform of a projection and a slice through the 2D Fourier transform of the object. This theorem provides the foundation for exact inversion methods such as filtered backprojection (FBP), which corrects the shortcomings of basic backprojection by applying an appropriate filtering step in the frequency domain. Modern algorithms build on this analytical foundation, adapting it for digital implementation while managing various practical constraints and approximations. A detailed review of these methods lies beyond the scope of this work. To understand the data at hand, it is sufficient to understand the general principle of how projections are acquired and used to estimate the internal structure of an object. The interested reader is referred to Prince et al. (2006, Chapter 6) for a comprehensive explanation of these concepts.

2.1.3 The Hounsfield Unit

Even if the reconstructed images represent the linear attenuation mapping over the volume, the scientific community has defined a specific intensity scale named the Hounsfield Unit (HU)

The HU is a standardized scale used in CT to quantify the attenuation of X-rays in different tissues. It is defined relative to the attenuation of water, which is assigned a value of 0 HU. In this case, air is assigned a value of -1000 HU.

In regards to the attenuation coefficient μ the corresponding HU of a CT can be derived using Equation 2.3.

$$\mathrm{HU}_{\mathrm{CT}} = 1000 \cdot \left(\frac{\mu_{CT} - \mu_{water}}{\mu_{water}} \right) \tag{2.3}$$

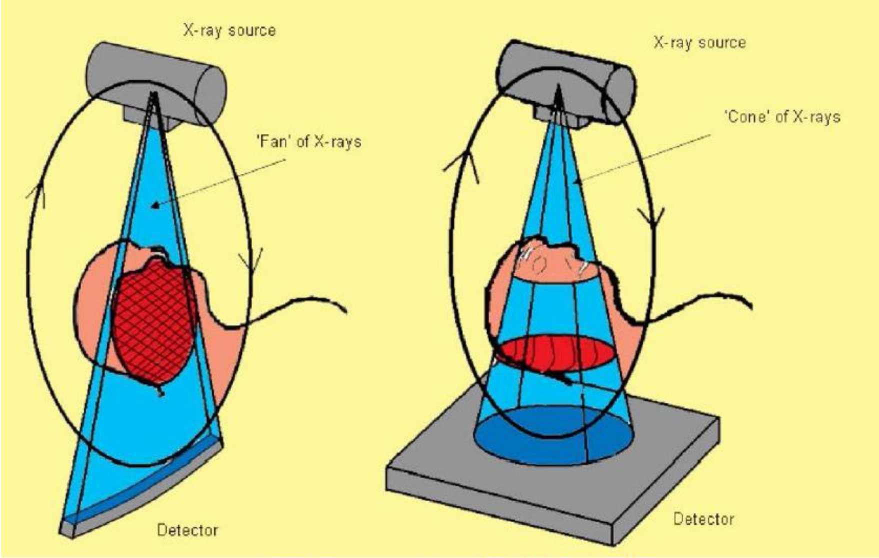
The Table 2.1, adapted from Greenway et al. (2025), provides a list of common materials and their corresponding Hounsfield Unit values.

Tissue/Material	**Typical Attenuation (HU)**
Air	-1000
Bone (cortical)	> 1000
Bone (trabecular)	300 to 800
Brain (gray matter)	40
Brain (white matter)	30
Subcutaneous fat	-100 to -115
Liver	45 to 50
Lungs	-950 to -650
Metal	> 3000
Muscle	45 to 50
Renal cortex	25 to 30
Spleen	40 to 45
Water	0 (by definition)

Table 2.1. Typical CT HU values for various tissues and materials. Water has a value of 0, as it is the calibration material. Adapted from Greenway et al. (2025)

Although CT images fundamentally reconstruct the linear attenuation coefficient μ, the values of μ depend on the effective energy of the X-ray beam, which varies between different scanners and acquisition settings. As a result, the same tissue may appear to have different μ values under different imaging conditions.

To overcome this variability, the Hounsfield Unit (HU) scale therefore provides a standardized reference based on water (0 HU) and air (-1000 HU). This normalization allows clinicians and researchers to compare CT values across different systems, enabling consistent interpretation



© J Can Dent Assoc 2006; 72(1); 75-80

Figure 2.1. Basic setup of a CT scanner (left) versus a CBCT scanner (right). The X-ray source and detector rotate around the patient in a fan-beam geometry for the CT scanner and in cone-beam geometry for the cbct scanner. Figure from Endoblog (2011).

of tissue densities and facilitating the use of thresholds for segmentation, diagnosis, and treatment planning.

2.1.4 Computed Tomography

Concerning the CT, collection of the measurements is typically achieved by rotating an X-ray source and a detector array within a structure known as a gantry. In the last generation of CT, the X-ray source and a curved detector rotate around the patient, while the patient is gradually moved through the gantry along the longitudinal axis. This helical motion allows for the acquisition of a dense set of projections, making accurate volumetric reconstruction possible. This setup is known as a fan-beam data sampling strategy. Figure 2.1 illustrates the basic setup of a CT scanner, where the X-ray source and detector, in fan-beam geometry, rotate around the patient.

Each rotation of the X-ray tube and detector pair captures multiple projections at different angles, which are organized into a dataset known as a sinogram. This sinogram encodes the line integrals of the attenuation coefficients along various paths through the body. Once the full set of projections is collected, image reconstruction algorithms are applied to estimate the internal distribution of attenuation coefficients. In CT, this reconstruction is typically performed slice by slice using algorithms such as Filtered Backprojection, as stated before. The final output can be visualized as a stack of slices that can be seen as a volumetric image, with each voxel assigned an attenuation value expressed in Hounsfield Units.

2.1.5 Cone-Beam Computed Tomography

The difference between CT and CBCT lies in the geometry of the X-ray source and detector. In CBCT, the X-ray source and detector are not arranged in a fan-beam configuration, but rather in a cone-beam configuration. This can be seen in Figure 2.1. This means that the X-ray source emits a cone-shaped beam of X-rays, while the detector is said to be a flat panel detector. The cone-beam geometry allows for a larger field of view and the ability to capture a larger volume of data in a single rotation compared to the fan-beam geometry used in traditional CT. However, it is not sufficient to do exact 3D reconstruction.

Different reconstruction algorithms are used in CBCT. One of the most common algorithms is the Feldkamp-Davis-Kress (FDK) algorithm, which is a modification of the FBP algorithm. The FDK algorithm takes into account the cone-beam geometry.

However, due to the nature of the cone-beam geometry and insuffisant data for places that are far away from the midkdle slice, CBCT images usually are of a poorer quality than those obtained from conventional CT (Ghaznavi et al. 2025).

In fact, CBCT images are particularly susceptible to a range of artefacts that can compromise image quality. Some of these arise from discrepancies between the mathematical models used in reconstruction and the actual physical conditions, while others result from inherent limitations of the imaging process. Common artefacts include beam hardening, scatter, extinction artefacts, and motion artefacts.

Beam hardening occurs when low-energy X-rays are absorbed more than high-energy ones as the beam passes through matter. As a result, the average energy of the beam increases with depth, a phenomenon known as beam hardening. This effect is present in all materials but becomes especially significant in dense structures such as bone or metal implants. It leads to characteristic artefacts in the reconstructed image, including cupping artefacts, where the center of a uniform object appears artificially darker, and streak artefacts that often appear between dense regions. Scatter artefacts occur due to deflected X-ray photons that deviate from their primary trajectory and reach the detector, causing image degradation. Extinction artefacts are observed when highly attenuating materials completely block the X-ray beam, resulting in missing projection data. Motion artefacts emerge when patient movement occurs during acquisition, leading to blurring or double structures in the reconstruction. These artefacts are exacerbated in CBCT due to the use of a cone-shaped X-ray beam and flat panel detectors, which increase sensitivity to scatter and geometric distortions as well as the FDK algorithm that is an approximation of the cone-beam geometry (Schulze et al. 2011).

2.1.6 CBCT Artefacts Correction

Although CBCT systems are increasingly integrated into adaptive proton therapy workflows as explained in Chapter 1, the artefacts described above limit their direct use for quantitative dose calculations. To mitigate these limitations, correction strategies are applied at two levels (Ghaznavi et al. 2025).

Hardware Correction

Hardware-based corrections aim to reduce artefact generation at the source. Using physical improvements such as copper filters to reduce beam hardening as they make the beam more homogeneous. Other improvements include the use of bowtie filters, that attenuates the beam in a way that compensates for the shape of the patient and anti-scatter grids that reduce the scatter at the detector.

However, these measures alone are insufficient to fully restore image accuracy. Therefore, software-based corrections were developed to address further the remaining artefacts.

Software Correction

These include methods like Anatomical Image Correction (AIC) and Deformable Image Registration (DIR), which adjust planning CTs to match CBCTs and improve Hounsfield Unit (HU) accuracy for dose calculation. However these methods may be less reliable when large anatomical changes occur.

As introduced in Chapter 1, deep learning (DL) methods such as U-Net and GANs have recently been used to generate high-quality sCT images directly from CBCTs. These approaches are faster and can reduce artefacts more effectively, but their performance depends on careful training and validation (Thummerer, Zaffino, et al. 2020).

2.2 Deep Learning and Neural Networks

This section aims to introduce the fundamentals of Deep Learning (DL). This field of computer science extends traditional machine learning with larger models with more parameters, a class of differentiable architectures, and significantly larger datasets. This section will be mainly based on the course of INFO8010 - Deep Learning, given by Louppe (2024).

These explanations are essential, as they form the core of this master's thesis. However, for a deeper understanding of the concepts, it is recommended to refer to the course material (Louppe 2024). At a high level, Deep Learning can be viewed as a composition of modular building blocks that can be assembled in a structured manner to perform inference tasks.

The first section introduces neural networks, the fundamental building block and backbone of Deep Learning. This is followed by an overview of the various components that are going to be used in this thesis. These components include convolutional blocks, transposed convolutional blocks, pooling blocks, ReLU activation function, and skip connections.

2.2.1 Neural Networks

Neural Networks are a cornerstone of Deep Learning. Their origins trace back to the Threshold Logic Unit (TLU), introduced by McCulloch et al. (1943), which modeled a simple computational unit performing logical operations:

$$f(\mathbf{x}) = 1_{\{\sum_i w_i x_i + b \geq 0\}} \tag{2.4}$$

where x_i are Boolean inputs, w_i are weights, and b is a bias.

Rosenblatt (1958) extended this idea with the Perceptron, supporting real-valued inputs in a biomimetic approach, where w_i mimics synaptic weights and x_i represents firing rates it then goes through the sign function to output a binary value. The Perceptron can be represented as:

$$f(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b) \tag{2.5}$$

where $\mathbf{x}$ is the input vector, $\mathbf{w}$ is the weight vector, and b is the bias.

Moreover, let's introduce the sigmoid function, a non-linear activation function that allows the model to learn complex patterns by introducing non-linearity. :

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.6}$$

The Figure 2.2 shows the Equation 2.6. One can see that it resemble a soft heavy side function.

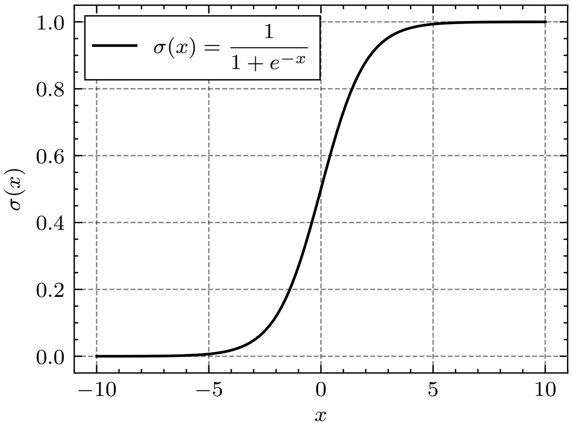
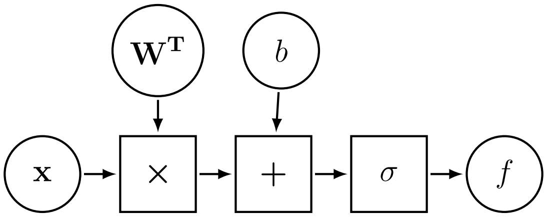


Figure 2.2. Examplification of the sigmoid function. One can see that it looks like a soft step function. It allows the model to learn complex patterns by introducing non-linearity.

These unit can be assembled in parallel to form the layer of a neural network where the sign function is replaced by the element-wise sigmoid function. This can be represented in a neat and easily understandable way as a computational graph:



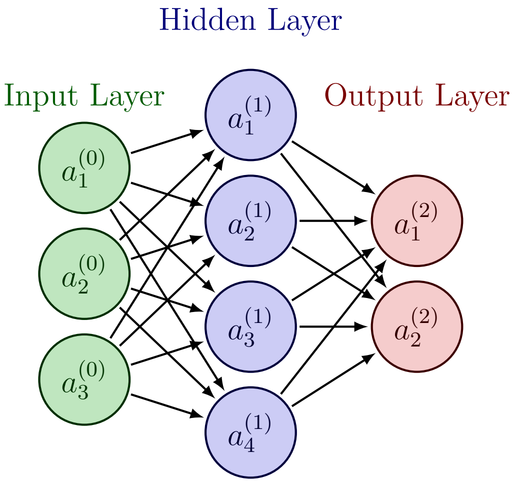


Figure 2.3. A simple neural network with 3 layers. The input layer, the hidden layer and the output layer.

Finally, by arranging these layers in series, one can create a Neural Network as can be seen in the Figure 2.3.

With the help of an optimization algorithm, such as stochastic gradient descent, the weights and biases of the network can be adjusted to minimize a loss function. This process is known as training the neural network.

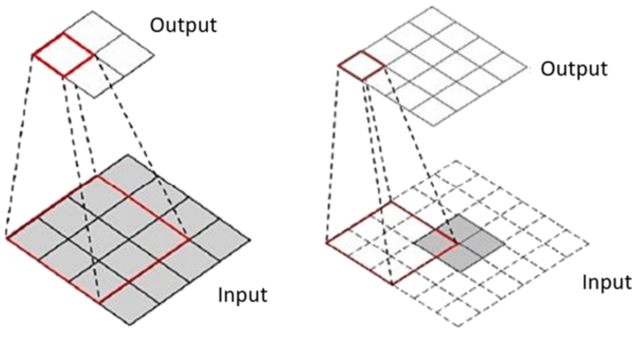
2.2.2 Important Building Blocks

Since then, researchers have developed a variety of architectures and techniques to improve the performance of these neural networks. This section will cover the building blocks that are relevant to this thesis and were developed with the idea of extending the use of these neural networks.

Convolutional Block & Transposed Convolutional Block

A convolutional block constitutes one of the fundamental elements in modern neural networks, particularly within image processing architectures. It consists of learnable kernels that are locally applied across an input tensor in a sliding-window fashion, preserving the spatial structure of the data. Formally, given an input tensor $x \in \mathbb{R}^{C\times H\times W}$ and a kernel $u \in \mathbb{R}^{C\times h\times w}$, the discrete convolution operation produces an output feature map $x * u \in \mathbb{R}^{H'\times W'}$. Each output value corresponds to a weighted sum of a localized input region, effectively enabling the model to extract local patterns such as edges or textures.

Beyond standard convolution, transposed convolution, also known as up-convolution, is commonly used for upsampling in decoder architectures. Instead of reducing spatial resolution like standard convolution, transposed convolution increases it by projecting each input location into a larger output space, using learnable weights arranged to simulate the reverse of a convolution operation. Both convolutional and transposed convolutional are illutrated in Figure 2.4.

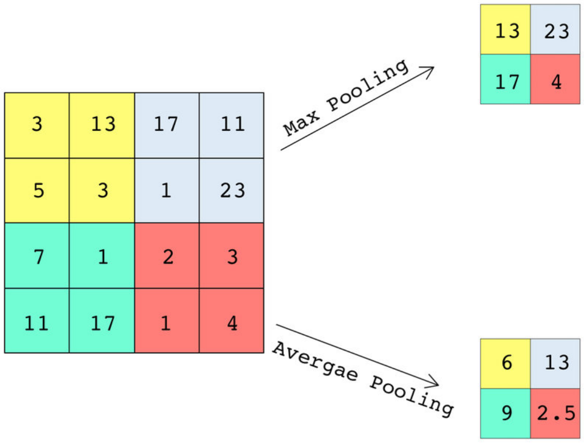


(a) Convolution (b) Deconvolution

Figure 2.4. Illustration of a convolutional block (a) and up-convolutional block (b). The convolutional block applies a kernel to the input tensor, while the up-convolutional block increases the spatial dimensions of the feature map. Figure from Pan et al. (2022).

Pooling Block

A pooling block is a downsampling operation commonly used in convolutional neural networks to reduce the spatial dimensions of feature maps. It processes each channel of the input tensor separately by applying a fixed-size sliding window over the spatial dimensions. Within each window, a simple, non-learnable function such as the maximum value (max pooling) or the average value (average pooling) is computed. This results in a tensor with reduced height and width, which improves computational efficiency and increases the receptive field of subsequent layers. Additionally, pooling introduces a degree of translation invariance, allowing the network to become less sensitive to small shifts or distortions in the input.



3	13	17	11
5	3	1	23
7	1	2	3
11	17	1	4

13	23
17	4

6	13
9	2.5

Figure 2.5. Illustration of a 2x2 max and average pooling. Figure from Aljaafari (2018).

ReLU Activation Function

The Rectified Linear Unit (ReLU) is a non-linear activation function widely used in deep neural networks in replacement of sigmoid.

Mathematically, the ReLU function is defined as:

$$\mathrm{ReLU}(x) = \max(0, x), \tag{2.7}$$

where the operation is applied element-wise to each component of the input tensor.

The main advantage of ReLU is that it avoids the vanishing gradient problem for positive inputs, allowing for more effective gradient propagation through deep networks. However, it may suffer from the so-called dying ReLU problem, where some neurons can become inactive if they output zero consistently, leading to zero gradients during training.

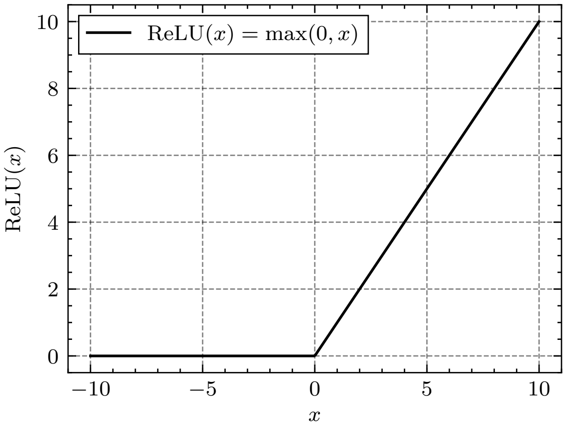


Figure 2.6. Illustration of the ReLU activation function. The function outputs zero for negative inputs and the input itself for positive inputs. This non-linearity allows the model to learn complex patterns.

Skip Connections

Skip connections, also referred to as residual connections, are architectural components that enable information to bypass one or more intermediate layers in a neural network. These connections were popularized by the ResNet architecture (K. He et al. 2016), which demonstrated their ability to mitigate optimization difficulties such as the vanishing gradient problem, particularly in very deep networks.

Formally, given an input x, a residual block computes an output of the form:

$$h(x) = x + f(x), \tag{2.8}$$

where $f(x)$ denotes a sequence of operations. The addition is performed element-wise, which requires that the dimensions of x and $f(x)$ be compatible.

The underlying intuition is that, rather than learning a direct mapping from x to $h(x)$, the network learns a residual transformation $f(x) = h(x) - x$. This residual formulation has been shown to facilitate convergence and improve training stability.

Moreover, skip connections provide alternative pathways for the gradient during backpropagation, allowing it to flow more efficiently through the network. This property is particularly beneficial in deep architectures, where gradient tend to vanish.

Chapter 3

Methodology

This chapter presents the methodological framework developed to address the central research question of this thesis: *Can state-of-the-art deep learning-based uncertainty quantification methods provide trustworthy and interpretable outputs when applied to sCT generation from CBCT images?* The overall objective is to assess the added value of uncertainty information, both epistemic and aleatoric, in improving the robustness, reliability, and clinical interpretability of sCT models.

To this end, the methodology is structured to address each of the aims outlined in the introduction. It begins by presenting the datasets used in this work. First, the SynthRad2023 CBCT to CT task 2 dataset is introduced alongside with the caveats it comes with and efforts to make it more usable. Then, a fully synthetic dataset, curated by IBA, is explained alongside the pipeline used to generate it. The data splitting strategy, a standard train-validation-test split being the same for both dataset, is introduced alongside the data augmentation pipeline used when training, validating and testing the model.

Next, two reference model architectures that will be compared are presented. One is commonly used in the literature, namely the U-Net (Ronneberger et al. 2015). The other, the U-Net++ (Zhou, Siddiquee, et al. 2019), is the SynthRad2023 challenge winner. They are chosen in order to establish a strong performance baseline. It is then shown how uncertainty estimation techniques are integrated into the modeling pipeline. Epistemic uncertainty is approximated using Monte Carlo Dropout (Gal et al. 2016), while aleatoric uncertainty is introduced via heteroscedastic Gaussian modeling (Kendall and Gal 2017).

The chapter then proceeds with a detailed description of the training setup, including the loss functions, the optimizer configuration, learning rate scheduling, and regularization strategies. Particular care is taken to maintain fair and consistent training conditions across all experiments. The hardware and software environment used for model development and training are also reported to support reproducibility.

Subsequently, evaluation protocols are outlined, covering three main aspects: the reconstruction accuracy, the perceptual quality, and the uncertainty estimation reliability. Metrics such as MAE, RMSE, SSIM, PSNR, the Pearson correlation coefficient (PCC), and the

Expected Calibration Error (ECE) are used to assess the effectiveness of the methods. These metrics, used in the field, allows to assess both the predictive performance and the quality of uncertainty estimation.

3.1 Data Overview and Preparation

This section presents the two datasets used in this thesis. The first is the SynthRad2023 challenge dataset, referred to as the real setting, which contains clinical CBCT and CT image pairs. Its main characteristics and the preprocessing steps applied by the original authors are first described. This is followed by additional preprocessing applied either by the IBA team or in the context of this work to address specific issues identified in the dataset, such as misalignments and scanner inconsistencies. These corrections were necessary to improve data quality and reduce sources of noise that could negatively affect model training, especially for uncertainty estimation tasks.

The second dataset, referred to as the synthetic dataset, is introduced afterward. It was specifically created to validate the proposed approach under a fully controlled environment. This allows for a clearer interpretation of the results and provides a complementary view to the experiments conducted on real clinical data.

Although the results will be presented starting with the synthetic dataset, this section begins with the real dataset. This ordering highlights the challenges encountered in real clinical data and motivates the need for a synthetic counterpart to ensure the robustness and interpretability of the proposed uncertainty estimation methods.

Moreover, these datasets serve complementary objectives in the overall evaluation of the proposed approach. The simulated dataset, by offering a fully controlled environment, enables a more direct and interpretable analysis of uncertainty estimation quality. In particular, it facilitates comparisons with existing literature to evaluate calibration, reliability, and disentanglement of aleatoric and epistemic uncertainties. On the other hand, the real dataset provides a practical benchmark to assess whether the theoretical principles underlying uncertainty quantification methods hold when applied to complex, imperfect clinical data. It plays a key role in evaluating the applicability and robustness of these methods in realistic treatment planning conditions.

3.1.1 Datasets and Data Processing

SynthRad2023

The first dataset used in this thesis was originally created for the SynthRad2023 Grand Challenge (Thummerer, Bijl, Galapon, et al. 2023). This large-scale, multi-center initiative was designed to benchmark synthetic CT generation methods for adaptive radiotherapy. It contains imaging data from a total of 1080 patients, acquired across three different Dutch University Medical Centers. For clarity, these institutions will be referred to as *Center A*, *Center B*, and *Center C* throughout this work.

The challenge is organized into two main tasks. The first, referred to as Task 1, focuses on sCT generation from MRI images. As this task is not relevant to the objectives of this thesis, it will not be discussed further.

The second task, named Task 2, is the one used in this work and involves the generation of sCT from CBCT scans. This task is further divided into two anatomical sites: the pelvis and the brain. Each anatomical site includes imaging data from 270 patients, resulting in a total of 540 patients for Task 2.

The data in Task 2 were collected from the three aforementioned medical centers, each using its own scanner configurations for CT and CBCT acquisition. These configurations are summarized in Table 3.1.

Each of the three centers contributed to the dataset with 60 patients per subset. This makes up a total of 180 patients per anatomical site, uniformly distributed across the centers.

	CT Scanner(s)	**CBCT Scanner(s)**
Center A	Philips Brilliance Big Bore or Siemens Biograph20 PET-CT	Elekta XVI
Center B	Siemens SOMATOM Definition AS	Elekta XVI or IBA Proteus+
Center C	Siemens Avanto Fit 1.5T or Siemens MAGNETOM Vida Fit 3.0T	Elekta XVI

Table 3.1. Scanner configurations for CT and CBCT imaging across the three different centers (*Center A*, *Center B*, and *Center C*) available in the SynthRad2023 dataset.

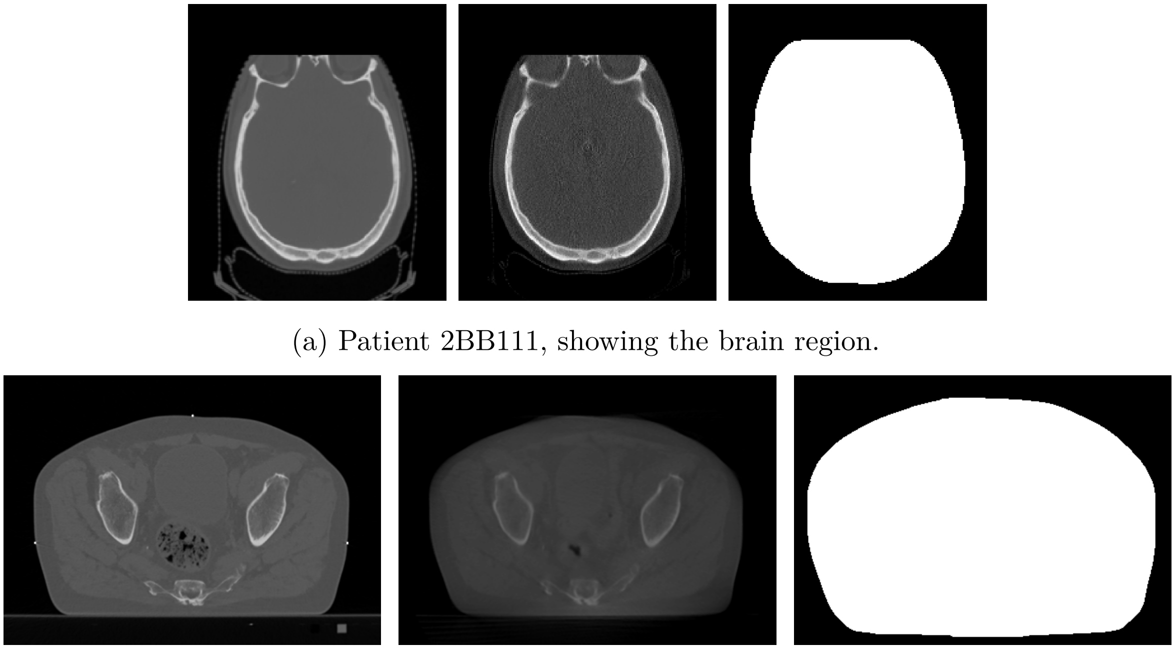
Each patient in the dataset has a paired CBCT and CT scan provided in NIfTI format alongside a binary mask delineating the patient outline. The mask was dilated to include surrounding margins. The CBCT scan serves as the model input, while the CT scan is used as the ground truth for training and evaluation purposes.

All relevant information regarding the dataset is available in the original publication (Thummerer, Bijl, Galapon Jr, et al. 2023). The dataset itself is publicly available on Zenodo (Thummerer, Bijl, Galapon, et al. 2023) under a Creative Commons Attribution license. It was originally divided by the authors into training, validation, and testing subsets. However, as the validation and test set are not publicly released, this thesis considers the combination of the training and validation sets as the full available dataset.

Figure 3.1 shows some example of corresponding CBCT and CT slices from patient *2BB111* and *2PB048* in the dataset, alongside with the binary mask. The binary mask is also shown, axially, for reference.

SynthRad2023 Preprocessing

The authors of the SynthRad2023 dataset applied several preprocessing steps, which are described in this section. These initial steps were further extended with additional preprocessing operations introduced by IBA and during the course of this work. The aim of these was to



(a) Patient 2BB111, showing the brain region.

(b) Patient 2PB048, showing the pelvis region.

Figure 3.1. Examples of CBCT-CT pairs present in the SynthRad2023. Visualized are the, CT (left), CBCT (center) and binary mask (right) for each patient.

address specific limitations in the raw data and to improve the quality and consistency of the images for training and evaluation purposes.

SynthRad2023 Authors Prepocessing

The SynthRAD2023 dataset underwent a preprocessing pipeline developed by it's authors to ensure data uniformity and facilitate model training. Initially, all CT and CBCT scans were converted from DICOM to compressed NIfTI format, optimizing storage and compatibility with deep learning frameworks. To standardize spatial resolution, images were resampled to a voxel size of 1x1x1 mm^3 for brain scans and 1x1x2.5 mm^3 for pelvic scans. Rigid registration was then performed to align CBCT images with their corresponding CT scans, ensuring anatomical correspondence between modalities. It was performed using Elastix (Klein et al. 2010) with the parameters specified in Thummerer, Bijl, Galapon Jr, et al. (2023). For brain images, additional anonymization was conducted by defacing, which involved removing voxels anterior and inferior to the eyes to protect patient identity. A binary mask delineating the patient outline was generated using thresholding and morphological operations, which was then dilated to include surrounding air margins. Finally, all images and masks were cropped to the bounding box of the dilated mask with an added margin of 20 voxels, reducing file size and focusing on the region of interest (Thummerer, Bijl, Galapon Jr, et al. 2023).

Detailed demographic information such as patient age and gender is not directly accessible from the released dataset files as they are in NIfTI format and not DICOM. However, summary statistics have been reported in the original SynthRad2023 publication (Thummerer, Bijl, Galapon Jr, et al. 2023). According to the authors, the age of patients, mostly adults, in the dataset ranges from 3 to 93 years, with a mean of 65 years. The overall gender distribution consists of 57.4% male and 42.6% female subjects for the brain subset. Furthermore, the pelvis subset is notably skewed towards male patients, due to the inclusion of prostate cancer cases, with 81.9% of patients being male and only 18.1% being female.

Moreover, this dataset present some additional challenges due to its real nature. These challenge introduces noise in the training signal, reducing the alignment between the loss function and the true objective. This renders training more difficult and the algorithms

One of the challenges is visible in Table 3.2 that provides a summary of the HU values for CT and CBCT images across the brain and pelvis regions in centers A, B, and C. The values from center A seem to be shifted by a bias of 1000 compared to the other centers. It seems like an error slipped through the preprocessing pipeline of the authors of the dataset. Hence, the values of the CBCT images from center A will be corrected by subtracting 1000 to all HU values.

Center	Region	Modality	Mean	Std	Min	Q1	Median	Q3	Max
A	brain	CT	-686.8	514.2	-1024.0	-1000.0	-995.0	-270.0	3000.0
A	brain	CBCT	420.5	544.7	0.0	0.0	216.0	641.0	3000.0
B	brain	CT	-517.5	608.2	-1024.0	-999.0	-973.0	32.0	3000.0
B	brain	CBCT	-658.1	518.6	-1000.0	-1000.0	-1000.0	-342.4	2000.0
C	brain	CT	-690.4	520.8	-1024.0	-1000.0	-1000.0	-110.0	3000.0
C	brain	CBCT	-711.5	475.1	-1024.0	-1000.0	-1000.0	-282.0	3000.0
A	pelvis	CT	-444.0	489.3	-1024.0	-990.0	-114.0	-15.0	3000.0
A	pelvis	CBCT	312.9	323.7	0.0	0.0	141.6	631.5	3000.0
B	pelvis	CT	-421.9	495.7	-1562.0	-990.0	-103.0	13.0	3000.0
B	pelvis	CBCT	-626.7	353.3	-1024.0	-1022.0	-473.0	-300.0	3000.0
C	pelvis	CT	-432.5	507.6	-1024.0	-1000.0	-109.0	12.0	3000.0
C	pelvis	CBCT	-514.7	439.6	-1024.0	-1000.0	-272.0	-150.0	3000.0

Table 3.2. Summary statistics of HU values for CT and CBCT images across brain and pelvis regions in centers A, B, and C. The table illustrates clear differences in intensity distributions between CT and CBCT for the same region in between centers. Notably, CBCT ranges from center A seem to be shifted by a bias of 1000, which is not the case for the other centers and will need correction.

Another challenge, that induces substantial noise and problems is the misalignment of patients in between CBCT and CT. This means that even if the anatomy is correponsing in between both modalities, the pixel in these do not correspond one to one. To paliate this issue and provided with the dataset files, an Excel file provides information about scanner parameters as well as notes for each patient. These notes are particularly useful for identifying and excluding these patients with incorrect rigid registration or the presence of implants. For example the patient 2PA003 has an implant in the hip that could cause some artefacts as can be seen in Figure 3.2. While the deep learning model should be able to learn to ignore these implants it cannot be said in the case of bad registration. For example, the patient 2PA009 is said to be not perfectly rigidly registered as well as the patient 2PA010. This is important to note, because as can be seen in Figure 3.3, the absolute difference axially between the CT and CBCT is notable as the patient is not perfectly aligned in between the two scans. A model trained on these images solely would not be able to learn the true correspondence between the two scans. This plays a big role in the noise of the dataset.

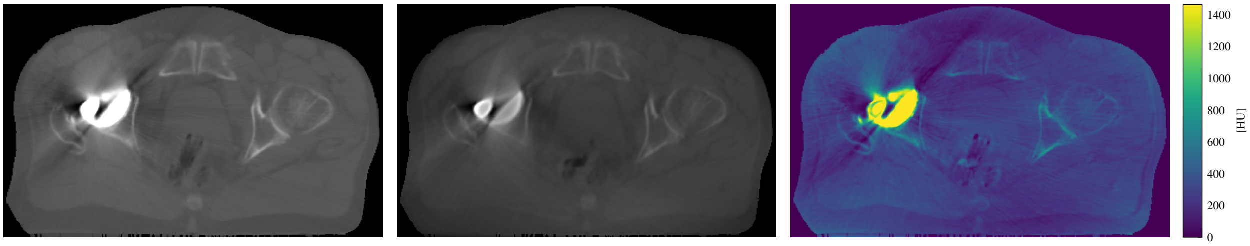
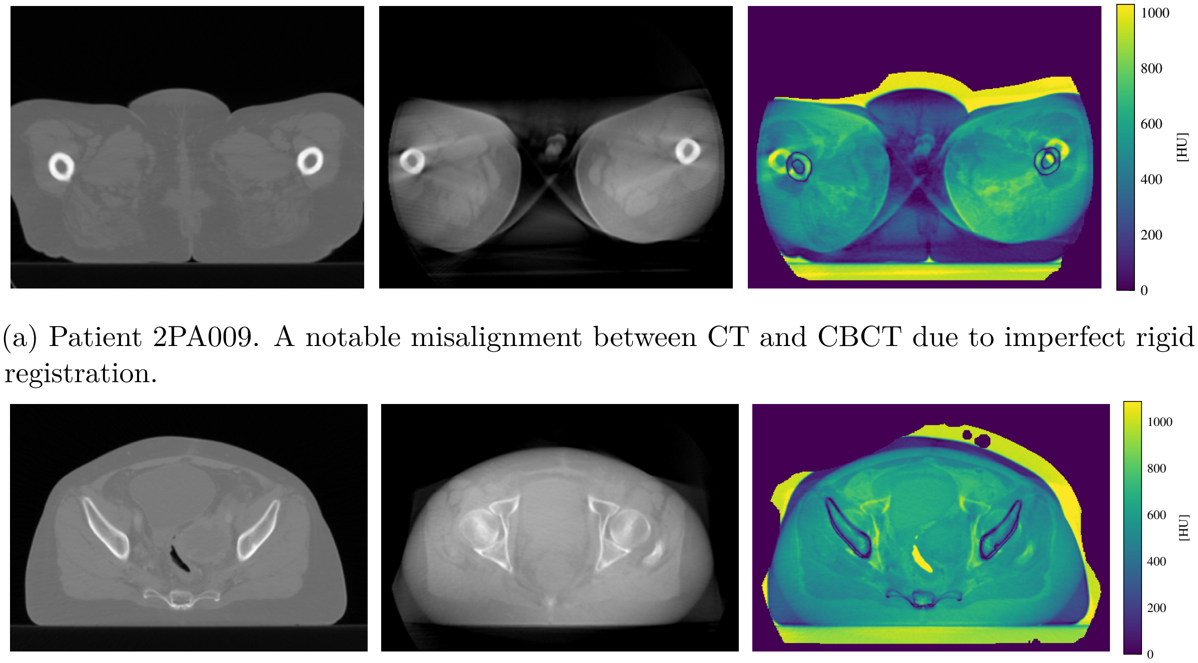


Figure 3.2. Patient 2PA003. The implant in the hip is visible in the CT (left) and CBCT (center) images. The absolute difference (right) shows a large discrepancy at the implant location.



(a) Patient 2PA009. A notable misalignment between CT and CBCT due to imperfect rigid registration.

(b) Patient 2PA010 . Misalignment is visible across the pelvis, particularly in bony structures.

Figure 3.3. Examples of CBCT-CT pairs with registration issues. Visualized are the CT (left), CBCT (center), and absolute difference computed within the mask (right) for each patient. The window levels are ajusted for CT and CBCT to see easily the discrepancies. Misregistration leads to artificially high intensity difference, where a model will not be able to learn the true correspondence between the two scans. Moreover, as can be seen in the absolute difference, the mask is partially wrong and will introduce unnecessary error.

Another notable source of variability in the dataset arises from anatomical changes between the CT and CBCT acquisitions. This is the case even if the patient is correctly registred, particularly in relation to cavity presence and soft tissue configuration. According to the authors of the dataset, the time interval between the planning CT and the follow-up CBCT scan can be as up to two months. During this period, clinically relevant changes may occur. Such change could be as the filling or emptying of rectal or bladder cavities, shifts in soft tissue distribution, or differences in patient positioning. These changes introduce structural discrepancies that are not corresponding to imaging noise or reconstruction artefacts, but to true anatomical variation.

This temporal separation between the "input" (CBCT) and the "ground truth" (CT) significantly complicates the regression tasks. For example, the presence of a rectal cavity in the CT that is absent in the corresponding CBCT scan leads to an apparent soft tissue mismatch, which may be incorrectly interpreted by the loss function as a model error. Such cases penalize the model for failing to predict tissue or cavities structures that, in fact, no longer exist or only appear at inference time.

SynthRad2023 Further Preprocessing

The goal of this preprocessing was to reduce the noise in the dataset in order to assess the impact on regression and to improve the general quality of the data to make it more suitable for the task at hand. The following preprocessing steps were applied to the SynthRad2023 dataset by IBA and are not part of the original contributions of this thesis. A brief overview of the preprocessing pipeline is provided below as well as in Figure 3.4. The first step in the pipeline rigidly registered the CT images to the CBCT images using their internal software that optimizes mutual information. Then, to account for the deformations that are not able to be captured by the rigid registration, a second non-rigid deformation was performed. Namely, Deformable registration using morphon algorithm (Janssens et al. 2011) which enables non-rigid warping of the CT image into the CBCT space while preserving anatomical coherence. In parallel, the pipeline then proceeds with CBCT correction stage. This step involve intensity standardization by matching HU values between CBCT and CT as well as filtering low frequency artefacts with a Gaussian filter, producing the corrected CBCT. The combination of this corrected CBCT with the deformed CT image yields the final virtual CT (vCT) image. The mask of the patient was also recomputed, effectively leading to a new mask that is more precise and better aligned. More details about this pipeline can be found in Veiga et al. (2016).

The goal of this preprocessing pipeline is to make the CT and CBCT images as similar as possible. By correcting both the overall position of the patient and the local differences in anatomy the result is a virtual CT (vCT) that hopefully closely matches the CBCT. This vCT is used as a cleaner and more realistic reference image. With this improved alignment and intensity correction, the aim is that any difference between the model's prediction and the ground truth is more likely to come from the model itself, and not from noise or errors in the data. As can be seen in Figure 3.5, the difference between the CBCT and the vCT is much smaller as the one with the CT from Figure 3.3. However, this pipeline is not perfect, and the dataset still present some issue. For example, *2PA010* still showed a significant misalignment between the two images as can be seen in Figure 3.5. However, upon further investigation, no other patient showed such a large misalignment and therefore this example was simply

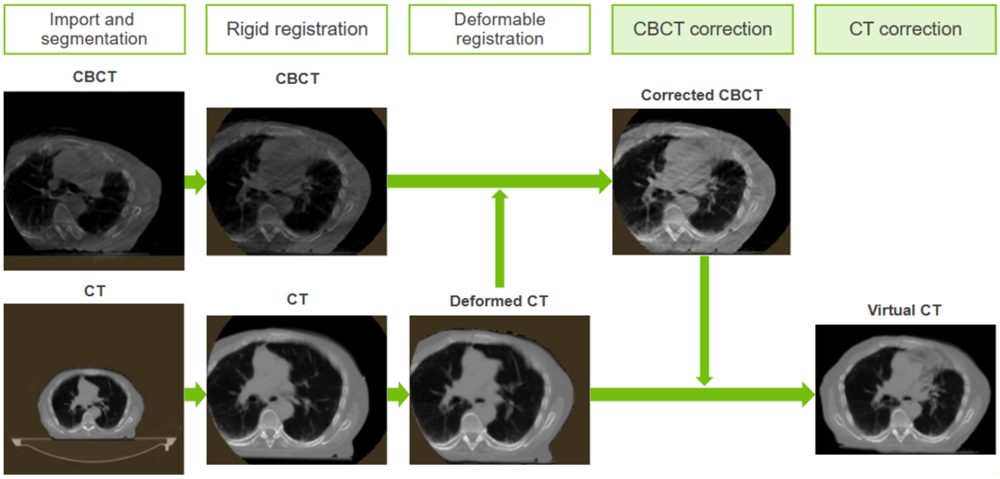


Figure 3.4. Visualisation of the pipeline to create a Virtual CT from a CBCT and a CT image. This pipeline was applied to the SynthRad2023 by IBA.

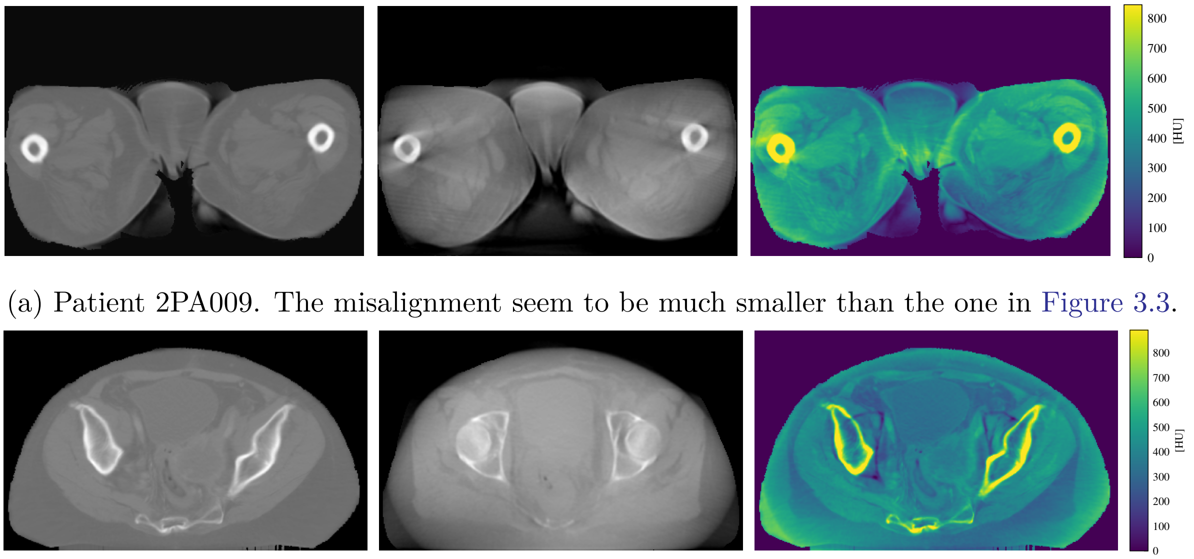
removed from the dataset.

The final SynthRad2023 dataset therefore consists of 179 patients. The input data correspond to the original CBCT images from the dataset, while the target data are preprocessed CT scans referred to as virtual CT (vCT). In the remainder of this thesis, the terms vCT and CT will be used interchangeably unless otherwise specified when talking about the real setting.

Synthetic Dataset

The second dataset used in this work is named the Simulated Dataset. This is because it is generated by IBA using their internal pipeline as can be seen in Figure 3.8. These simulated data are therefore not generated as part of this work but were provided by IBA for the purposes of this thesis. The goal of this dataset is to produce CBCT images that resemble the real IBA CBCT scans as closely as possible. To achieve this, a CT volume is first selected and subjected to a forward projection process. Forward projection simulates the physical acquisition of X-ray projections by modeling how X-rays would traverse the CT volume and be attenuated by the different tissue densities, effectively emulating the measurements that a real CBCT scanner would capture. A range of various noise and physical effects is applied to these projections to make them match CBCT projections. These synthetic projections are then reconstructed into a volumetric image through backward projection, generating the simulated CBCT. Because the images are generated from the same source data, they are perfectly registered by design, thereby eliminating any misregistration issues as can be seen in Figure 3.7. Furthermore, since the CT and CBCT are derived simultaneously through simulation, no anatomical differences exist between them, unlike in the SynthRad2023 dataset were difference in anatomy were present due to the time between scans.

The CT used to generate this dataset came from the SynthRad2023, and the simulated CBCT as well as the CT scan were resampled to 1x1x2.5mm^2 to enable comparison with the original dataset. The generated dataset also came in form of a CT scan, a CBCT scan and a mask



(a) Patient 2PA009. The misalignment seem to be much smaller than the one in Figure 3.3.

(b) Patient 2PA010. Misalignment is still visible.

Figure 3.5. Examples of CBCT-vCT pairs were the original dataset presented some registration issues. Visualized are the virtual CT (left), CBCT (center), and absolute difference computed within the mask (right) for each patient. The window levels are ajusted for CT and CBCT to see easily the discrepancies. Some misregistration is still visible for some patients, but the overall difference is much smaller than the one in Figure 3.3. The mask also seems to be much more precise, as can be seen in the absolute difference.

visualized in Figure 3.6. These are available in mhd format, a widely used format in medical imaging.

The generation of this dataset required careful tuning of numerous simulator hyperparameters. For simplicity, certain elements such as the couch position, which needed to be manually specified per patient, were omitted from the simulation pipeline, and thus the resulting scans do not include it. For each CT volume, simulated CBCT projections were provided by IBA at varying levels of scatter (30%, 40%, and 50%) and tube currents (10, 32, 80, and 125 mA), all acquired at a fixed tube voltage of 120 kV and an exposure time of 12 ms.

To introduce variability representative of real-world acquisition conditions, each volume was assigned simulated scatter levels of 30%, 40%, or 50%. This controlled manipulation of scatter conditions enhances the dataset's variability by exposing the model to different noise profiles. The effects of scatter levels on projection profiles are illustrated in Figure 3.9. This plot shows a CT slice at fixed current and show what the variation of scatter does. The line plot alongside helps in understanding that the scatter is the highest at the center of the slice.

Simulated CBCT scans were generated at tube current levels of 10, 32, 80, and 125 mA, all at a constant tube voltage of 120 kV and an exposure time of 12 ms. While most current levels were retained, the scans simulated at 10 mA were excluded due to excessive noise and limited clinical relevance. An overview of the current settings and their impact on acquisition is provided in Figure 3.10. One can see that the 10mA variant presents fat too much noise and is deemed to be irrelevant.

The dataset finally represented 130 patients, with each of them having 9 used CBCT. this made up for 1170 different CBCT scan to train the models with corresponding to their respective 130 CT scans as the CBCT with lower 10 mA current were not used as they were deemed too noisy.

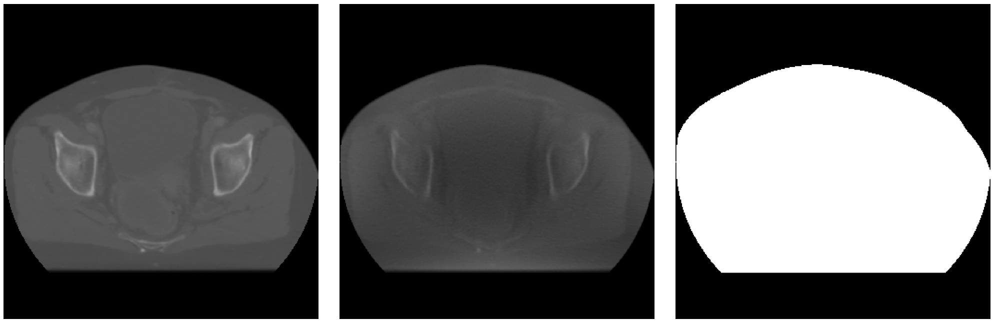


Figure 3.6. Patient 2PA010. Example of CBCT-CT pair from the simulated dataset. Visualized are the, CT (left), CBCT (center) and binary mask (right).

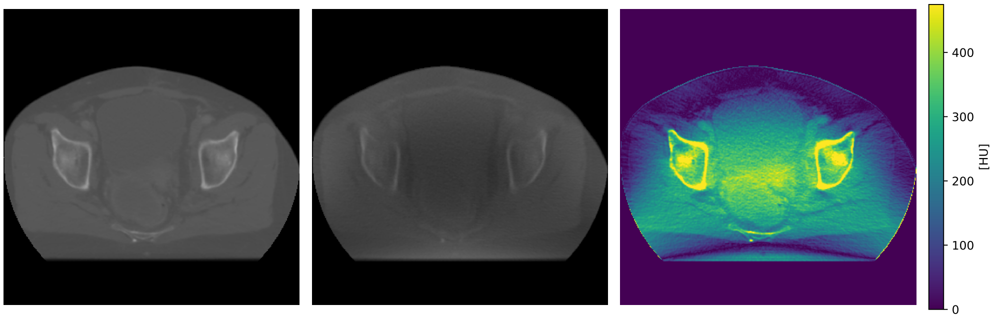


Figure 3.7. Patient 2PA010. One can see that there is no misregistration in the simulated dataset. In contrast, this specific patient had to be removed from the SynthRad2023 dataset.

3.1.2 Data Split and Data Augmentation

The dataset is divided into training, validation, and test sets using a stratified patient-level split, ensuring that slices from the same patient do not appear in multiple subsets, thus avoiding data leakage. Each split was also balanced to proportionally represent the different medical centers included in the dataset, ensuring institutional diversity across the sets. To ensure reproducibility, CSV files listing patient IDs for each split were created and used consistently throughout the experiments although this will be described in more detail in subsection 3.6.3.

The training set consisted of 70% of the patients, while the validation and test sets each contained 15% of the patients. This split was designed to provide a robust training set while maintaining a sufficient number of samples for validation and testing.

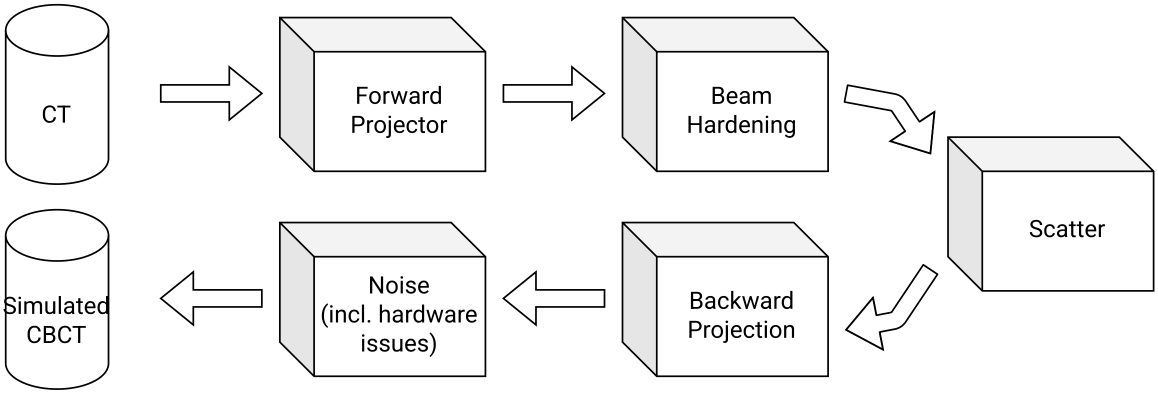


Figure 3.8. Simplified simulation of the IBA Simulated CBCT Pipeline. The chosen CT is first forward projected in order to add the chosen physical noise and artefacts that are proper to IBA CBCT's. It is then backward projected to get the simulated CBCT. One of the main challenge is ensuring that the resulting simulated CBCT can be considered as a sample of the real IBA CBCT's.

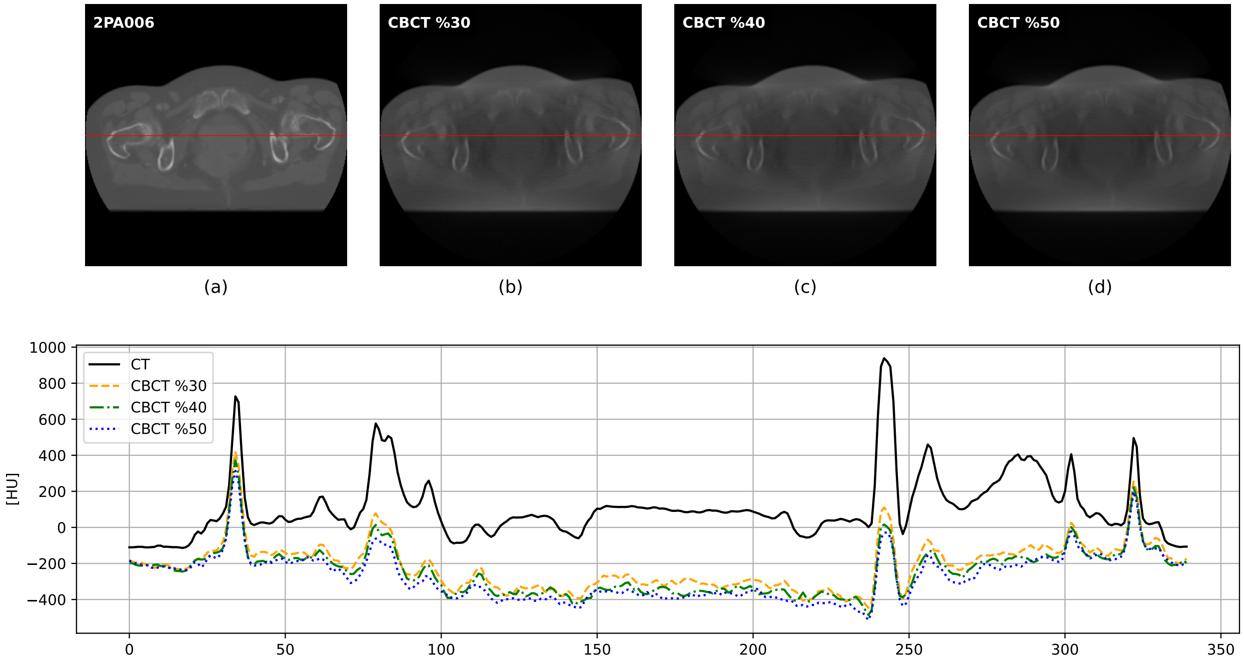


Figure 3.9. HU profiles extracted from (a) CT of patient 2PA006 for various simulated scatter levels: (b) CBCT with 30% residual scatter, (c) CBCT with 40% residual scatter, and (d) CBCT with 50% residual scatter. The profiles are compared to the corresponding CT reference profile. The increasing scatter fraction introduces greater degradation in the HU values, notably in the center. All acquisitions were performed using a simulated tube voltage of 120kV, an exposure time of 12ms and a current of 125mA.

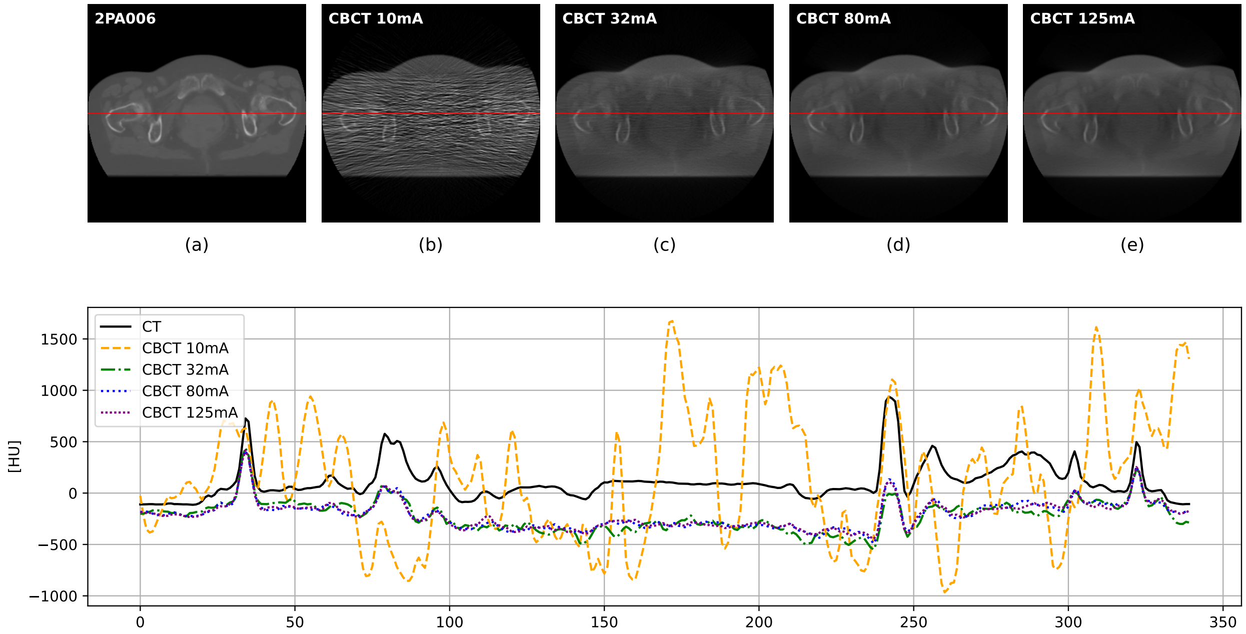


Figure 3.10. HU profiles extracted from (a) CT of patient 2PA006 for various simulated tube currents: (b) 10mA, (c) 32mA, (d) 80mA, and (e) 125mA. Higher tube currents lead to improved signal fidelity and reduced noise, with the 125mA profile exhibiting the closest match to the CT reference. All acquisitions were performed using a simulated tube voltage of 120kV, an exposure time of 12ms and a scatter of 30%

After splitting and preprocessing, the dataset was organized into a series of 2D slices, representing the axial view of each patient. This approach effectively treated the data as individual 2D axial slices rather than entire volumetric patient scans. This is necessary for the training of the models, as a 3D model would require a significant amount of memory and computational resources, which were not available for this work. Beyond practical considerations, the use of 2D slice-based approaches is well established in the literature, especially for tasks involving large datasets. Although this approach discards some of the spatial continuity present in the volumetric data, it offers a good trade-off between model complexity and computational feasibility, while still allowing for meaningful spatial context within each slice (Zhou, Sodha, et al. 2019). Furthermore, the model will follow a 2.5D approach, where each prediction uses information from adjacent slices, thus preserving some of the 3D context. The full modeling strategy is detailed in Section 3.2.

On top of that and for the training phase, several data augmentation techniques were applied. These included random vertical and horizontal flips, random affine transformations with slight rotations, translations, and scaling, as well as a center crop to ensure consistency in image dimensions. If the image was smaller than the target size, it was padded with the value of air in HU. These transformations were designed to simulate variations in the dataset while preserving the anatomical relevance of the images. The Figure 3.11 shows some examples of the data augmentation applied during the training phase.

Concerning the validation and the test phase, only the center crop was applied, ensuring that all images were resized to be compatible with the model's architecture while preserving sufficient information by avoiding excessive cropping.

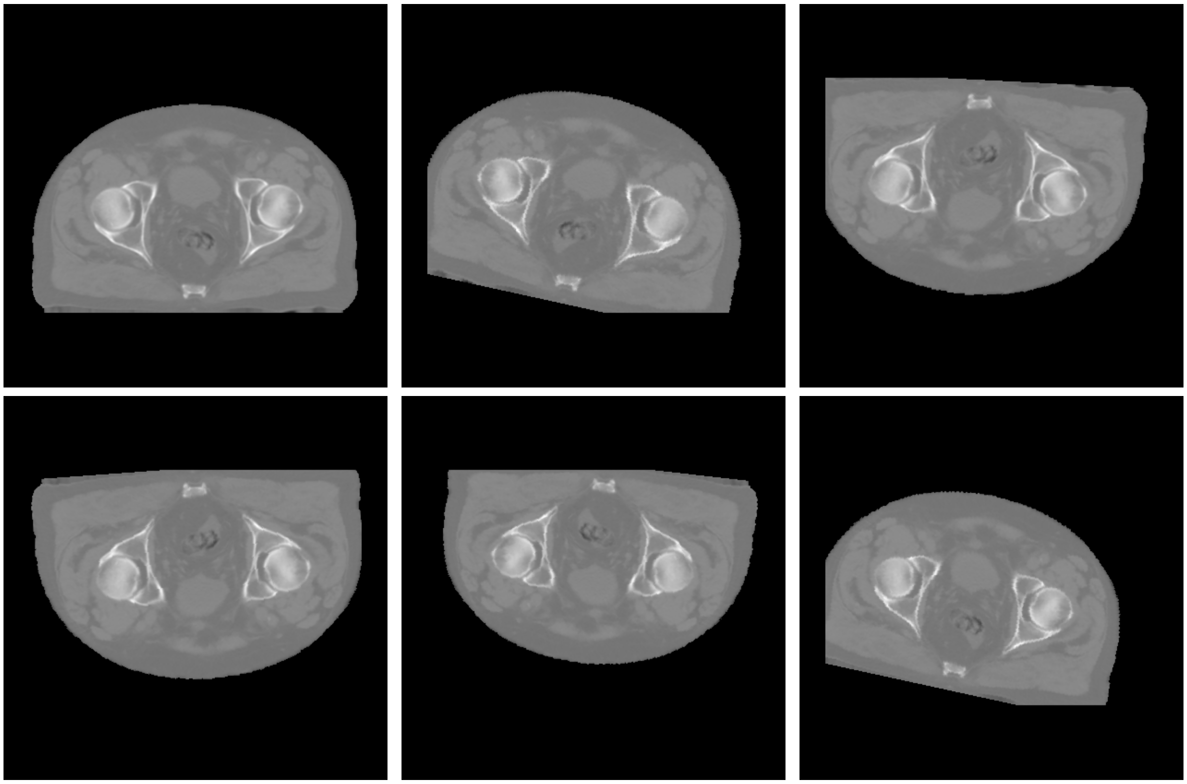


Figure 3.11. Examples of data augmentation. The original image, as used in the validation or test phase, is shown on the top left. Other frames show some examples of data augmentation with flips, rotations and translations.

Finally, each of the slices was clipped between -1000 and 3000, to ensure that no outliers or bug would be present during training or evaluation. As a final step, the images were normalized between -1 and 1, using the minimax scaling method. This was done because as will be seen in Section 3.2, the architecture uses a tanh activation function at the output layer, that maps the output between -1 and 1.

3.2 Model Architecture and Baselines

This study explores two convolutional neural network architectures for generating synthetic CT images from CBCT scans: the U-Net and its advanced variant, U-Net++. The aim of this study is not to achieve peak predictive performance, but rather to evaluate the relevance and applicability of uncertainty quantification methods in this context. For this reason, well-established models were deliberately chosen, as they should have enough expressivity for the regression problem at hand while being well understood.

To guide the choice of model for the uncertainty estimation framework, a preliminary comparison will be conducted using the real CBCT dataset. This choice is motivated by the need to identify a model that offers a good trade-off between performance and computational efficiency. As the simulated dataset is a noise-free and perfectly registered version of the real data, it is expected that the insights gained from the real-world setting will transfer effectively to the idealized one. The performance of both models will therefore be assessed in terms of reconstruction accuracy, and the model demonstrating sufficient performance with reduced computational burden will be retained for the remainder of this work.

This section will also present the two baselines used for comparison when computing the metrics in order to have a better understanding of the results. Specifically, the first is the Water baseline and the second is the raw CBCT.

3.2.1 Investigated Architectures

This section will present the two architectures that are explored in this work, namely the UNet and the UNet++.

U-Net

The U-Net architecture, originally introduced by Ronneberger et al. (2015) for biomedical image segmentation, is a convolutional neural network that adopts an encoder-decoder structure with skip connections. It was designed to perform well even when trained on limited datasets, a characteristic that aligns well with medical imaging tasks.

The encoder path consists of a sequence of convolutions, each followed by a rectified linear unit (ReLU) activation and a downsampling operation, usually max-pooling. This contracting path allows the network to capture increasingly abstract representations by reducing spatial dimensions while increasing the number of feature maps.

The decoder path symmetrically mirrors the encoder. It performs upsampling via transposed convolutions, and progressively reconstructs the spatial resolution of the output. Each decoding step incorporates skip connections from the corresponding encoder layer. The goal of the skip connections is to concatenate the high-resolution features comming from the encoder with the upsampled features in the decoder. This is done with the goal of preserving spatial information that may be lost during downsampling.

Figure 3.12 shows a typical U-Net architecture. This Figure was taken from the original paper by Ronneberger et al. (2015) and illustrates the encoder-decoder structure with skip connections.

This architecture has proven effectiveness in various medical imaging tasks, such as the one of Ronneberger et al. (2015). It was also leveraged by Thummerer, Zaffino, et al. (2020) for generating synthetic CT images from CBCT scans using the SynthRad2023 dataset. Hence, this type of architecture was selected as it is well-established in the field and was previously used in this setting.

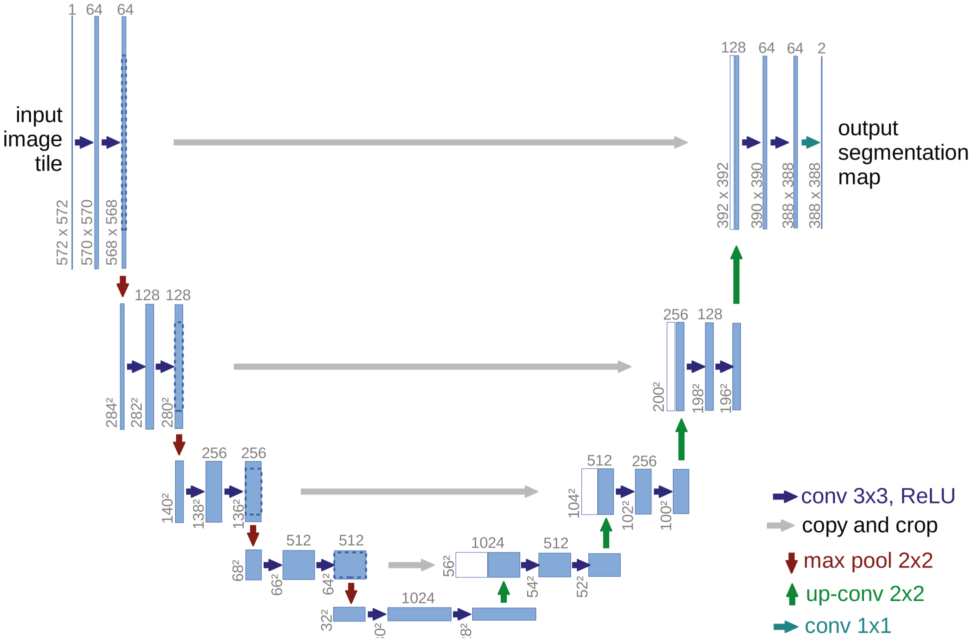


Figure 3.12. U-Net architecture. Taken from Ronneberger et al. (2015). The architecture follows a symmetric encoder-decoder structure with skip connections. Arrows indicate the operations applied at each step: convolution (blue), pooling (red), up-convolution (green) and concatenation (white)

U-Net++

UNet++, proposed by Zhou, Siddiquee, et al. (2019), is a redesigned variant of the original U-Net architecture that addresses two important structural limitations: the restrictive nature of standard skip connections and the fixed-depth design of the network. In UNet++, skip pathways are no longer simple shortcuts between encoder and decoder layers of equal depth. Instead, each skip connection is replaced by a series of nested convolutional blocks that gradually refine encoder features before fusing them with the decoder. This results in a denser, more expressive network topology capable of capturing and combining multiscale contextual information.

This architecture was chosen to be Investigated as it happens that the SynthRad2023 challenge winners used it in their winning solution (Huijben et al. 2024; Zhang et al. 2023). Hence, it is of interest to compare the performance of this architecture with the simpler U-Net architecture.

Segmentation Models PyTorch & Actual Implementation

Practically, both of these architectures were implemented using the `Segmentation Models PyTorch` library (Yakubovskiy 2020), which provides a wide range of encoder-decoder architectures originally developed for image segmentation tasks. Although the original use case of the library is segmentation, the encoder-decoder structure of these models is adaptable to regression tasks since both require pixel-level predictions. The advantages of the pre-existing

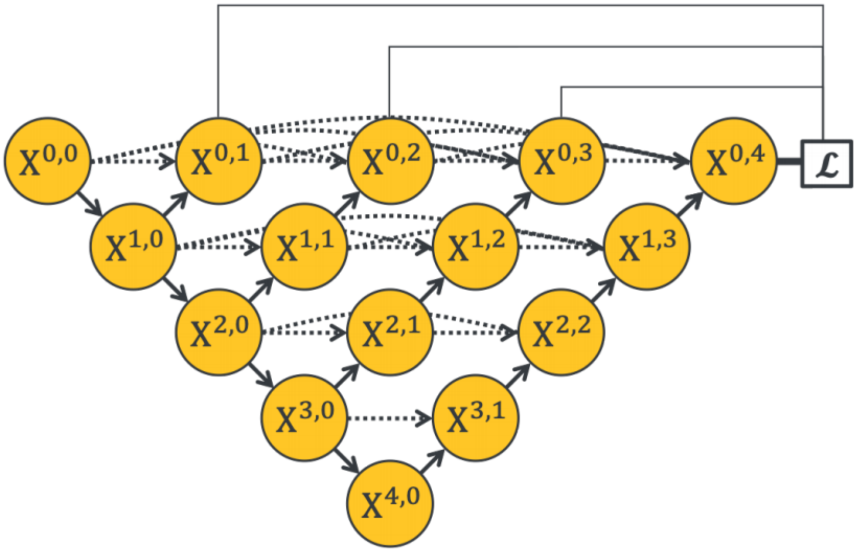


Figure 3.13. U-Net++ architecture. The figure illustrates the nested skip pathways between encoder and decoder stages, where each node $X^{i,j}$ represents a convolutional block. Taken from Zhou, Siddiquee, et al. (2019).

implementations include modularity, compatibility with pretrained backbones, and extensive community support, making them a practical choice for this work. Moreover, monkey patching allows for efficient code modification as this library is open-source.

The hyperparameters of the models were set to be in accordance with the SynthRad2023 challenge winners. Additional hyperparameters were determined with the guidelines proposed in the work of Fabibombo (2023), which explored sCT generation.

All of the hyperparameters used in this study are summarized in Table 3.3. Both the U-Net and U-Net++ architectures are configured as 2.5D models, meaning they process multiple adjacent slices simultaneously to generate a synthetic CT image for the central slice. Specifically, the input consists of the CBCT slice that corresponds to the CT slice and two neighboring slices on either side, resulting in an input volume of five slices as shown in Figure 3.14. Accordingly, the `in_channels` parameter is set to 5. This configuration follows the recommendations of the SynthRad2023 challenge winners, who found this input depth to be optimal for the synthetic CT generation task.

The encoder backbone used for both models is a Res2Net-based architecture available through the `timm` library (Timm 2020), and both models are initialized with ImageNet-pretrained weights. This initialization provides the encoder with a strong set of low-level features as edges learned from the ImageNet dataset, which is then fine-tuned to better adapt to synthetic CT generation. The encoder depth is set to 5, corresponding to five stages of down-sampling, which is the value used by Fabibombo (2023).

The decoder is configured with channel dimensions of (256, 128, 64, 32, 16) across its five stages. Batch normalization is disabled in the decoder path, as well as no attention mechanism is applied in either model, consistent with the setup adopted in prior work. The number of

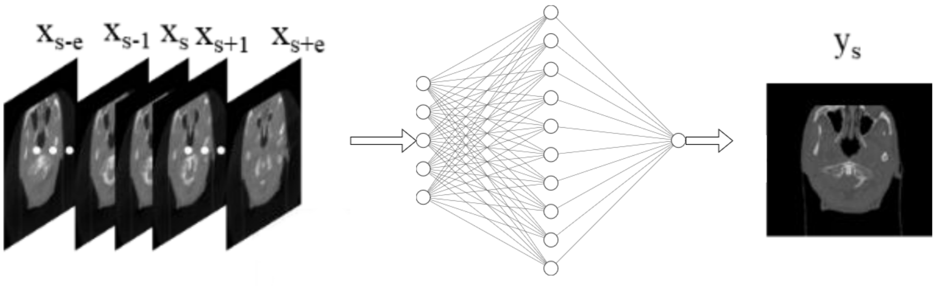


Figure 3.14. Illustration of the 2.5D model input strategy used in this work. The central slice X_s is predicted using a stack of adjacent slices $(X_{s-e}, \dots, X_{s+e})$ as input, preserving local 3D context. Adapted from Zhang et al. (2023).

output channels (`classes`) is set to 1, as CT data is made up of 1 channel, and the final activation function is set to hyperbolic tangent to constrain the output range between the chosen minimax values.

Although both models share identical architectural hyperparameters, U-Net++ introduces a denser structure through its nested skip connections, resulting in a substantially larger number of trainable parameters. This added complexity increases computational requirements during both training and inference. Given that uncertainty quantification methods typically require multiple network evaluations, computational cost becomes a critical factor. For this reason, the comparison between U-Net and U-Net++ is conducted not just to maximize predictive accuracy, but to determine whether the simpler U-Net architecture offers a sufficient trade-off between performance and efficiency. If found adequate, the U-Net will be retained as the reference model for subsequent uncertainty estimation experiments as a strong and well known approach.

Hyperparameter	**U-Net**	**U-Net++**
`encoder_name`	`timm-res2net101_26w_4s`	`timm-res2net101_26w_4s`
`encoder_depth`	5	5
`encoder_weights`	`imagenet`	`imagenet`
`decoder_channels`	(256, 128, 64, 32, 16)	(256, 128, 64, 32, 16)
`decoder_use_batchnorm`	`False`	`False`
`decoder_attention_type`	`None`	`False`
`in_channels`	5	5
`classes`	1	1
`activation`	`Tanh`	`Tanh`

Table 3.3. Hyperparameters used for the U-Net and U-Net++ architectures. These are set given the guidelines of the SynthRad2023 challenge winners and Fabibombo (2023).

3.2.2 Baselines

Baselines play a crucial role in evaluating the performance of machine learning models by providing objective reference points against which more advanced methods can be assessed. In this work, two baselines were used, a simpler and dummy baseline as well as a more refined one. These baselines were selected to ensure consistency with prior work, particularly Thummerer, Zaffino, et al. (2020), whose pipeline and evaluation design are well established in the context of sCT generation from CBCT. Moreover, it is explained why the Corrected CBCT, a baseline often used when assessing sCT regression performance, cannot be used in the present work.

The first baseline, the Water Baseline, assigns a uniform Hounsfield Unit value corresponding to water (0 HU) across the entire synthetic CT volume. While simplistic, this baseline provides a lower bound of performance, indicating what can be achieved without using any input information. A model failing to surpass this baseline suggests either an inability to learn meaningful representations or that the data itself is not informative enough.

The second baseline, Raw CBCT, is the unprocessed Cone Beam CT. This serves as a practical reference for the starting point of the input data. If a model performs no better than this baseline, it implies that the network is not improving upon the inherent limitations and artefacts of the CBCT scan.

Each of these baseline is of interest. The water baseline basically shows the worst case scenario, a model that fails completely at predicting anything. The second baseline, the raw CBCT, shows how much a model can learn from the data itself and reflect the quality of the data.

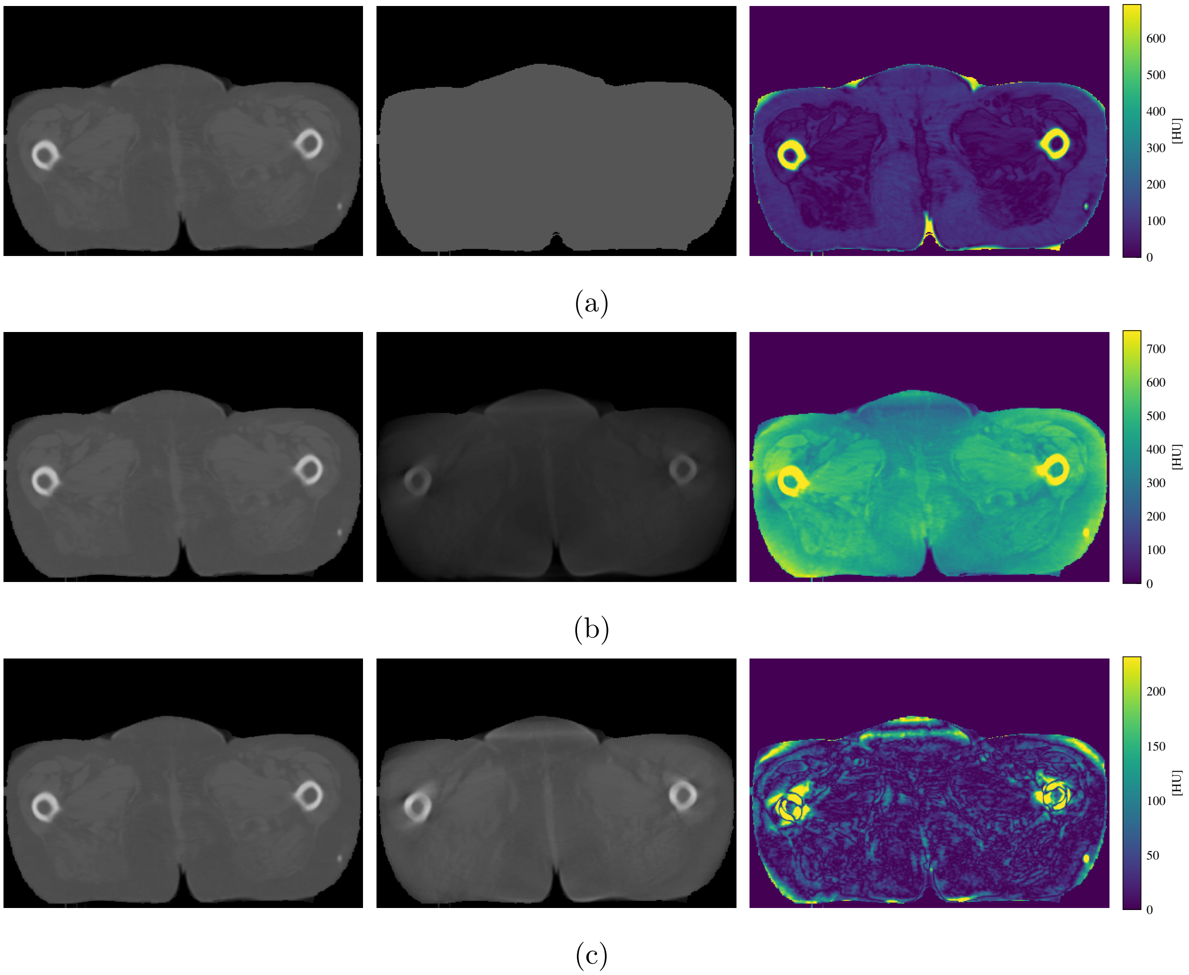
Usually, a third and last baseline is usually used is the the Corrected CBCT. This Corrected CBCT is obtained by correcting the range of HU using the CT as well as correcting low frequency artefacts with gaussian filters. However, in this work, the ground truth for one of the two datasets is the Virtual CT, which is partly obtained using this Corrected CBCT. Therefore, to avoid any potential biases in the analysis, this baseline will not be used in this work.

3.3 Uncertainty Estimation Framework

This section discusses the uncertainty estimation framework used in this thesis. The approach follows the foundational framework introduced by Kendall and Gal (2017). subsection 3.3.1 details the practical implementation of the epistemic uncertainty estimation technique, while subsection 3.3.2 explains how aleatoric uncertainty is modeled within the network.

3.3.1 Epistemic Uncertainty Estimation

Following the methods described in Section 1.2.2, epistemic uncertainty can be modeled via Monte Carlo Dropout, which approximates Bayesian inference by sampling multiple stochastic forward passes by leveraging dropout Srivastava et al. (2014) at inference time. Dropout is a regularization technique that aims to prevent overfitting by randomly setting a proportion



(a)

(b)

(c)

Figure 3.15. Example of the three baselines on patient *2PA009*. The window is adjusted to see pixels value from -1000HU to 2000HU for all modalities. The first row (a) shows the water baseline, which assigns a uniform Hounsfield Unit value of 0 across the entire synthetic CT volume. The second row (b) shows the raw CBCT, which is the unprocessed Cone Beam CT. The last row (c) shows the Corrected CBCT, which is used to obtain the Virtual CT and is therefore not leveraged in this work.

of the network's activations to zero during training. This proportion is referred to as the dropout rate.

Dropout layers were therefore introduced in the chosen model architecture. Following the work of Kendall, Badrinarayanan, et al. (2015), Dropout was applied primarily to the deeper layers of the encoder and at the start of the decoding process, rather than uniformly across the entire network. This configuration was found to be optimal in their experiments, as it allows the model to regularize deeper abstract feature representations, which are typically more sensitive to overfitting, while preserving the integrity of low-level features extracted in the early layers. Applying dropout in shallow layers was shown to act as an overly strong regularizer and degrade performance, likely because these layers extract fundamental visual patterns, such as edges, that are relatively stable and less prone to noise.

To define the appropriate dropout rate, the values of *0.1*, *0.2*, *0.3*, *0.4* and *0.5* were empirically tested. The final dropout rate used for evaluation was selected based on validation performance.

To determine the optimal dropout rate, the number of forward passes need to be determined. This number is an hyperparameter that dictates how many stochastic forward passes are performed during inference to determine the uncertainty map as well as the prediction. It is set to *50* in Kendall, Badrinarayanan, et al. (2015). This is what Kendall and Gal (2017) also used in their work. However, in a more recent work, Galapon Jr et al. (2024) used *10* forward passes and shows that it is sufficient in the case of MRI to CT synthesis. In this work, a preliminary evaluation will be made, in the real setting, for the same reasons as the one exposed in the Section 3.2, being that the values found in the real case are expected to be applicable to the same idealized case. This rate will be chosen by observing the Mean Absolute Error for a validation patient across numerous number of Monte-Carlo samples. The observations made in this work joins this conclusion and therefore the number of passes is set to *10*.

Let's denote by $\hat{y}_t(x)$ the prediction of the model on input x during the t-th stochastic forward pass with dropout active, where $t = 1, \ldots, T$ and T is the total number of samples. The Monte Carlo estimate of the predictive mean is given by:

$$\bar{y}(x) = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t(x), \tag{3.1}$$

and the corresponding predictive variance, used as an approximation of epistemic uncertainty, is computed as:

$$\mathrm{Var}[\hat{y}(x)] \approx \sigma^2 + \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t(x)^2 - \bar{y}(x)^2. \tag{3.2}$$

With the σ^2 term being the aleatoric uncertainty. When using epistemic only uncertainty quantification, data is supposed homoscedastic and as such, this term is simply overlooked as in the work of Galapon Jr et al. (2024). To have a better grasp of the uncertainty, the final map is the square root of the predictive variance, effectively giving the standard deviation of the predictions. This is done to have values in HU, which is more interpretable given the context. The mean of the predictions is used as the final prediction (Kendall and Gal 2017).

3.3.2 Aleatoric Uncertainty Estimation

Aleatoric uncertainty is modeled using a heteroscedastic loss function that will be described in Section 3.4 and follows the work of Kendall and Gal (2017).

To enable the estimation, we therefore model each of the pixel as coming from a gaussian with a mean and a variance. The model is thus trained to jointly regress both the predicted synthetic CT value as the mean, $\mu(x)$, and the corresponding data-dependent variance $\sigma^2(x)$. In the segmentation model pytorch library, this is achieved by modifying only the final convolutional layer of the decoder. No additional encoder or decoder path is instantiated.

Two mathematical tricks are applied to ensure that the predicted variance remains positive and numerically stable throughout training. Firstly, the model does not directly output $\sigma^2(x)$. Instead the network predicts either $\log \sigma^2(x)$, or alternatively, the raw variance logits are passed through a `Softplus` activation function, defined in Equation 3.3 which guarantees non-negativity and can be visualized in Figure 3.16. This activation function is a smooth approximation of the ReLU activation. Secondly, a small constant ϵ is added to the predicted variance to ensure numerical stability and avoid division by zero during training. This is particularly important when the predicted variance is very small.

$$\text{Softplus}(x) = \log(1 + e^x) \tag{3.3}$$

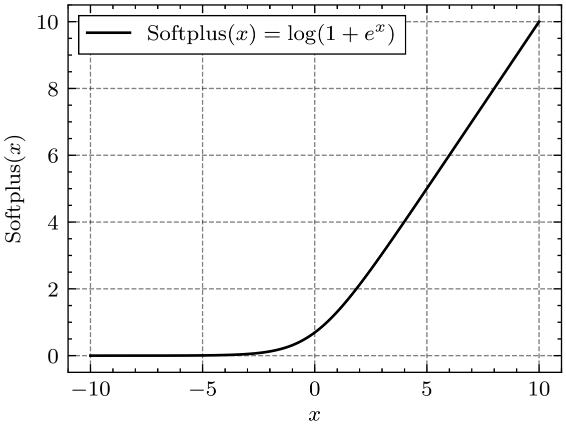


Figure 3.16. Softplus activation function, a smooth approximation of ReLU. It is used to ensure that the predicted variance remains positive and numerically stable throughout training.

With this setup, the total variance of the model's predictions can be expressed as the sum of the aleatoric and epistemic uncertainties. The aleatoric uncertainty is represented by the mean predicted variance $\sigma^2(x)$ over the T stochastic forward passes, while the epistemic uncertainty is captured by the variance of the predictions across these passes. Equation 3.4 shows how the total variance is obtained.

$$\mathrm{Var}[\hat{y}(x)] \approx \frac{1}{T}\sum_{t=1}^{T}\sigma_t^2(x) + \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t(x)^2 - \bar{y}(x)^2, \tag{3.4}$$

3.4 Training Strategy

This section describes the training procedure, including the loss functions, the optimisation procedure with the optimizers and learning rate scheduling, the training setup with the number of epochs, batch size. The hardware and software used for training will be described in Section 3.6.

Loss Functions

Two distinct loss functions are employed in this study, depending on whether aleatoric uncertainty is modeled. When the network is trained without modeling heteroscedasticity, a variation of the Mean Absolute Error (MAE) loss is used. In contrast, when aleatoric uncertainty is explicitly modeled, a heteroscedastic Gaussian Negative Log-Likelihood loss is adopted.

Deterministic Loss. For both the deterministic baseline and the epistemic uncertainty model, a weighted Mean Absolute Error (MAE) loss is employed. The motivation behind this choice lies in the large differences in HU ranges across tissue types in CT imaging. In particular, bone regions have substantially higher HU values compared to soft tissues, which would dominate the loss. The weighting is therefore applied to balance the contribution of soft tissues and bones, so that errors in both regions are treated more equally during training. The combined MAE loss is defined as:

$$\mathcal{L}_{\mathrm{MAE}} = \lambda_1 \cdot \frac{1}{|\Omega_{\mathrm{soft}}|}\sum_{(i,j)\in\Omega_{\mathrm{soft}}} |y_{i,j} - \hat{y}_{i,j}| + \lambda_2 \cdot \frac{1}{|\Omega_{\mathrm{other}}|}\sum_{(i,j)\in\Omega_{\mathrm{other}}} |y_{i,j} - \hat{y}_{i,j}|, \tag{3.5}$$

where Ω_{soft} and Ω_{other} denote the sets of pixels corresponding to soft tissue and other tissue classes, respectively, using Table 3.2. The weights λ_1 and λ_2 control the relative importance of each region. In this work, λ_1 and λ_2 are respectively set to 2 and 1.

Heteroscedastic Loss. To model aleatoric uncertainty, a heteroscedastic Gaussian Negative Log-Likelihood (NLL) loss is used. The network is trained to predict both the mean $\mu(x)$ and the variance $\sigma^2(x)$ for each pixel, modeling the conditional distribution $p(y|x) = \mathcal{N}(\mu(x), \sigma^2(x))$. The loss is given by:

$$\mathcal{L}_{\mathrm{NLL}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}\log\sigma_i^2(x) + \frac{(y_i - \mu_i(x))^2}{2\sigma_i^2(x)}\right], \tag{3.6}$$

where N is the number of valid pixels. To ensure numerical stability and positivity of the variance, the predicted logits are either passed through a `Softplus` activation. Moreover, a stable implementation of this loss is present in pytorch and is used in this work.

Regularization

Regularization is applied differently depending on whether the model includes uncertainty estimation. In the deterministic baseline model, dropout is explicitly disabled, and L2 weight regularization is employed instead, with a regularization factor set to 0.01. This choice ensures the model remains fully deterministic while still benefiting from regularization. In contrast, when uncertainty estimation is introduced, specifically epistemic uncertainty, dropout is activated and serves both as a regularization mechanism and as a means to approximate Bayesian inference. In this case, no additional L2 regularization is applied.

3.4.1 Optimization Procedure

This section covers all aspects related to the optimization procedure, including the choice of optimizer, learning rate configuration, and scheduling strategy used during model training.

Optimizer Configuration

The training process relies on the well-known Adam optimizer (Kingma et al. 2014). The learning rate is set to $\eta = 10^{-4}$, following standard practice for regression tasks. The default values for the Adam optimizer's internal parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$) are retained, as they have been shown to provide stable convergence across a wide range of tasks.

Learning Rate Scheduling

The learning rate is divided by 2 every 17 epochs during training, following a predefined schedule. This allows the model to make larger adjustments in the early stages and finer updates as it approaches convergence.

3.4.2 Training Setup

This subsection outlines the key parameters used during training, including the number of epochs and the batch size configuration.

Epochs and Batch Size

The number of training epochs was set to *50* for all models. This value was chosen to ensure convergence of the training loss, as proposed by Fabibombo (2023). No overfitting was observed during training, likely due to the variability in the input data and the regularization techniques employed. Although the validation loss was monitored throughout training, no early stopping criterion was applied.

The batch size was set to either *10* or *24*, depending on the memory constraints of the hardware available (see subsection 3.6.2). In both cases, the total training time remained comparable, averaging approximately 8 hours for the *50* training epochs.

3.5 Metrics And Evaluation

This section introduces the evaluation framework used to assess both the predictive performance and the reliability of the uncertainty estimates produced by the proposed models. It begins by defining the metrics used to quantify reconstruction accuracy, perceptual image quality, and the calibration and informativeness of uncertainty estimates. It then describes how the model is evaluated under out of distribution conditions by analyzing selected samples that include unusual anatomical content or artificial artefacts. Finally, the section presents an experiment designed to validate whether the model correctly captures acquisition noise through aleatoric uncertainty, using varying levels of simulated scatter.

3.5.1 Evaluation Metrics

This section outlines the evaluation metrics used to assess the performance of the proposed models and uncertainty quantification methods. The evaluation is divided into three main categories: reconstruction accuracy, perceptual quality, and uncertainty evaluation.

Let us denote the following quantities:

- $x \in \mathbb{R}^{H \times W}$: a CBCT input slice of height H and width W,
- $y(x) \in \mathbb{R}^{H \times W}$: the corresponding ground truth synthetic CT (sCT),
- $\hat{y}_t(x) \in \mathbb{R}^{H \times W}$: the t-th stochastic prediction of the model under Monte Carlo Dropout,
- T: the number of stochastic forward passes,
- $\bar{y}(x) = \frac{1}{T} \sum_{t=1}^{T} \hat{y}_t(x)$: the mean prediction across the T samples,
- $\sigma_t^2(x) \in \mathbb{R}^{H \times W}$: the predicted variance (aleatoric uncertainty) for sample t,
- $\sigma^2(x) = \frac{1}{T} \sum_{t=1}^{T} \sigma_t^2(x)$: the mean predicted variance over the T samples,
- $\mathcal{M}(x) \in \{0, 1\}^{H \times W}$: a binary mask indicating valid voxels to consider in the evaluation.
- $|\mathcal{M}| = \sum_{\forall x} \mathcal{M}(x)$: the total number of valid voxels within the mask.

Unless otherwise stated, all metrics are computed only over the region defined by the mask $\mathcal{M}(x)$.

Reconstruction Accuracy Metrics

To quantitatively evaluate how closely the predicted synthetic CT volumes match the ground truth, two standard regression metrics are used, namely the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE).

The MAE, defined in Equation 3.7, computes the average of the absolute voxel-wise differences between the predicted and reference CT values within the valid region defined by the mask. It provides a direct and interpretable measure of the typical prediction error in Hounsfield Units (HU), treating all errors equally regardless of their magnitude.

$$\text{MAE} \quad = \quad \frac{1}{|\mathcal{M}|}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathcal{M}_{i,j}\cdot|y_{i,j}-\bar{y}_{i,j}| \tag{3.7}$$

The RMSE, defined in Equation 3.8, computes the square root of the average of the squared differences between the predicted and ground truth CT values, again restricted to the mask. Unlike MAE, RMSE penalizes larger errors more strongly due to the squaring operation. As such, it is more sensitive to outliers and can better reflect situations where large deviations are particularly undesirable.

$$\text{RMSE} \quad = \quad \sqrt{\frac{1}{|\mathcal{M}|}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathcal{M}_{i,j}\cdot(y_{i,j}-\bar{y}_{i,j})^2} \tag{3.8}$$

Perceptual and Structural Quality Metrics

To assess the structural fidelity and perceptual quality of the generated synthetic CT images, two widely adopted image quality metrics are employed: the Structural Similarity Index Measure (SSIM) (Wang et al. 2004) and the Peak Signal-to-Noise Ratio (PSNR).

The SSIM quantifies image similarity by combining three components. The luminance l, contrast c and structure s. The luminance component compares the mean intensity of two images, the contrast component compares their standard deviations, and the structure component evaluates the correlation between them. It provides a score between -1 and 1, where 1 indicates perfect similarity. Unlike pixel-wise error metrics, SSIM is designed to reflect perceptual image quality and structural consistency as perceived by the human visual system. Equation 3.9 defines the SSIM.

$$\text{SSIM} = [l(y,\bar{y})]^{\alpha}\cdot[c(y,\bar{y})]^{\beta}\cdot[s(y,\bar{y})]^{\gamma}\,, \tag{3.9}$$

where α, β, γ are positive constants controlling the importance of each component. These values are typically set to 1 to signal equal importance across the three components.

The PSNR measures the ratio between the maximum possible pixel intensity and the MSE between the predicted and ground truth images, expressed in decibels (dB). It is computed when the data is normalized between -1 and 1, the PSNR peak value is therefore *1* and is given by Equation 3.10. Higher PSNR values indicate better image quality and lower distortion.

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{peak value}^2}{\text{MSE}}\right) = 10 \cdot \log_{10}\left(\frac{1}{\text{MSE}}\right) \tag{3.10}$$

Uncertainty Evaluation Metrics

To assess the quality and reliability of the uncertainty estimates the use of the Pearson Correlation Coefficient (PCC) and the Expected Calibration Error (ECE) are proposed, the former being introduced by Galapon Jr et al. (2024) and the latter by Kendall and Gal (2017) as calibration plots.

The PCC measures the linear correlation between two variables, providing a value between -1 and 1. A value of 1 indicates perfect positive correlation, while -1 indicates perfect negative correlation. A value of 0 indicates no correlation. It can be used to evaluate if the predicted uncertainty is correlated with the actual error.

The Expected Calibration Error (ECE) is a metric that quantifies the discrepancy between predicted confidence and observed accuracy. It is computed by dividing the predictions into several bins based on their predicted confidence. Within each bin, the average predicted confidence is compared to the actual accuracy (or error), and the absolute difference is recorded. The final ECE score is the weighted average of these differences, with weights proportional to the number of samples in each bin. This process, known as binning, plays a crucial role in how calibration is assessed. Two binning strategies are commonly used: uniform binning, where the uncertainty is evenly divided into fixed-width bins, and quantile binning, where bins are constructed such that each contains roughly the same number of samples. This work adopts both strategies to assess the robustness of the calibration analysis. A lower ECE indicates better calibration, meaning that the predicted uncertainties more accurately reflect the model's actual performance. To complement the ECE, calibration plots are also presented to provide a visual comparison between predicted confidence and observed accuracy.

3.5.2 Out of Distribution Samples

Out of distribution (OOD) samples refer to inputs that lie outside the distribution of the data used during model training. These samples often include features that are rare or absent in the training set, such as metal implants, unusual anatomical variations, or severe noise and artefacts. In medical imaging, detecting such cases is essential, as models may become unreliable when exposed to unfamiliar patterns that challenge their learned representations.

To evaluate the model's behavior under these conditions, some OOD samples are either carefully selected from each dataset or created by introducing a 800 HU square in order to mimic data corruption. These samples are chosen to ensure they were not part of the training data. The analysis investigates whether the uncertainty modeling can qualitatively detect anomalies by assessing if unfamiliar regions in the input correspond to increased predictive uncertainty in the resulting maps.

3.5.3 Evaluation of Scatter

The simulated dataset enables controlled experiments on the effect of scatter, by artificially varying its intensity across three distinct levels: 30%, 40%, and 50%. These levels are separable and serve as a proxy for increasing degrees of inherent noise in the CBCT input.

To assess whether the uncertainty estimation aligns with theoretical expectations, the experiment evaluates the average predicted aleatoric uncertainty across the three scatter levels. Since higher scatter is associated with greater acquisition noise, the mean aleatoric uncertainty is expected to increase accordingly. Comparing these means thus provides a direct test of whether the model's uncertainty estimates correctly reflect the noise level present in the data.

In addition to computing the average aleatoric uncertainty at each scatter level, a statistical analysis is conducted to formally assess whether the observed differences in means are significant. Specifically, Welch's t-test is applied pairwise across the three distributions corresponding to 30%, 40%, and 50% scatter levels. This test is chosen because it does not assume equal variances and is appropriate when comparing sample means from potentially heterogeneous distributions. A significant result would support the hypothesis that the model distinguishes between different levels of acquisition noise through its aleatoric uncertainty estimates.

3.6 Hardware and Software

This section outlines the computational infrastructure and software environment used for the development, training, and evaluation of the models. Experiments were conducted on two distinct workstations: one provided by IBA, running Ubuntu 20.04, and the other being the author's personal machine operating under Fedora 41. Although the hardware configurations and distributions differ, both systems are based on Linux kernels and leverages NVidia CUDA. To ensure reproducibility and consistency across experiments, an identical software environment was maintained on both machines.

3.6.1 Software Environment

A python virtual environment was created to isolate the project dependencies. The python version used was 3.10.12, and the following libraries were installed, alongside with their respective versions:

- `torch` 2.3.0
- `torchvision` 0.18.0
- `torchmetrics` 1.4.0.post0
- `numpy` 1.26.4
- `SimpleITK` 2.4.0
- `segmentation-models-pytorch` 0.4.0

The `torch` and `torchvision` libraries were installed with CUDA 12.1 support, enabling GPU acceleration for deep learning tasks. The `torchmetrics` library was used for some of the evaluation metrics, while `SimpleITK` was employed to manipulate the CT and CBCT scans. The `segmentation-models-pytorch` library was used to implement the architectures, as stated in Section 3.2.

A more detailed list of the libraries and their versions can be found in the `requirements.txt` file, which is included in the project's GitHub repository (Delporte 2025).

3.6.2 Hardware Configuration

The first worksation, provided by IBA, is equipped with an Intel Xeon X5650 CPU, 48 GB of DDR3 RAM and an Quadro P6000 24GB GPU. The second workstation, the author's personal machine, is powered by an AMD Ryzen 7 7800X3D CPU, 64 GB of DDR5 RAM and an RTX 3080 10GB GPU.

The reason behind the use of the two different workstation is purely practical. The first workstation was used for the majority of the training and evaluation, while the second one was used when the first one was not available. Even though training times were roughly the same on both GPU, inference time on the newer RTX 3080 were significantly faster. However, the lack of VRAM on the RTX 3080 limited the batch size, generally dividing it by 2 in regards to the Quadro P6000, hence the equivalence in training time.

The large RAM on both machines allowed for efficient data handling, enabling the loading of the whole dataset in heap. This meant that the system was not bottlencked by slower SSD or HDD speeds and no complex data input pipeline management needed to be implemented.

3.6.3 Reproducibility and Experiment Tracking

Some level of reproducibility is achieved through the use of a virtual environment, which ensures that the same versions of libraries are used across different machines. To enhance reproducibility as much as possible, good practice were followed. The first one is to use a fixed random seed across all libraries if possible. This was done by setting the random seed in Python, NumPy and PyTorch. Particular care is taken to set the random state when applying data augmentation on the mask and the images, ensuring that the same transformations are applied to both. Moreover, the use of Weight and Biases (Biewald 2020) for experiment tracking allows for easy comparison of different runs and hyperparameter settings. Each run is logged with its respective configuration. Finally, the use of a version control system, specifically Git through GitHub, allows for tracking changes in the codebase and ensuring that the same code is used across different experiments. The code developed for this thesis is available on GitHub (Delporte 2025).

Chapter 4

Results

This chapter presents the results obtained by applying the methodology detailed in Chapter 3. It is organized into three experimental phases: a preliminary step, a simulated setting, and a real setting.

The preliminary step includes two analyses. The first addresses model selection, providing a rigorous justification for choosing the UNet architecture over UNet++ given the context of this thesis, where a simpler and well known model might be beneficial. The second determines the optimal number of Monte Carlo Dropout samples required to achieve stable epistemic uncertainty estimations.

Then, the simulated setting offers a controlled environment to evaluate the proposed framework. It allows for verification of core modeling assumptions and enables a detailed assessment of both reconstruction performance and uncertainty estimation under idealized conditions.

Finally, the real setting introduces a clinically realistic scenario, characterized by acquisition noise, more artefacts, and registration imperfections. This phase assesses the framework's robustness and practical reliability when confronted with the complexities of real-world medical imaging data.

For each of both setting, the raw performance of baseline models, the deterministic baseline (U-Net without uncertainty modeling), the Monte Carlo Dropout model capturing epistemic uncertainty, and the heteroscedastic regression model designed to represent aleatoric and epistemic uncertainty is examined. Subsequently, uncertainty estimation itself is evaluated, assessing both the correlation between predicted uncertainty and observed reconstruction errors, and the calibration of these uncertainty predictions.

4.1 The Preliminary Step

This section establishes the experimental foundations by selecting the neural network architecture and determining the appropriate Monte Carlo sampling rate for uncertainty estimation. These choices ensure a consistent and computationally efficient setup for all subsequent analyses.

4.1.1 Model Selection and Sampling Rate

This subsection focuses on selecting the most suitable neural network architecture for uncertainty estimation, as well as determining the Monte Carlo sampling rate used during inference. Both evaluations are conducted using the real dataset to ensure that the choices made reflect the conditions encountered in practical applications. Since the simulated dataset is a simplified and idealized version of the real one, it is assumed that the choices validated on the real dataset, as introduced in Section 3.1.1, are also appropriate for use in the simulated setting, even if not necessarily optimal. Both evaluations are performed on the validation set to avoid introducing selection bias when reporting the results.

Model Architecture Selection

Table 4.1 compares the performance of the UNet and UNet++, on the validation set of the real dataset. Both UNet and UNet++ achieve comparable performance across all four evaluation metrics, with UNet++ showing marginally better results. However, the improvements remain small in magnitude. In terms of model complexity, UNet++ comprises approximately 69 million trainable parameters compared to 52 million for UNet, and requires roughly twice the training time per epoch (24 minutes versus 12 minutes).

Despite its slightly superior performance, the increased complexity and training cost of UNet++ may not be justified in this context. The objective of this work is not to maximize predictive accuracy but to investigate uncertainty estimation techniques. Therefore, the simpler UNet architecture is retained for the remainder of the study. Its faster training, lower computational burden, and widespread use in the literature make it a more practical and interpretable choice for this application.

	UNet	**UNet++**	**WB**	**CBCT**
↓ MAE [HU]	35 ± 13.917	**32 ± 8.665**	102 ± 23.779	283 ± 94.873
↓ RMSE [HU]	73 ± 25.101	**67 ± 26.057**	179 ± 67.226	302 ± 85.832
↑ PSNR [dB]	35 ± 3.746	**36 ± 3.848**	18 ± 0.711	24 ± 2.572
↑ SSIM	0.97 ± 0.0162	**0.98 ± 0.0109**	0.71 ± 0.0380	0.92 ± 0.0357
# trainable params	≈ **52M**	≈ 69M	–	–
Time / epoch [min]	**12**	24	–	–

Table 4.1. Comparison of UNet and UNet++

Monte Carlo Sampling Rate Selection

As illustrated in Figure 4.1, the Mean Absolute Error stabilizes for all evaluated dropout rates once the number of Monte Carlo samples reaches approximately 10. Increasing the sample count beyond this threshold does not yield noticeable performance improvements, although it does increase computational time.

This observation aligns with the findings of Galapon Jr et al. (2024), who also identified 10 samples as a practical balance between estimation accuracy and inference time in the

MRI-to-CT domain. Consequently, all subsequent experiments in this study are conducted using 10 Monte Carlo samples to ensure computational efficiency without compromising performance.

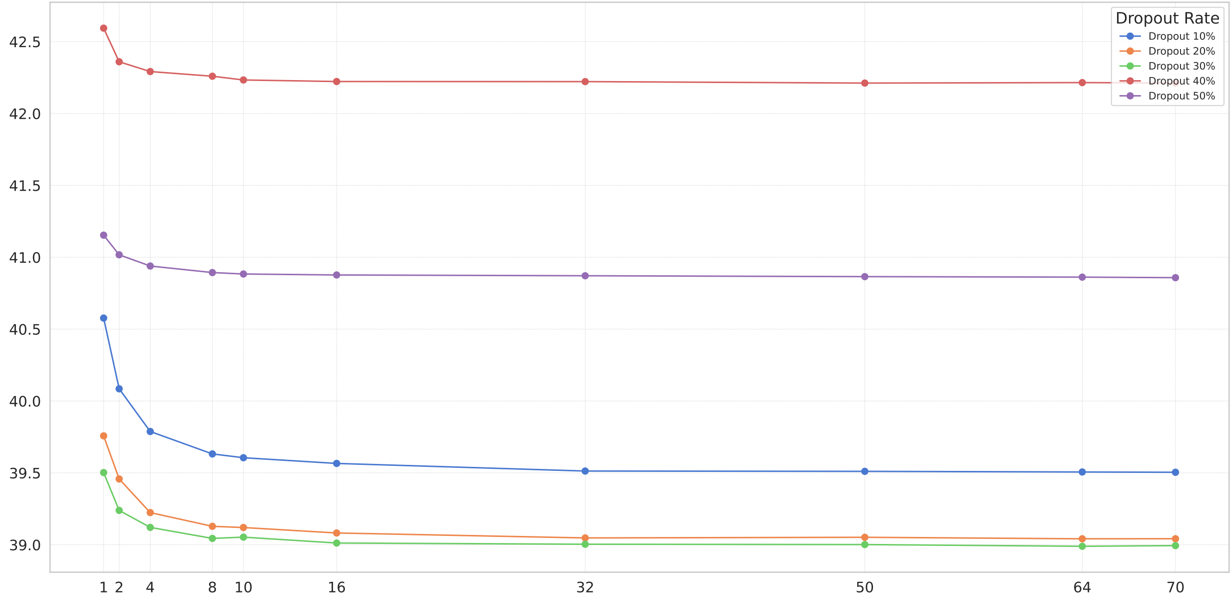


Figure 4.1. Comparison of the MAE for 2PA009 accross Monte Carlo Samples. The plot shows the MAE (y-axis) for different dropout rates (10%, 20%, 30%, 40%, and 50%) as the number of Monte Carlo samples increases (x-axis). The optimal number of samples is highlighted at 10, where the MAE stabilizes across all dropout rates. This joins the observation of Galapon Jr et al. (2024).

4.2 The Simulated Setting

This section presents the results obtained by applying the proposed framework to the simulated dataset, under idealized conditions. The objective is to assess both the reconstruction performance of the models and the quality of the uncertainty estimates produced. To this end, multiple evaluation criteria are considered: voxel-wise accuracy (MAE, RMSE), image quality (PSNR, SSIM), and uncertainty-related metrics, including Pearson correlation between predicted uncertainty and absolute error, as well as Expected Calibration Error (ECE) computed with the two binning strategies.

The first part of the analysis focuses on a baseline UNet trained purely for regression, without modeling uncertainty. This allows us to establish a reference in terms of deterministic reconstruction performance using standard image similarity metrics. It is also compared to the baselines to understand to which extent the model has learned to do what it is asked to do.

The second part evaluates the framework incorporating epistemic uncertainty only, modeled via Monte Carlo Dropout. The network's predictive performance is assessed under stochastic inference, and the resulting uncertainty maps are analyzed both quantitatively (via ECE and Pearson correlation with absolute error) and qualitatively, with the help of the visual inspection of such uncertainty estimation maps.

The final part extends the analysis to the full uncertainty aware framework, where both epistemic and aleatoric components are modeled using MCD and heteroscedastic regression. The evaluation follows the same structure, enabling direct comparison with the previous approaches. This setting is designed to test whether joint modeling improves calibration, interpretability, and error correlation beyond what is achieved by epistemic modeling alone.

4.2.1 Baseline Model

This subsection presents the performance of the baseline U-Net model trained without explicit uncertainty quantification. Both quantitative and qualitative results are reported using the simulated dataset, which offers clean ground truth CT images for direct comparison. The aim is to evaluate how well the U-Net reconstructs synthetic CT (sCT) images from CBCT inputs in idealized conditions, and to identify the model's strengths and limitations prior to integrating uncertainty modeling.

Quantitative Results

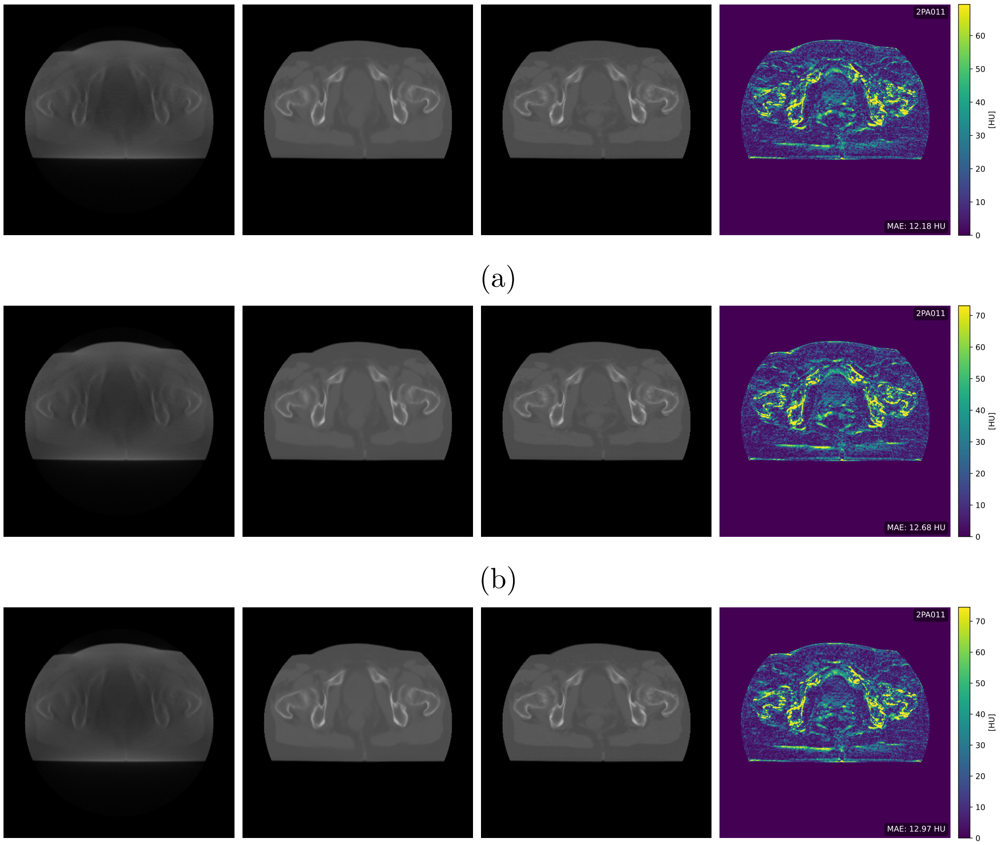
Table 4.2 shows the performance of the U-Net model for sCT generation without uncertainty quantification compared to two baselines: the original CBCT input and the water baseline. The U-Net outperforms both baselines on all reported metrics. The mean absolute error and root mean square error are significantly lower for U-Net with a respective value of 22.49 ± 2.39 and 36.31 ± 4.11. The structural similarity index is close to 1, indicating near perfect similarity with the input CT. The PSNR is high at 43.21 ± 1.01. This confirms that the U-Net coherent sCT images that aligns well in regards to the CT with the simulated dataset as training data.

	U-Net	**WB**	**CBCT**
↓ MAE [HU]	$\mathbf{22.49 \pm 2.39}$	105.13 ± 18.10	202.82 ± 32.33
↓ RMSE [HU]	$\mathbf{36.31 \pm 4.11}$	188.10 ± 36.70	245.71 ± 34.34
↑ PSNR [dB]	$\mathbf{43.21 \pm 1.01}$	15.47 ± 0.41	26.66 ± 1.37
↑ SSIM	$\mathbf{0.995 \pm 0.001}$	0.520 ± 0.042	0.889 ± 0.021
# trainable params	≈ 52M	–	–
Time / epoch [min]	12	–	–

Table 4.2. Quantitative comparison of sCT reconstruction from synthetic CBCT using U-Net, compared to water baseline (WB) and original CBCT input. U-Net shows far better performance than the two other baselines, highlighting the accurate reconstruction of sCT from simulated data.

Qualitative Results

As shown in Figure 4.2, increasing the scatter level from 30% to 50% with constant simulation acquisition settings of 125mA, 120kV, and 12ms produces visible degradation in the CBCT images. However, the resulting impact on the sCT reconstruction error remains limited. For the same slice, the MAE increases only slightly, from 12.18HU at 30% to 12.97HU at 50%, which remains small relative to the full HU range, i.e. from –1000 to 3000HU. When doing the same experiment by fixing the scatter at 30% and varying the current from 32mA, 85mA to 125mA as illustrated in Figure 4.3, the CBCT is clearly noisier when using less current and therefore giving a lower dose, rendering reconstruction harder. This impact is also visible in the reconstruction error, that is incrementally higher the lower the current. Finally, Figure 4.4 shows that in the context of the synthetic dataset, where there is no misregistration or anatomical discrepancy, the model is able to accurately reconstruct cavities, even small ones composed of only a few voxels.



(a)

(b)

(c)

Figure 4.2. Comparison of regression when increasing scatter while keeping the current at 125mA for 2PA011 with CBCT (left), sCT (middle left), CT (middle right) and absolute difference (right). The absolute difference map is accompanied with it's mean value in the bottom right corner. The same model is evaluated with scatter levels of (a) 30%, (b) 40%, and (c) 50%. The result is similar on all slices although it rises to some extent for the slices with higher scatter.

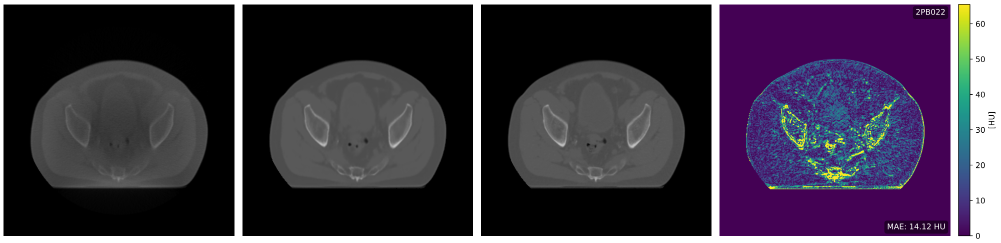
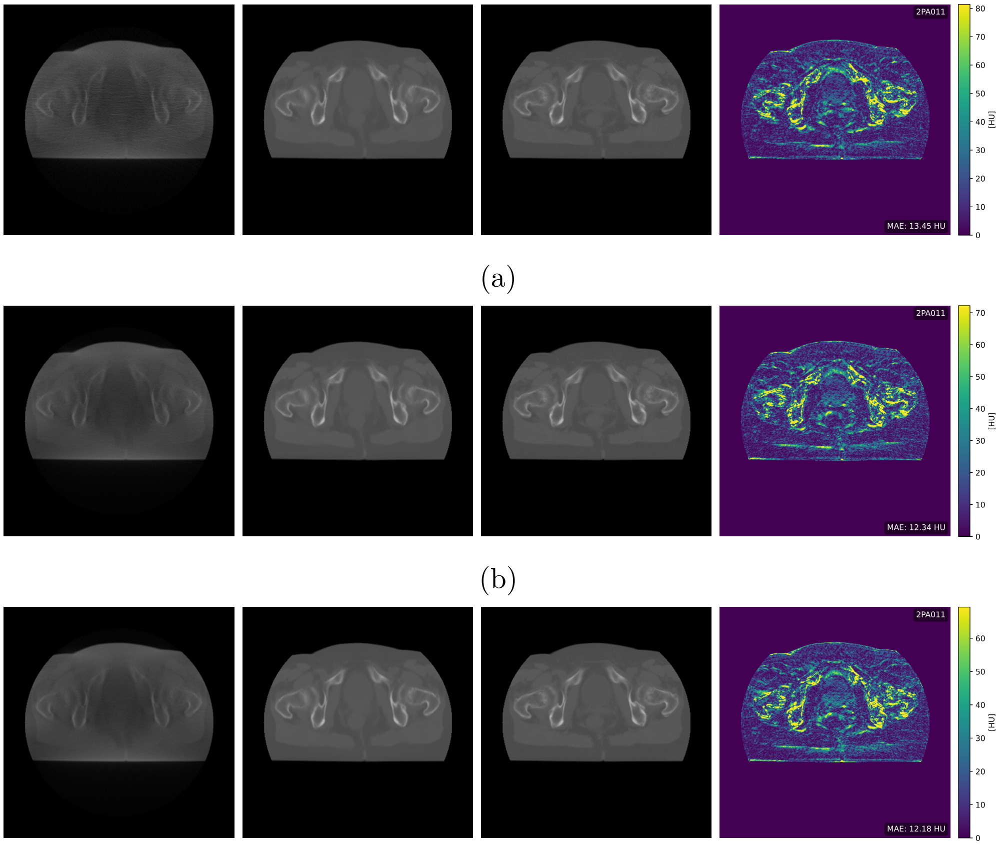


Figure 4.4. Example illustrating that, in the absence of misregistration and anatomical discrepancies, even small low-density cavities, typically challenging to reconstruct in sCT, are accurately recovered. Displayed for patient 2PB022: CBCT (left), sCT (middle left), ground truth CT (middle right), and absolute error (right).



(a)

(b)

(c)

Figure 4.3. Comparison of regression when increasing current while keeping the scatter at 30% for 2PA011 with CBCT (left), sCT (middle left), CT (middle right) and absolute difference (right). The absolute difference map is accompanied with it's mean value in the bottom right corner. The same model is evaluated with current of (a) 32mA, (b) 85mA, and (c) 125mA. One can see that the higher the current, the better quality is the CBCT and the MAE decreases accordingly. This is because when using higher current, the dose delivered by the X-Ray tube to the detector will be higher, enabling less noisy reconstruction.

4.2.2 Uncertainty via Monte Carlo Dropout

This section evaluates the impact of epistemic uncertainty modeling using Monte Carlo Dropout (MCD) applied during training and inference. The analysis focuses on how different dropout rates affect both predictive performance and the quality of the resulting uncertainty maps. Models are assessed quantitatively through the proposed metrics and uncertainty calibration measures, and qualitatively through visual inspection of uncertainty maps.

Regression Quantitative Results

The performance of the model is evaluated under different dropout rates. As shown in Table 4.3, the baseline model achieves a MAE of 22.49 HU and an RMSE of 36.31 HU. Introducing dropout as a regularization mean, during training, at 10% and 30% improves both MAE and RMSE. The best result is obtained with a 30% dropout rate, achieving the lowest MAE (21.08 ± 2.29) and RMSE (34.05 ± 3.77). This suggests that moderate regularization enhances generalization while preserving performance. At higher dropout rates of 40% and 50%, performance drops back to baseline level, indicating that too much dropout may lead to underfitting. PSNR and SSIM values remain high across all configurations, confirming that the structural integrity of the output is preserved.

Dropout Rate	**↓ MAE [HU]**	**↓ RMSE [HU]**	**↑ PSNR [dB]**	**↑ SSIM**
Baseline	22.49 ± 2.39	36.31 ± 4.11	43.21 ± 1.01	0.995 ± 0.001
10%	21.22 ± 2.05	34.39 ± 3.42	43.65 ± 0.93	0.9960 ± 0.0010
20%	22.13 ± 2.02	36.02 ± 3.52	43.26 ± 0.90	0.9957 ± 0.0010
30%	$\mathbf{21.08 \pm 2.29}$	$\mathbf{34.05 \pm 3.77}$	$\mathbf{43.74 \pm 1.01}$	$\mathbf{0.9959 \pm 0.0010}$
40%	22.27 ± 2.38	36.26 ± 3.83	43.21 ± 0.96	0.9955 ± 0.0011
50%	22.31 ± 2.36	36.32 ± 3.81	43.20 ± 0.96	0.9955 ± 0.0011

Table 4.3. Comparison of regression performance with MC Dropout and the baseline. A dropout level of 30% yields the best results across all evaluated metrics.

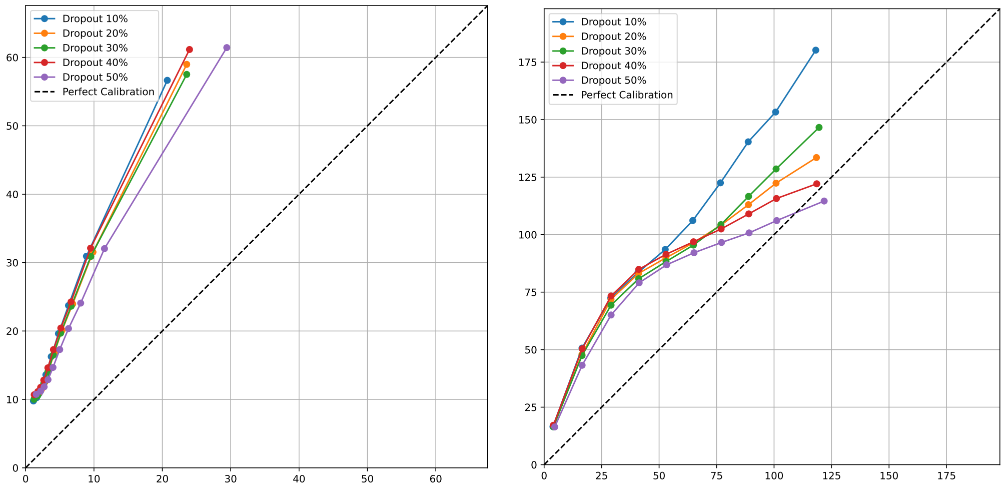
Uncertainty Quantitative Results

Table 4.4 presents the uncertainty estimation results for different dropout rates. All dropout levels yield similar moderate performance in terms of correlation with the prediction error and calibration metrics. The Pearson correlation coefficients are in the same range (around 0.54), and the ECE values do not show a clear optimal dropout rate. However, since the best regression performance was obtained with a dropout rate of 30%, this value is selected for further experiments. In particular, heteroscedastic uncertainty modeling will be explored using this dropout configuration.

Visually, the calibration curves presented in Figure 4.5 indicate that models using dropout consistently exhibit overconfidence, regardless of the binning strategy employed. When applying quantile binning, which ensures that each bin contains an equal number of samples, the predicted uncertainty aligns reasonably well with the observed error in the lower error range (0–30 HU), although still displaying slight overconfidence. As the error increases, the calibration curve deviates more from the diagonal, revealing that the model underestimates eevn more the true error in these regions. In contrast, uniform binning, which distributes the bins evenly across the uncertainty range, reveals a characteristic bell-shaped miscalibration curve. In this case, the model remains slightly overconfident at the extremities of the uncertainty range, but the overconfidence is most pronounced in the mid-range, where the predicted uncertainty systematically underestimates the actual error.

Dropout Rate	↑ **Pearson CorrCoeff**	↓ **ECE-Q**	↓ **ECE-U**
10%	**0.548 ± 0.037**	14.89	42.92
20%	0.542 ± 0.028	14.92	28.72
30%	0.541 ± 0.030	14.63	28.79
40%	0.541 ± 0.027	15.57	26.89
50%	0.544 ± 0.028	**14.24**	**21.59**

Table 4.4. Correlation between predicted uncertainty and absolute error, and Expected Calibration Error (ECE), for different dropout rates in MC Dropout. While results are comparable, a dropout rate of 30% is retained as a balanced compromise and for consistency with Table 4.3.



(a) (b)

Figure 4.5. Calibration curves for different dropout rates in MC Dropout using (a) quantile binning and (b) uniform binning. In both cases, the x-axis represents the predicted uncertainty and the y-axis the corresponding observed absolute error. In (a), most data fall in between 0 and 30 HU, where the model is slightly overconfident, with increasing miscalibration at higher error levels. In (b), the model appears better calibrated at the extremes of the uncertainty range but shows overconfidence in the mid-range, as evidenced by the bell-shaped deviation from the diagonal.

Qualitative Results

Figure 4.6 is an example of the input, output, ground truth alongside the absolute error and the predicted uncertainty map. The uncertainty map is the standard deviation of the predictions over 10 Monte-Carlo samples. The uncertainty map is visually correlated with the error map but joins the observations on calibration curves and seem largely overconfident, with uncertainty estimates shifted towards 0-40HU whereas the error goes to ≈ 90 HU.

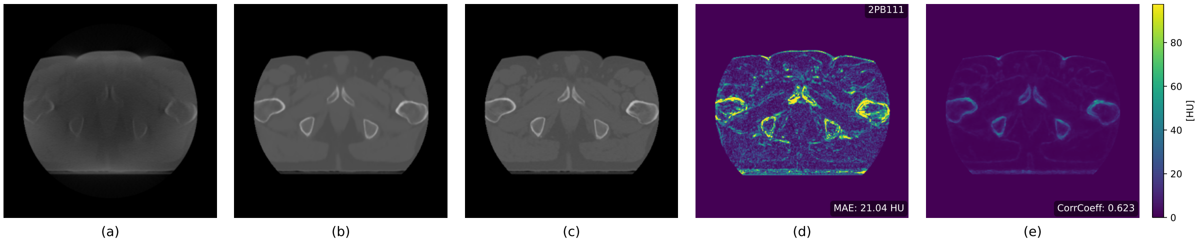


Figure 4.6. Slice prediction for 2PB111. (a) Input CBCT image. (b) Predicted sCT. (c) CT. (d) Absolute error map between (b) and (c), with the MAE displayed in the bottom right corner (21.04 HU). (e) Epistemic uncertainty map estimated via MCD. The uncertainty map seem to be visually correlated to the error map but the model seems largely overconfident.

Another example, shown in Figure 4.7, presents more contrasted results. The epistemic uncertainty map visually overlaps with several structures present in the absolute error map. However, a high-error region visible in the absolute error map in the central soft tissue area, is not reflected in the uncertainty map. This is the case even if this high error region tends to be around the 40 HU.

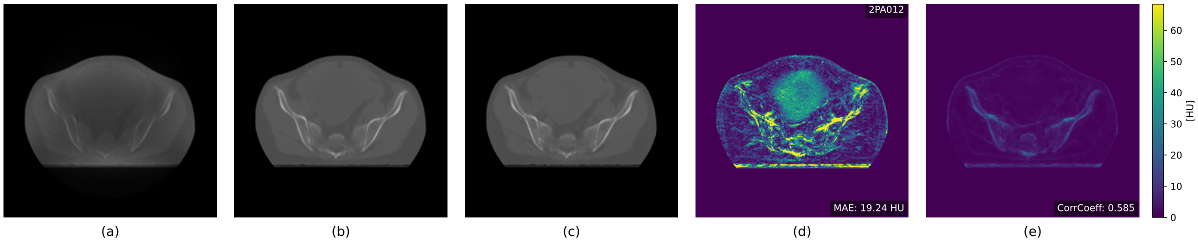


Figure 4.7. Slice prediction for 2PA012. (a) Input CBCT image. (b) Predicted sCT. (c) CT. (d) Absolute error map between (b) and (c), with the MAE displayed in the bottom right corner (19.24 HU). (e) Epistemic uncertainty map estimated via MCD. A high error region is not visible in the predicted uncertainty map.

Figure 4.8 presents an example where the epistemic uncertainty map visibly captures anatomical structures such as bone and internal cavities outlines, which are also present in the absolute error map. The uncertainty values still appear overconfident in magnitude. Additionally, both error and uncertainty are elevated along the outer skin boundary, indicating that edge regions are more prone to reconstruction errors.

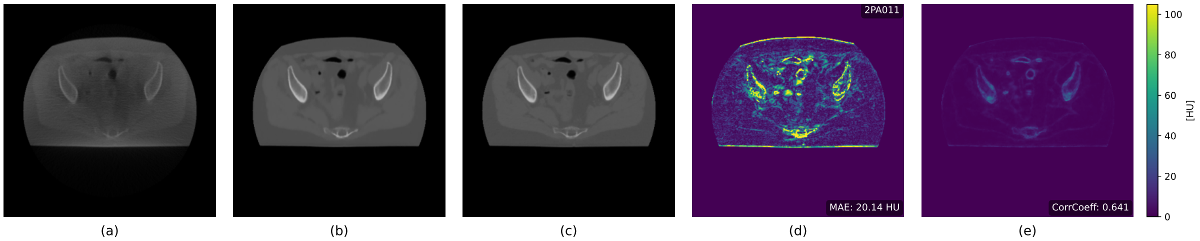


Figure 4.8. Slice prediction for 2PA011. (a) Input CBCT image. (b) Predicted sCT. (c) CT. (d) Absolute error map between (b) and (c), with the MAE displayed in the bottom right corner (20.14 HU). (e) Epistemic uncertainty map estimated via MCD. Anatomical structures like cavities and bones remain visible in both error and uncertainty maps, though the uncertainty map shows overconfidence. The skin contour shows higher prediction error and uncertainty, indicating that boundary regions are more susceptible to reconstruction errors.

Out Of Distribution Results

Finally, Figure 4.9 and Figure 4.10 show examples of the model's response to out-of-distribution inputs. In the first case (Figure 4.9), the CBCT contains a small metal structure that was never seen during training. The uncertainty map clearly shows a local spike at the location of the metal, indicating that the model identifies this region as unusual. Although the predicted uncertainty remains overconfident in magnitude, the local peak suggests that the model is surprised by the presence of this high-intensity point.

Concerning the second OOD sample (Figure 4.10), an artificial square with a value of 800 HU was embedded in an otherwise standard CBCT image. This experiment simulates the presence of corrupted acquisition data, as such a shape would not naturally occur in human anatomy. The model's prediction appears visually disturbed by this input, with unusual rectangular patterns visible in the output. The uncertainty map successfully flags the large inserted square but does not react to the secondary rectangular reflections appearing elsewhere in the image. This suggests that while the model detects the main source of corruption, it does not consistently flag all resulting anomalies.

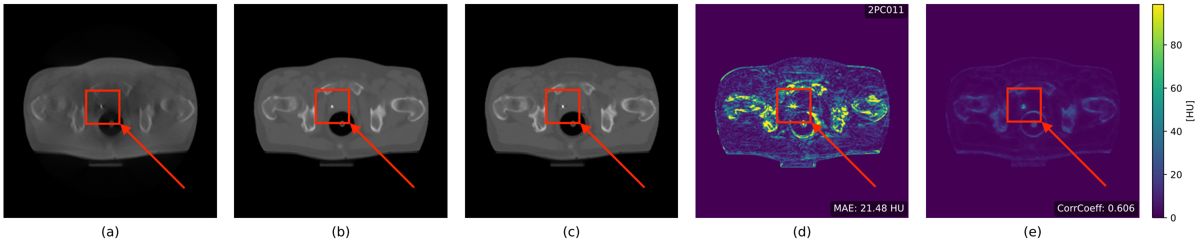


Figure 4.9. Slice prediction for 2PC011. (a) Input CBCT image. (b) Predicted sCT. (c) CT. (d) Absolute error map between (b) and (c), with the MAE displayed in the bottom right corner (21.48 HU). (e) Epistemic uncertainty map estimated via MCD. This CBCT contains a small metal artefact not seen during training. A peak in the uncertainty map correlates with the artefact location, indicating that the model detects it as out-of-distribution, despite overall overconfidence.

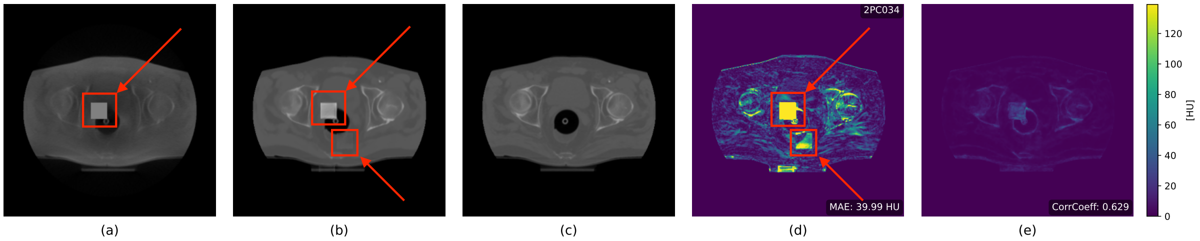


Figure 4.10. Slice prediction for 2PC011. (a) Input CBCT image. (b) Predicted sCT. (c) CT. (d) Absolute error map between (b) and (c), with the MAE displayed in the bottom right corner (39.99 HU). (e) Epistemic uncertainty map estimated via MCD. This example includes an artificially inserted square od 800 HU to simulate data corruption. The uncertainty map successfully flags the main inserted region, but fails to detect secondary rectangles that appear in the prediction. This suggests that the model identifies the primary anomaly but does not capture all effects of corrupted input.

The use of Monte Carlo Dropout in this setting yields two main observations. First, from a predictive standpoint, introducing dropout during training acts as an effective regularizer: moderate dropout rates, particularly 30%, lead to improved regression metrics compared to the baseline model. Second, from an uncertainty estimation perspective, the generated uncertainty maps show partial alignment with the prediction error, especially in regions of anatomical complexity or image noise. However, the overall calibration remains suboptimal, with consistent overconfidence across all dropout configurations. This is particularly visible in the calibration plots, where quantile binning reveals increasing miscalibration in higher error regions, while uniform binning displays a bell-shaped deviation, suggesting underestimation in mid-uncertainty ranges. Qualitatively, the maps capture obvious structures and occasionally react to out-of-distribution artefacts, such as metal or inserted anomalies, indicating that the model can detect unfamiliar features. Nonetheless, these maps often underestimate the magnitude of the true error and fail to comprehensively flag all error sources.

After evaluating epistemic uncertainty only using MCD, the full framework combining both epistemic and aleatoric components is investigated to see whether jointly modeling both types of uncertainty improves reliability and calibration.

4.2.3 Uncertainty via Heteroscedasticity and MCD

This section evaluates the impact of combining epistemic and aleatoric uncertainty through the integration of Monte Carlo Dropout (MCD) and heteroscedastic modeling. The analysis focuses on whether this joint framework improves predictive performance and produces better calibrated. The framework is assessed quantitatively using regression metrics, calibration errors, and correlation with prediction error, and qualitatively through visual inspection of uncertainty maps under both standard and out of distribution conditions.

Regression and Uncertainty Quantitative Results

Table 4.5 presents the quantitative performance of the model when both epistemic and aleatoric uncertainty are estimated jointly through Monte Carlo Dropout and heteroscedastic modeling. Compared to the baseline model and the epistemic-only framework, the combined approach slightly improves the regression performance, reaching a MAE of 20.65 ± 2.00 and an RMSE of 33.97 ± 3.76, with a PSNR of 43.76 ± 1.03 and SSIM of 0.9961 ± 0.0010. These results confirm that uncertainty modeling does not degrade the predictive accuracy of the network and even provides marginal improvements.

In terms of uncertainty evaluation, the combined model exhibits stronger correlation with the actual error and better calibration metrics compared to epistemic modeling alone as can be seen in Table 4.6. Even the epistemic estimation from combined modeling seem to be more correlated, going from 0.541 ± 0.030 to 0.65 ± 0.102. The Pearson correlation coefficient between total predicted uncertainty and absolute error is 0.652 ± 0.106, while the ECE computed using quantile binning drops to 4.41. Notably, the aleatoric uncertainty estimate is best calibrated in both the uniform binning metric (6.86) and quantile binning metric (1.21), while epistemic-only uncertainty remains largely overconfident.

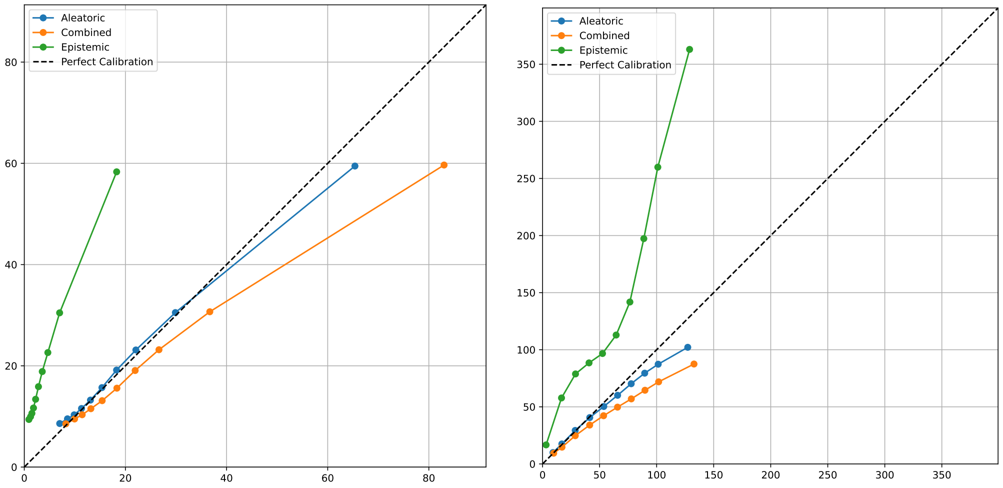
Uncertainty Type	↓ **MAE [HU]**	↓ **RMSE [HU]**	↑ **PSNR [dB]**	↑ **SSIM**
Baseline	22.49 ± 2.39	36.31 ± 4.11	43.21 ± 1.01	0.995 ± 0.001
Epistemic Dropout	21.08 ± 2.29	34.05 ± 3.77	43.74 ± 1.01	0.9959 ± 0.001
Combined	**20.65 ± 2.00**	**33.97 ± 3.76**	**43.76 ± 1.03**	**0.9961 ± 0.001**

Table 4.5. Quantitative regression results obtained with the heteroscedastic U-Net, modeling both epistemic and aleatoric uncertainty. The combined framework slightly outperforms both the baseline and the epistemic-only MC Dropout model. This confirms that uncertainty modeling does not compromise the regression performance and may enhance it.

Uncertainty Type	↑ **Pearson CorrCoeff**	↓ **ECE-Q**	↓ **ECE-U**
Epistemic Dropout	0.54 ± 0.030	14.63	28.79
Combined	**0.65 ± 0.106**	4.41	16.23
Epistemic only Combined	**0.65 ± 0.102**	15.72	81.16
Aleatoric only Combined	0.62 ± 0.105	**1.21**	**6.86**

Table 4.6. Uncertainty evaluation metrics for the explored frameworks. The combined model achieves the highest correlation with the prediction error and improved calibration. Aleatoric uncertainty alone gives the lowest calibration error, while epistemic uncertainty remains overconfident.

Figure 4.11 show the calibration curves for the combined uncertainty model, comparing the aleatoric, epistemic, and total (combined) components using both quantile and uniform binning strategies. In the quantile binning plot, the aleatoric component shows the best calibration, with predicted uncertainty closely tracking the observed error across most of the range. The combined uncertainty curve deviates moderately from the diagonal but still reflects a more cautious behavior in contrast to the epistemic component alone, which remains largely overconfident, particularly at high uncertainty levels. This pattern is even more pronounced in the uniform binning plot. In contrast, both the aleatoric and combined curves follow the diagonal more closely, indicating better overall calibration even if they seem to be too cautious at high uncertainty values. These plots support the quantitative findings from Table 4.6, confirming that the aleatoric uncertainty is best calibrated.



(a) (b)

Figure 4.11. Calibration curves for the aleatoric, epistemic, and combined uncertainty components using (a) quantile and (b) uniform binning, illustrating that aleatoric uncertainty is best calibrated while epistemic remains overconfident. Combining both components seem to lead to a over-cautious model.

Qualitative Results

The qualitative results are presented on the same slices as those used for the MC Dropout only model, to allow for direct visual comparison between the two uncertainty estimation approaches. The first example, shown in Figure 4.12, shows that the epistemic uncertainty map remains visually overconfident and appears similar in structure to the aleatoric map. The aleatoric and combined uncertainty maps are more aligned with the absolute error. For this slice, the correlation between the epistemic and aleatoric maps is 0.914. To further explore this observation, the correlation across the full dataset will be evaluated next. In addition, the controlled nature of the synthetic dataset allows the analysis of how aleatoric uncertainty evolves across different levels of scatter, in order to check if the model predictions follow the theoretical background.

Firstly, to investigate the relationship between epistemic and aleatoric uncertainty at the slice level, a log-log density plot is generated, comparing the voxel-wise predicted aleatoric uncertainty (x-axis) and epistemic uncertainty (y-axis) for a large subset of slices in the simulated dataset, as shown in Figure 4.13. Both axes are displayed on a logarithmic scale, and the color scale reflects the density of slice-level observations. It was computed using a subset of the whole training set for hardware reasons. The highest density of points lies along the perfect correlation line, indicating that, for most slices, the two uncertainty estimates are strongly correlated. Additionally, epistemic uncertainty values appear consistently lower than their aleatoric counterparts due to their overconfidence. This observation is supported by a high correlation coefficient of 0.903 ± 0.018 across all slices. These results suggest that

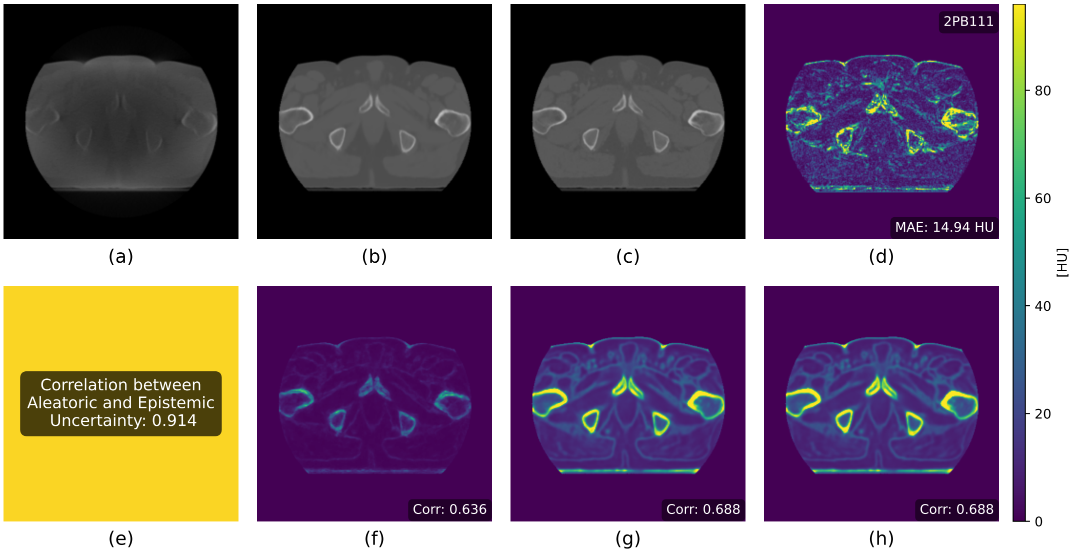


Figure 4.12. Slice prediction for 2PB111. (a) Input CBCT image. (b) Predicted sCT. (c) Ground truth CT. (d) Absolute error map between (b) and (c), with the MAE displayed in the bottom right corner. (e) Correlation between aleatoric and epistemic uncertainty for this slice. (f) Epistemic uncertainty map. (g) Aleatoric uncertainty map. (h) Combined uncertainty map. Pearson correlation with the absolute error map is indicated in the bottom right corner of (f), (g), and (h). The aleatoric and combined uncertainty maps align more closely with the error, while the epistemic map remains overconfident and visually similar to the aleatoric map.

epistemic uncertainty does not capture substantially distinct information from the aleatoric component.

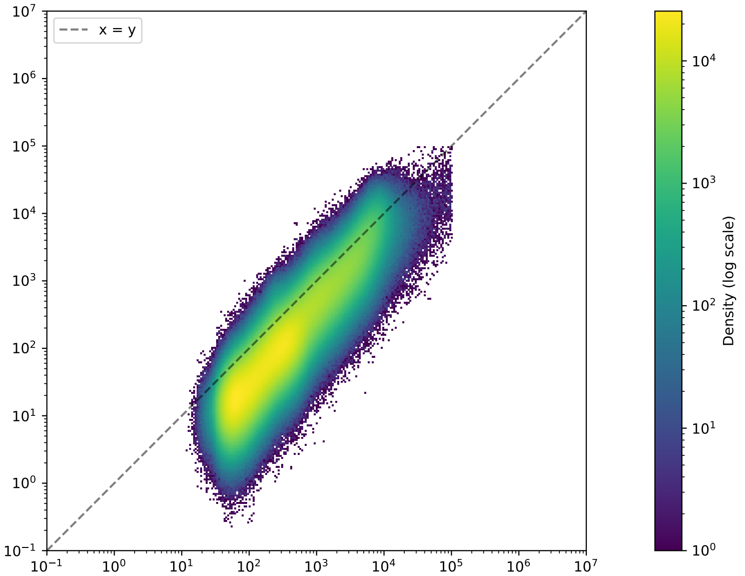


Figure 4.13. Log-log density plot comparing aleatoric uncertainty (x-axis) and epistemic uncertainty (y-axis) across the simulated dataset. The correlation coefficient of 0.903 ± 0.018 indicates a strong linear relationship between the two uncertainty types. This strong correlation is visually apparent as most density concentrates along the diagonal, with brighter regions indicating higher point concentrations on the logarithmic color scale. The epistemic values appear generally shifted, highlighting the model overconfidence

Secondly, the effect of increasing scatter artefacts in the inputs on aleatoric uncertainty was assessed by comparing the mean predicted uncertainty per slice across different scatter levels in the synthetic dataset. As shown in Figure 4.14, the histograms corresponding to 30%, 40%, and 50% scatter levels exhibit progressively higher mean aleatoric uncertainty values: 20.01, 20.15, and 20.40 HU, respectively. While the visual overlap between the distributions makes the increase less apparent, statistical testing using Welch's t-test, with a significance threshold of $\alpha = 0.05$, which compares the means of two groups while accounting for unequal variances, was chosen to ensure robustness of the results. It confirmed a significant difference between 40% and 50% ($p = 8.41 \times 10^{-4}$), and between 30% and 50% ($p = 5.74 \times 10^{-7}$). However, the difference between 30% and 40% was not statistically significant (p = 0.084). These findings indicate that the model captures an increasing trend in aleatoric uncertainty as input scatter rises, although the distinction is hard to notice.

Figure 4.15 illustrates that the overlap between epistemic and aleatoric uncertainty is particularly evident in the bony structures, where both maps display similar patterns. Additionally, the uncertainty maps appear to be well calibrated or cautious within the high-contrast bony regions.

The high-error region located in the central soft tissue area remains unflagged by either uncertainty component, reproducing the behavior observed in the MC Dropout-only framework.

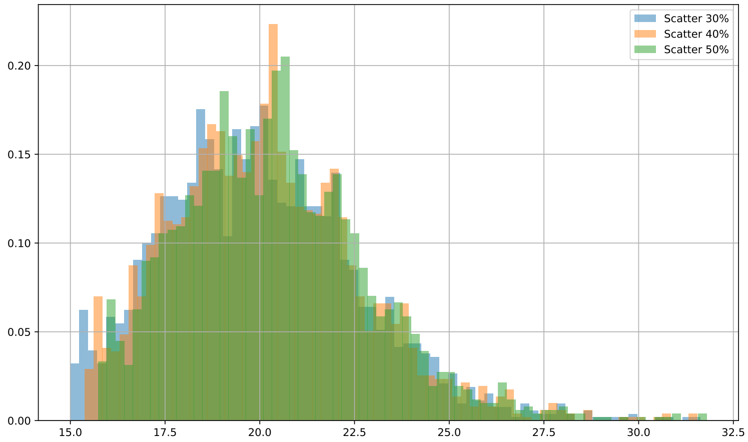


Figure 4.14. Histogram of mean aleatoric uncertainty per slice across scatter levels of 30%, 40%, and 50%, showing progressive increases (20.01, 20.15, and 20.40 HU respectively). Welch's t-tests ($\alpha = 0.05$) revealed significant differences between 40%-50% (p = 8.41×10^{-4}) and 30%-50% (p = 5.74×10^{-7}), but not between 30%-40% (p = 0.084). Results demonstrate that aleatoric uncertainty increases with input noise, though small changes may not be reliably detected.

The uncertainty seems to be lower than in surrounding regions for this specific area. The correlation values of 0.615 for both the aleatoric and combined maps and 0.565 for the epistemic map reflect a moderate alignment with the absolute error. These values are consistent with the visual observation that uncertainty partially correlates with error.

The next examples, Figure 4.16, clearly shows once again the overlap of epistemic uncertainty. However, the correlation values of 0.72 for both the aleatoric and combined maps and 0.669 for the epistemic map indicate a stronger correlation. This correlation is easily seen visually, with every uncertainty map flagging the extremities of bones and cavities as high uncertainty region. Moreover, the outline of the skin also seem to pose a problem that is flagged correctly by the maps to some extent. This observation also joins the one done in the MCD-only framework, although the aleatoric and combined maps seem better calibrated.

Out Of Distribution Results

In the next two examples, out of distribution scenarios are evaluated using the combined model. As shown in Figure 4.17, the presence of a small metal structure in the CBCT input is associated with a clear spike in the predicted uncertainty. This anomaly is consistently detected across the aleatoric, epistemic, and combined uncertainty maps, indicating that the model correctly identifies it as unusual. Notably, the aleatoric component is primarily responsible for this detection, as it displays the most pronounced local response at the metal location.

Compared to the MC Dropout-only model, the epistemic map in this heteroscedastic setting also exhibits a clearer peak at the anomaly site, suggesting that the model is more confident in its uncertainty when confronted with this unfamiliar input. However, since the combined uncertainty is strongly influenced by the aleatoric component, the overall interpretation aligns

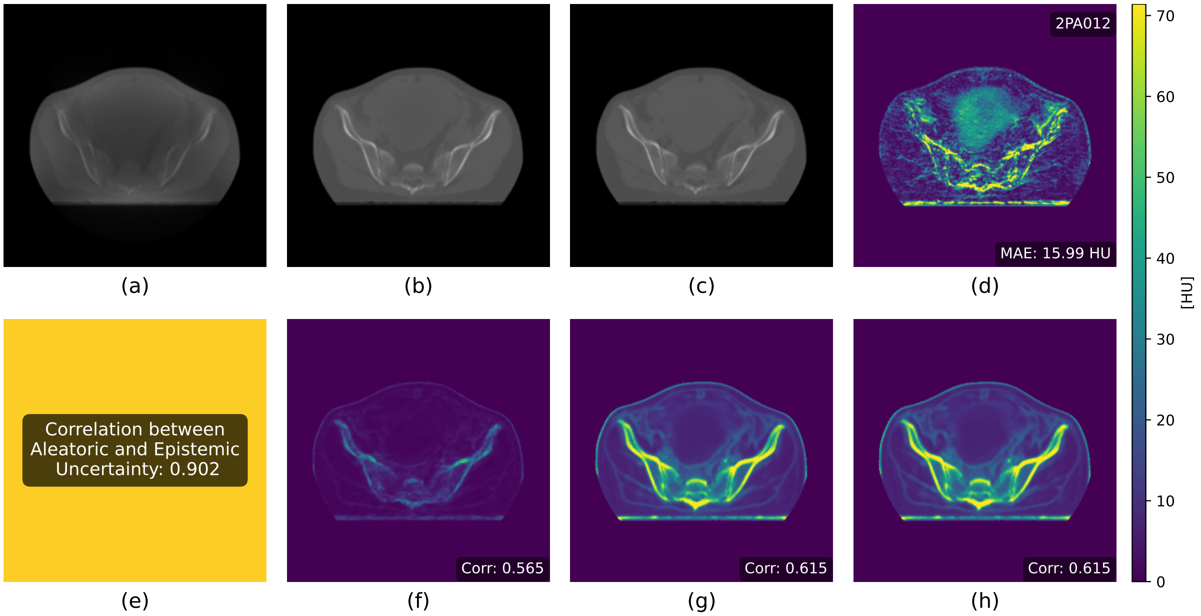


Figure 4.15. Slice prediction for 2A012. (a) Input CBCT image. (b) Predicted sCT. (c) Ground truth CT. (d) Absolute error map between (b) and (c), with the MAE displayed in the bottom right corner. (e) Correlation between aleatoric and epistemic uncertainty for this slice. (f) Epistemic uncertainty map. (g) Aleatoric uncertainty map. (h) Combined uncertainty map. Pearson correlation with the absolute error map is indicated in the bottom right corner of (f), (g), and (h). This slice highlights the collapse of epistemic and aleatoric uncertainty in the bony regions, with both maps showing similar patterns. A high-error central region remains undetected, and even underestimated, by all uncertainty maps.

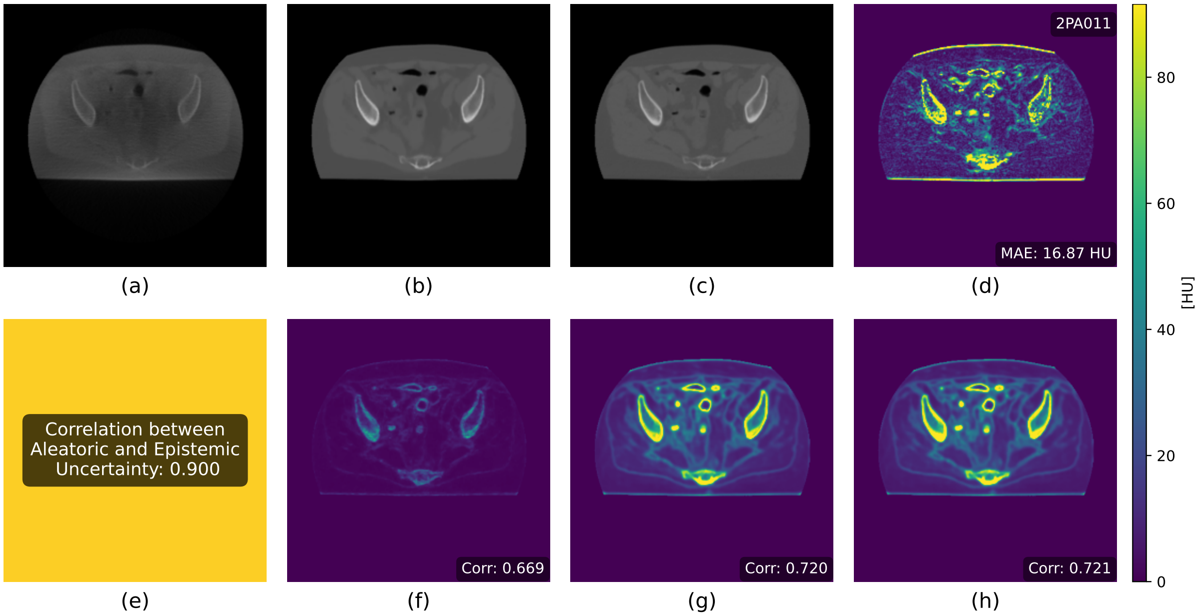


Figure 4.16. Slice prediction for 2PA011. (a) Input CBCT image. (b) Predicted sCT. (c) Ground truth CT. (d) Absolute error map between (b) and (c), with the MAE displayed in the bottom right corner. (e) Correlation between aleatoric and epistemic uncertainty for this slice. (f) Epistemic uncertainty map. (g) Aleatoric uncertainty map. (h) Combined uncertainty map. Pearson correlation with the absolute error map is indicated in the bottom right corner of (f), (g), and (h). This slice shows stronger alignment between uncertainty and error, with all maps highlighting bony edges, cavities, and skin contours. The correlation is higher than in previous examples, despite continued collapse of epistemic uncertainty.

with the one of the aleatoric estimate.

Last but not least, in Figure 4.18, the artificially embedded 800 HU square is once again added to mimic data corruption out-of-distribution. It is flagged by the uncertainty maps, especially the aleatoric component. However, two notable differences with the MCD-Only output this time, is that the aleatoric and combined map clearly defines the other corrupted part of the sCT as high uncertainty, with the uncertainty map agreeing to some extent. This was not the case previously where only the perturbation location was flagged. The other difference is that the corrupted square in the epistemic map shows a higher level in HU in regards to the MCD-only framework, just like last example, it seems much clearer.

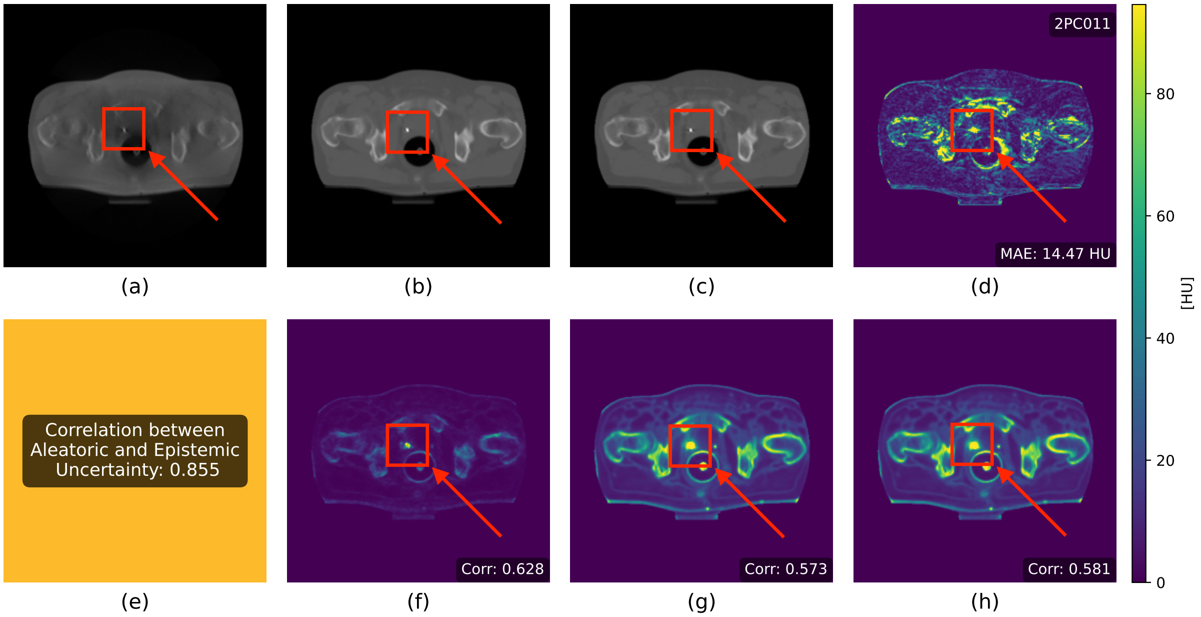


Figure 4.17. Slice prediction for 2PC011. (a) Input CBCT image. (b) Predicted sCT. (c) Ground truth CT. (d) Absolute error map between (b) and (c), with the MAE displayed in the bottom right corner. (e) Correlation between aleatoric and epistemic uncertainty for this slice. (f) Epistemic uncertainty map. (g) Aleatoric uncertainty map. (h) Combined uncertainty map. Pearson correlation with the absolute error map is indicated in the bottom right corner of (f), (g), and (h). This example features a small unseen metal artefact, which is effectively flagged by the aleatoric and combined maps. Compared to the MC Dropout-only model, the epistemic map displays a clearer and distinct spike, suggesting improved detection of out-of-distribution content.

4.3 The Real Setting

This section presents the results obtained by applying the proposed framework to the real modified SynthRad2023 task 2 dataset. As in the simulated setting, the section is structured in three parts. First, the baseline U-Net model is evaluated on the regression task without any uncertainty modeling. Second, the performance of the first uncertainty-aware framework, which estimates epistemic uncertainty only, is assessed both quantitatively and qualitatively.

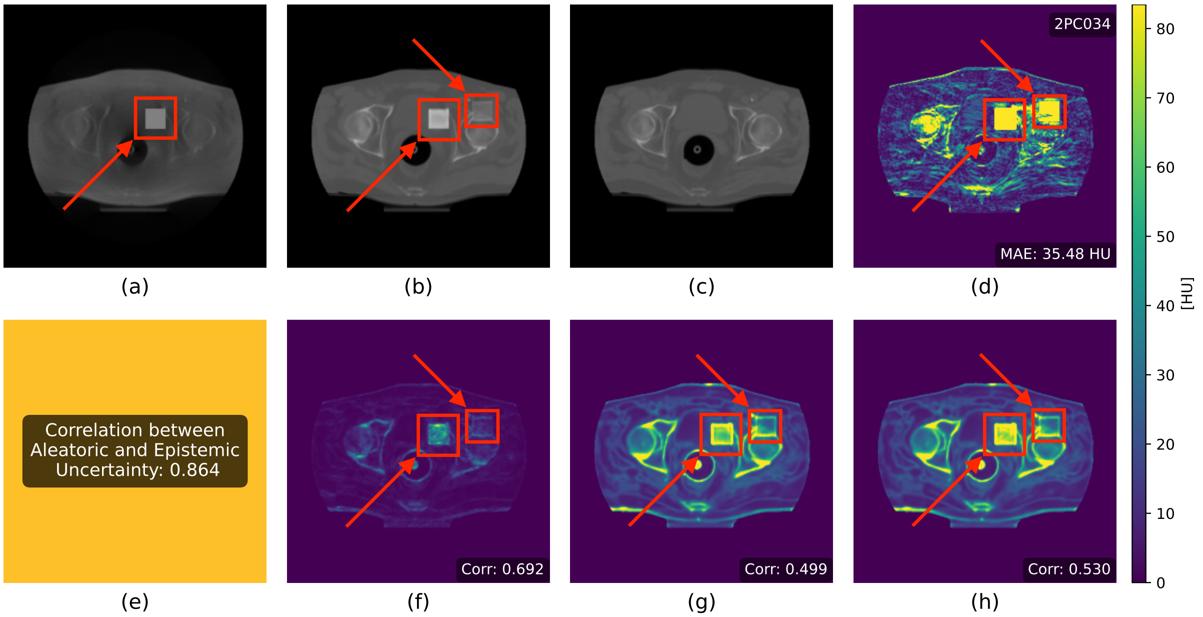


Figure 4.18. Slice prediction for 2PC034. (a) Input CBCT image. (b) Predicted sCT. (c) Ground truth CT. (d) Absolute error map between (b) and (c), with the MAE displayed in the bottom right corner. (e) Correlation between aleatoric and epistemic uncertainty for this slice. (f) Epistemic uncertainty map. (g) Aleatoric uncertainty map. (h) Combined uncertainty map. Pearson correlation with the absolute error map is indicated in the bottom right corner of (f), (g), and (h). This example includes an artificial 800 HU square to simulate corrupted data. The aleatoric and combined maps flag both the main perturbation and its secondary effects more clearly than in the MC Dropout-only model. Additionally, the epistemic map shows a more distinct response to the anomaly, further supporting improved out-of-distribution detection.

Finally, the second framework, combining epistemic and aleatoric uncertainty, is analyzed. In contrast to the previous section were the goal was to analyze the general behavior of the uncertainty, particular attention is given to examples where the real data shows acquisition noise or ground truth inaccuracies, in order to investigate how uncertainty estimates behave under such imperfect conditions.

4.3.1 Baseline Model

This section presents the performance of the baseline U-Net model trained without explicit uncertainty quantification on the real clinical dataset. Both quantitative and qualitative results are reported using paired CBCT and CT images acquired in separate sessions, which introduces several inherent limitations. These include spatial misregistration, anatomical differences due to patient movement or changes over time, and acquisition noise or artefacts. The goal is to evaluate how the model performs under realistic conditions, where ground truth is imperfect, and to highlight potential limitations that may affect reconstruction accuracy before uncertainty estimation is introduced.

Quantitative Results

On the test set, Table 4.7 presents the quantitative performance of the baseline U-Net architecture. The results are significantly worse than those obtained in the simulated setting reported in Table 4.2, where U-Net achieved MAE of 22.49 ± 2.39 HU and RMSE of 36.31 ± 4.11 HU), with a MAE of 34 ± 9.92 HU and a RMSE of 71 ± 31.10 HU. The PSNR and SSIM also drop from 43.21 ± 1.01 dB and 0.995 ± 0.001 in simulation to 36 ± 3.89 dB and 0.98 ± 0.0112, respectively. These lower performance metrics likely result from acquisition noise and ground truth imperfections of the real dataset, as discussed in Chapter 3.

	UNet	**WB**	**CBCT**
↓ MAE [HU]	**34 ± 9.917**	98 ± 14.379	278 ± 94.579
↓ RMSE [HU]	**71 ± 31.101**	175 ± 57.542	297 ± 94.042
↑ PSNR [dB]	**36 ± 3.889**	18 ± 0.786	24 ± 2.761
↑ SSIM	**0.98 ± 0.0112**	0.73 ± 0.0429	0.92 ± 0.039
# trainable params	≈ 52M	–	–
Time / epoch [min]	12	–	–

Table 4.7. Comparison of U-Net and the baselines on the modified SynthRad2023 task 2 dataset. The performance is notably lower with this dataset, likely due to the limitations discussed in Chapter 3.

Qualitative Results

The qualitative evaluation on the real dataset highlights a range of behaviors that depend heavily on the quality and characteristics of each test sample. The first case, shown in Figure 4.19, represents an example without apparent dataset artefacts. The U-Net model successfully corrects the HU range and removes most of the beam hardening visible in the

CBCT. However, soft tissues appear blurry with limited contrast, as seen in both the sCT and the corresponding error map.

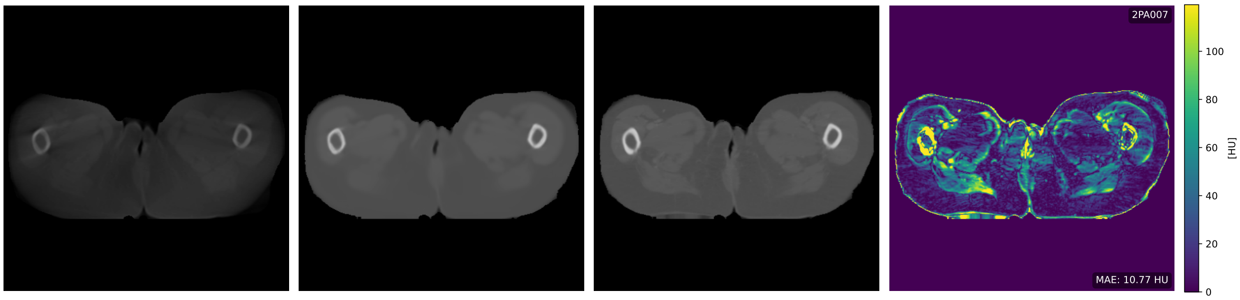


Figure 4.19. Patient 2PA007. CBCT input (left), U-Net sCT (middle left), vCT (middle right), and absolute error (right). One can see that the range of HU is well corrected by the model. However, soft tissues seem to present low contrast and be blurry, as can be seen in the sCT and the corresponding error map. The beam hardening present in the CBCT is mostly corrected, although a small residual streak remains at the left bone. This examples show the performance of the model on an example without a dataset problem.

In Figure 4.20, the artefact from the CBCT is mostly removed by the model. This correction, however, is accompanied by some degree of blurring. Despite this, the overall sCT quality is visually high, with good alignment of bone structures. Some soft tissues lack contrast, suggesting that the model fails to recover fine details in those regions.

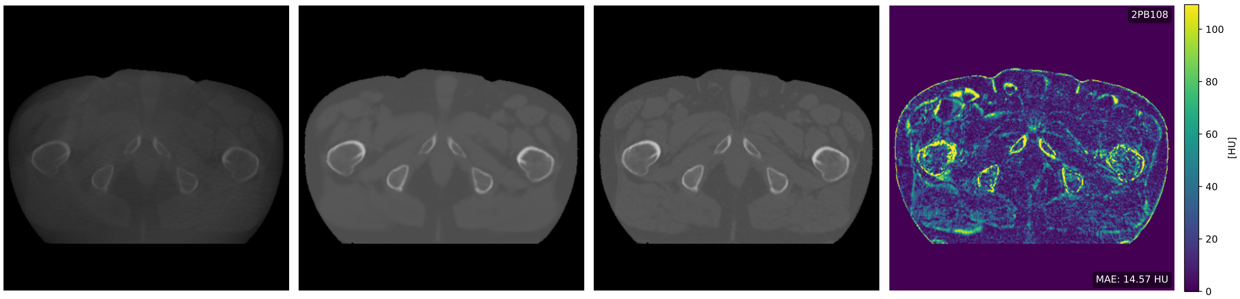


Figure 4.20. Patient 2PB108. CBCT input (left), U-Net sCT (middle left), vCT (middle right), and absolute error (right). In this example, the cupping artefact present in the original CBCT image appears to be largely corrected in the predicted sCT. However, this correction comes at the cost of some blurring. Overall, the reconstruction is visually of good quality, with bony structures well recovered and aligned. Nonetheless, certain soft tissue regions lack sufficient contrast, indicating that the model does not fully recover the fine intensity variations in those areas.

Figure 4.21 illustrates a case with a hip implant, which was not present in the training dataset and is thus treated as an out-of-distribution example. Severe metal artefact occurs around the implant, and despite correction efforts in the ground truth vCT, artefacts remain. Consequently, this case must be interpreted with caution, as the ground truth itself is unreliable.

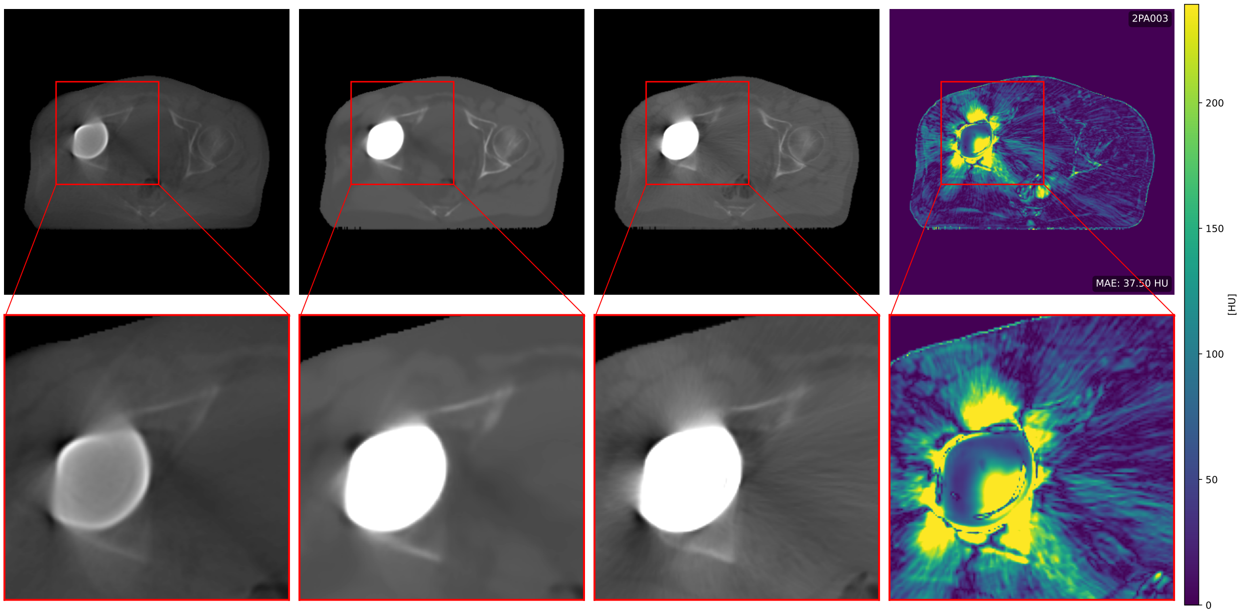


Figure 4.21. Patient 2PA003. CBCT input (left), U-Net sCT (middle left), vCT (middle right), and absolute error (right). Respective zooms of the region of interest are shown below each image. This slice represent a patient with a hip implant. Because no such patient is available in the train set, it can be used to represent an out of distribution example. However, severe beam hardening happens around the implant, and even with the correction steps applied by IBA, the ground truth CT exhibits the same problems. Therefore the results obtained with the uncertainty framework on this slice will have to be interpreted with caution.

In Figure 4.22, anatomical differences caused by the time gap between CBCT and vCT acquisitions are evident. The cavities change shape between modalities, and although IBA's pipeline attempts to correct such discrepancies, residual mismatches persist. As a result, the model's prediction appears blurred, and the errors are driven more by the dataset misalignment than by poor modeling.

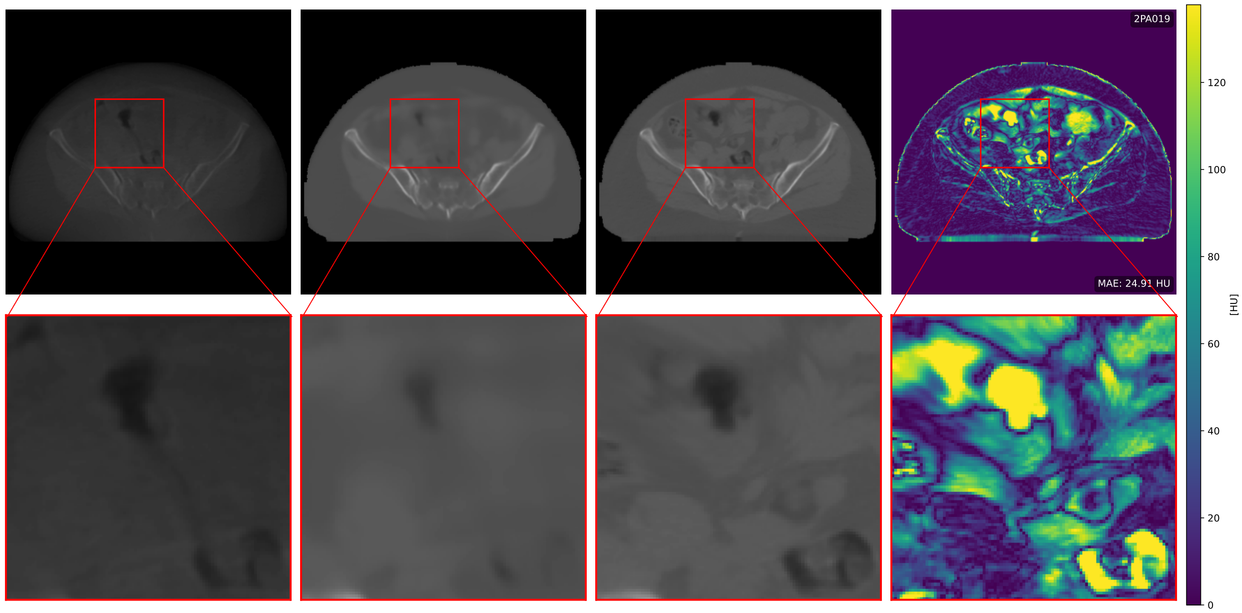


Figure 4.22. Patient 2PA019. CBCT input (left), U-Net sCT (middle left), vCT (middle right), and absolute error (right). Respective zooms of the region of interest are shown below each image. This example show how the cavities may change shape due to the time between the two modalities. These cannot all be corrected by IBA's pipeline and therefore the input CBCT and Ground Truth vCT do not correctly correspond to each other. The resulting prediction is therefore blurry, as the training set presents the same instances of problems leading to errors.

Another case of temporal misalignment is presented in Figure 4.23. Here, the input CBCT and ground truth vCT show distinctly different anatomies. Still, the U-Net manages to generate a reasonable reconstruction. Despite this, the absolute error remains high, emphasizing that the misalignment rather than model failure drives the discrepancy.

Finally, Figure 4.24 shows an input CBCT of extremely poor quality. The reconstruction quality drops significantly, particularly in soft tissues, which are not recovered appropriately. This example provides a useful basis for later analysis of how uncertainty measures behave under severely degraded input conditions. Overall, while the U-Net demonstrates reasonably good performance, its effectiveness is clearly constrained by the limitations of the real dataset, as illustrated across multiple representative cases. The results obtained through uncertainty estimation in the following sections must therefore be interpreted in light of these baseline reconstruction challenges.

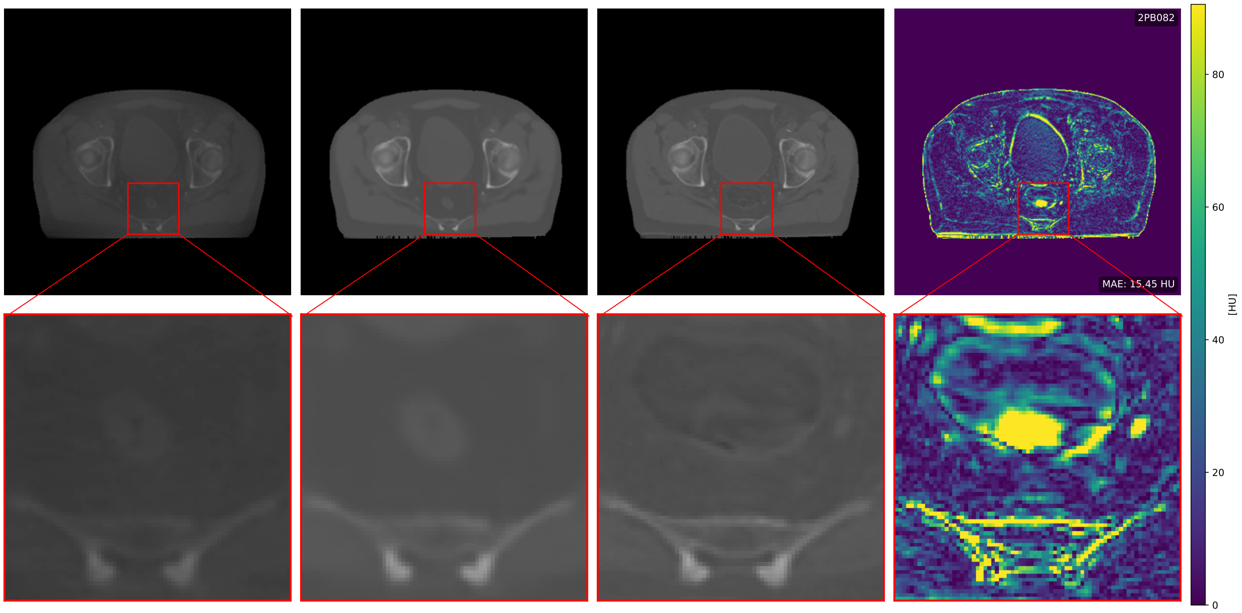


Figure 4.23. Patient 2PB082. CBCT input (left), U-Net sCT (middle left), vCT (middle right), and absolute error (right). Respective zooms of the region of interest are shown below each image. This example show yet another type of problem that is apparent in the real dataset. Due to the time between both modalities, the ground truth vCT do not present the same anatomy at all compared to the input CBCT. However in this example, the model appears to be reconstructing accurately the sCT, although the error is large nonetheless because of this discrepancy between the two scans.

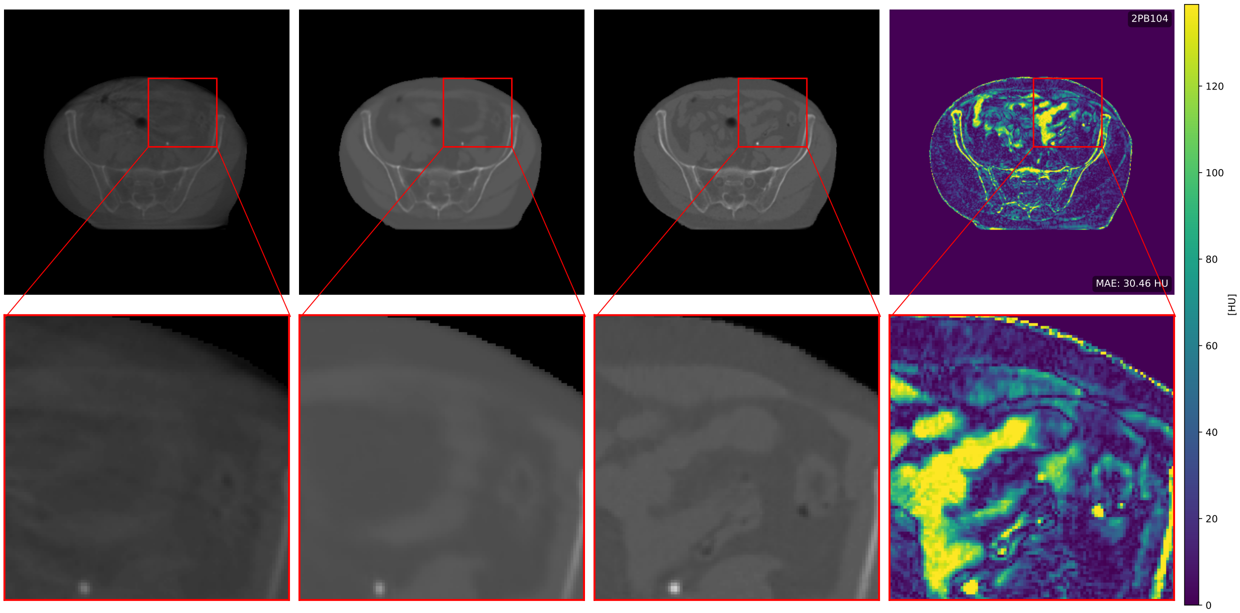


Figure 4.24. Patient 2PB082. CBCT input (left), U-Net sCT (middle left), vCT (middle right), and absolute error (right). Respective zooms of the region of interest are shown below each image. This example show an input CBCT that present a heavily degraded quality. The model is not able to reconstruct the soft tissues accordingly. It will be interesting to see how does the uncertainty measures react when confronted to such noisy inputs.

4.3.2 Uncertainty via Monte Carlo Dropout

This section evaluates the impact of epistemic uncertainty modeling using Monte Carlo Dropout (MCD) on sCT generation from real clinical CBCT images. Given the limitations of the real dataset, such as anatomical inconsistencies between modalities, registration errors, and acquisition artefacts, the analysis aims to assess how well epistemic uncertainty captures prediction confidence under these realistic conditions. Different dropout rates are explored to understand their influence on both regression performance and uncertainty quality. The models are evaluated quantitatively using the proposed metrics, and qualitatively through visual inspection of uncertainty maps.

Regression Quantitative Results

The results presented in Table 4.8 evaluate the impact of different dropout rates on the regression performance of the U-Net model applied to the real dataset with the Monte-Carlo Dropout uncertainty quantification. Across all configurations, performance metrics remain relatively stable, with only minor variations. Therefore, a dropout rate of 30% yields the best overall performance, achieving the lowest mean absolute error (33 ± 8.51 HU) and root mean squared error (67 ± 23.52 HU), while preserving high structural similarity (0.981 ± 0.014) and PSNR (36.41 ± 3.68 dB). Higher dropout rates (40–50%) result in a slight degradation of performance, likely due to underfitting. Based on these results, a 30% dropout rate is selected as the optimal configuration for the aleatoric uncertainty-aware model. This result suggests that this real dataset makes the choice of dropout rate less critical than in the simulated setting, as performance differences remain marginal across rates.

Dropout Rate	↓ **MAE [HU]**	↓ **RMSE [HU]**	↑ **PSNR [dB]**	↑ **SSIM**
Baseline	34 ± 9.917	71 ± 31.101	36 ± 3.889	0.98 ± 0.0112
10%	34 ± 8.96	67 ± 24.15	36 ± 3.72	0.981 ± 0.014
20%	33 ± 8.56	68 ± 23.89	36 ± 3.68	$\mathbf{0.981 \pm 0.014}$
30%	$\mathbf{33 \pm 8.51}$	$\mathbf{67 \pm 23.52}$	36 ± 3.68	$\mathbf{0.981 \pm 0.014}$
40%	34 ± 8.82	69 ± 24.94	36 ± 3.68	0.980 ± 0.014
50%	34 ± 8.98	69 ± 27.20	$\mathbf{36 \pm 3.79}$	0.980 ± 0.014

Table 4.8. Quantitative performance of U-Net with different dropout rates on the sCT generation task. Every dropout rate appears to yield comparable results, with 30% showing a slight improvement. It was therefore selected for the heteroscedastic model.

Uncertainty Quantitative Results

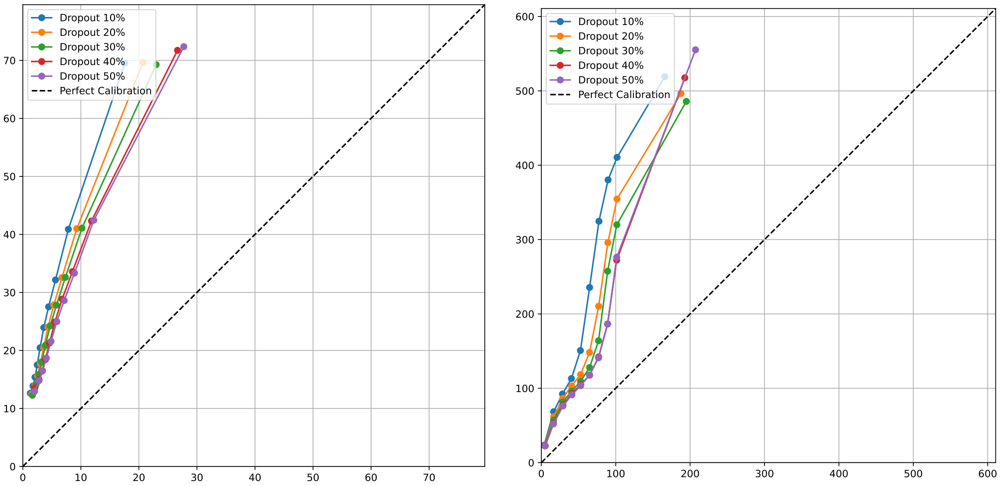
The correlations in Table 4.9 are all relatively close to each other, with a value of approximately 0.45, indicating a moderate correlation between predicted uncertainty and absolute error. The ECE values obtained using the quantile binning strategy support this interpretation, showing the same amount of moderate miscalibration across dropout rates. In contrast, the uniform binning approach yields significantly larger ECE values, likely due to the presence of noisy bins. Although the lowest ECE for both binning strategies is observed at 40–50% dropout, suggesting slightly better calibration at higher rates, a 30% dropout rate is retained for consistency with prior experiments and to limit retraining overhead, as performance differences remain small.

Dropout Rate	↑ **Pearson CorrCoeff**	↓ **ECE-Q**	↓ **ECE-U**
10%	0.45 ± 0.088	22.41	167.67
20%	0.45 ± 0.088	21.61	123.43
30%	$\mathbf{0.45 \pm 0.088}$	**21.14**	**105.11**
40%	0.44 ± 0.087	21.01	91.78
50%	0.44 ± 0.086	**20.71**	93.84

Table 4.9. Correlation between uncertainty and prediction error, and Expected Calibration Error (ECE), across different dropout rates. While correlation and quantile-based ECE values indicate constant moderate correlation and miscalibration for all rates, uniform binning shows higher values likely due to noisy bins. A 30% dropout rate is retained for consistency with Table 4.8 and to limit re-training time.

Qualitative Results

This section presents qualitative results illustrating the behavior of epistemic uncertainty, as estimated via Monte Carlo Dropout, on selected cases from the real dataset. In Figure 4.26,



(a) (b)

Figure 4.25. Calibration curves for different dropout rates in MC Dropout using (a) quantile binning and (b) uniform binning. In both cases, the x-axis represents the predicted uncertainty and the y-axis the corresponding observed absolute error. In (a), most uncertainty fall in between 0 and 30 HU. However, this proves to be highly overconfident with real error range being between 10 to 80 HU. In (b) and in contrast to Figure 4.5, the plot tends to show that the greater the error, the less calibrated every MCD model becomes with a really large error corresponding to large but not proportional uncertainty measure. This observation however joins the one of made in the simulated setting, with dropout seemingly being overconfident.

the uncertainty concentrates primarily around the bony structures and their boundaries with surrounding soft tissue.

This spatial distribution suggests that while the model captures some anatomical transitions, the uncertainty map remains overconfident and fails to reflect the full extent of the prediction error. The lack of alignment between high-uncertainty regions and actual high-error areas indicates limited reliability in this setting.

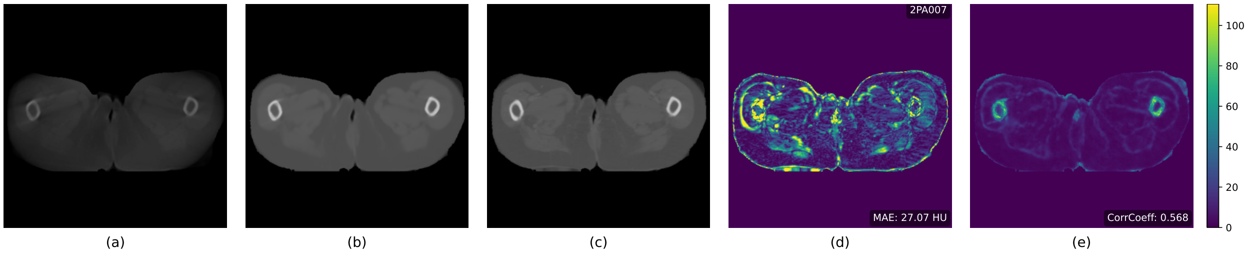


Figure 4.26. Slice prediction for 2PA007. (a) Input CBCT image. (b) Predicted sCT. (c) CT. (d) Absolute error map between (b) and (c), with the MAE displayed in the bottom right corner (21.04 HU). (e) Epistemic uncertainty map estimated via MCD. This example shows that the epistemic uncertainty is indeed overconfident and seem to be concentrated around bony structure, highlighting the boundary between bones and soft tissue.

Figure 4.27 highlights a case where cupping artefacts present in the CBCT input are largely corrected in the predicted sCT. However, the epistemic uncertainty map fails to reflect the associated error, suggesting limited sensitivity to such artefacts. Nonetheless, the correlation between uncertainty and absolute error remains relatively high (0.625), indicating that the model captures some error-related features.

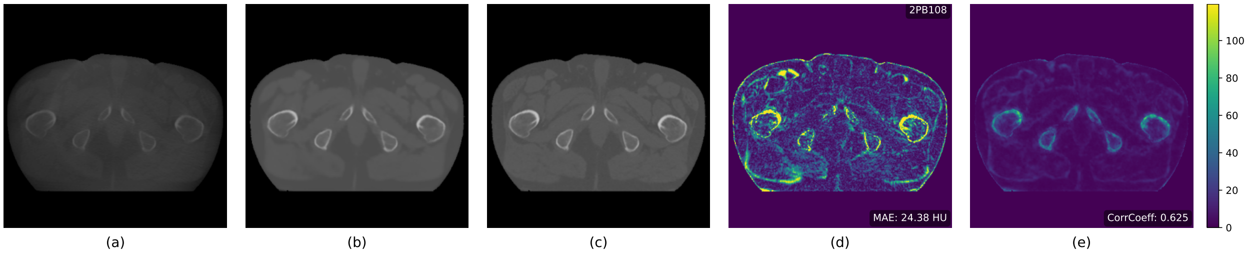


Figure 4.27. Slice prediction for 2PB108. (a) Input CBCT image. (b) Predicted sCT. (c) CT. (d) Absolute error map between (b) and (c), with the MAE displayed in the bottom right corner (21.04 HU). (e) Epistemic uncertainty map estimated via MCD. This example show that cupping artefact error is not detected by the epistemic uncertainty map, although visually as the correlation between the epistemic and absolute error is relatively high at 0.625.

Figure 4.28 demonstrates that, similar to observations in the simulated setting, epistemic uncertainty exhibits partial correlation with error in the bony regions but fails to flag several high-error zones. This further confirms the model's tendency toward overconfidence, particularly in regions where anatomical mismatch or acquisition noise is present.

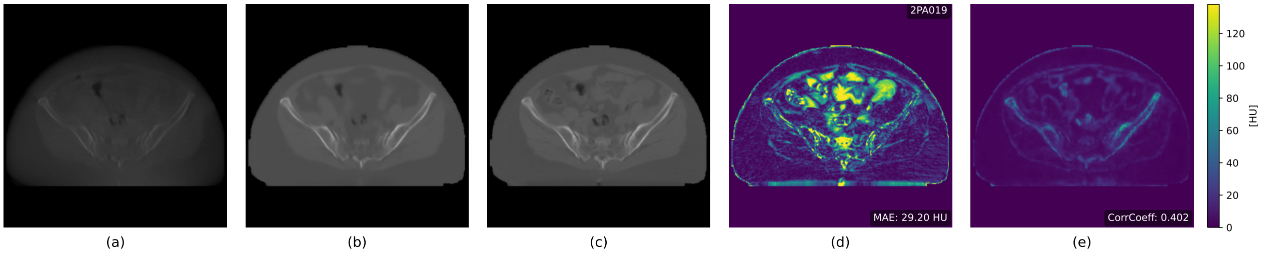


Figure 4.28. Slice prediction for 2PA019. (a) Input CBCT image. (b) Predicted sCT. (c) CT. (d) Absolute error map between (b) and (c), with the MAE displayed in the bottom right corner (21.04 HU). (e) Epistemic uncertainty map estimated via MCD. This example show that in the real setting, just as in the simulated setting, the epistemic uncertainty maps may behave in a way that they partially correlates to the error in regards to the bony structure, but they seem to completely miss some high-error regions and showing a high level of overconfidence in these regions.

Out Of Distribution results

Finally, Figure 4.29 presents a case of an out of distribution example featuring a hip implant absent from the training data. In this scenario, the epistemic uncertainty map successfully identifies the implant as anomalous, with uncertainty levels qualitatively matching the high prediction error in the affected region. This suggests that while epistemic uncertainty struggles in standard settings, it can offer meaningful signals in the presence of unseen data distributions.

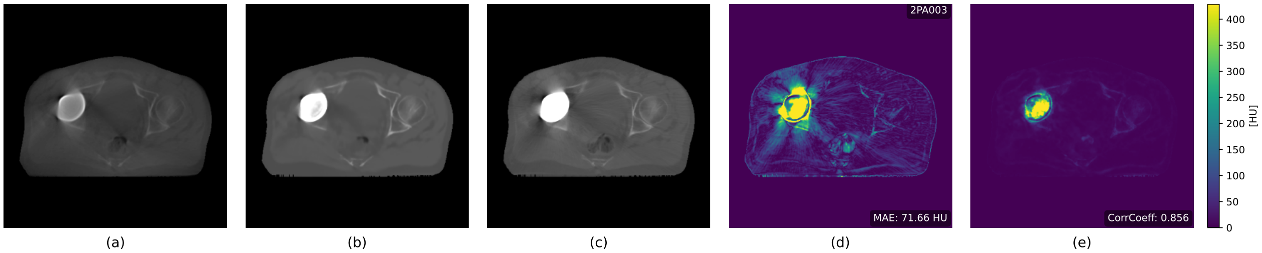


Figure 4.29. Slice prediction for 2PA003. (a) Input CBCT image. (b) Predicted sCT. (c) CT. (d) Absolute error map between (b) and (c), with the MAE displayed in the bottom right corner (21.04 HU). (e) Epistemic uncertainty map estimated via MCD. This example is present in an effort to show a real example of out of distribution data. As can be seen on the epistemic uncertainty map, the model effectively detects the implant that it has never seen as out of distribution. It can even be seen visually that the level of uncertainty seem calibrated with the actual error.

4.3.3 Uncertainty via Heteroscedasticity and MCD

This section evaluates the integration of epistemic and aleatoric uncertainty through a combined framework that leverages Monte Carlo Dropout (MCD) alongside heteroscedastic modeling. The goal is to determine whether jointly modeling both uncertainty types improves predictive accuracy, enhances the alignment between uncertainty estimates and actual errors, and results in better-calibrated uncertainty maps. The evaluation is conducted quantitatively using the proposed metrics, and qualitatively through visual inspection across representative cases from the real clinical dataset.

Regression and Uncertainty Quantitative Results

Table 4.10 reports the regression performance of the heteroscedastic U-Net on the real dataset. When compared to the baseline and the epistemic-only model, the combined framework achieves a marginal improvement in MAE, reducing it to 32 ± 8.73 HU. RMSE and SSIM values remain comparable across all models, with PSNR staying stable around 36 dB. These results indicate that modeling uncertainty, even in a noisier setting, does not degrade the reconstruction quality and may still provide slight benefits in terms of prediction accuracy. This joined the observations made in Section Section 4.2

Uncertainty evaluation metrics are presented in Table 4.11. The combined uncertainty demonstrates a stronger correlation with the absolute error (0.72 ± 0.095) compared to the epistemic-only model (0.45 ± 0.088), indicating improved error-awareness. Interestingly, the decomposition of the combined uncertainty reveals that the epistemic-only component correlates even more strongly with the error (0.77±0.06), but suffers from higher miscalibration. In contrast, the aleatoric-only component shows the best calibration performance, with the lowest ECE values of 2.35 for quantile binning and 6.45 for uniform binning.

Uncertainty Type	↓ **MAE [HU]**	↓ **RMSE [HU]**	↑ **PSNR [dB]**	↑ **SSIM**
Baseline	34 ± 9.917	71 ± 31.101	36 ± 3.889	0.98 ± 0.0112
Epistemic Dropout	33 ± 8.51	$\mathbf{67 \pm 23.52}$	36 ± 3.68	0.981 ± 0.014
Combined	$\mathbf{32 \pm 8.73}$	68 ± 25.43	$\mathbf{36 \pm 3.80}$	$\mathbf{0.981 \pm 0.013}$

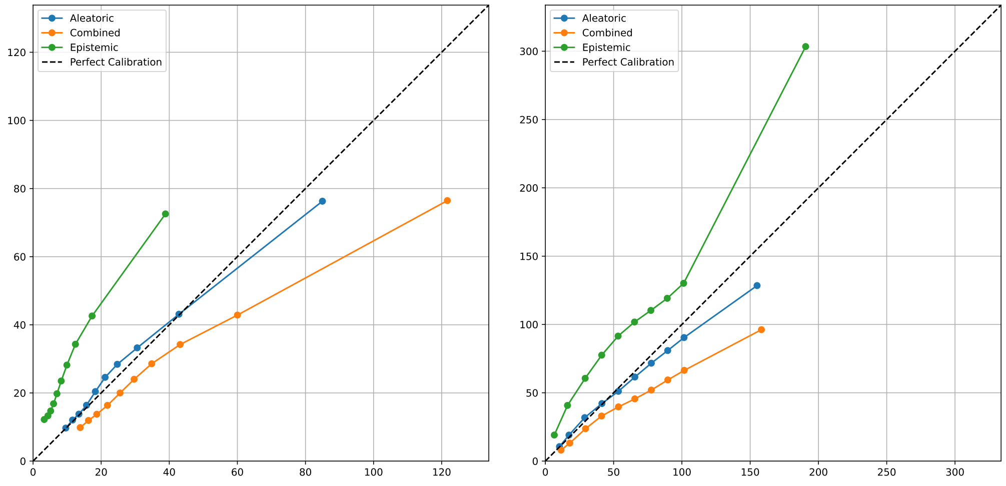
Table 4.10. Regression performance for uncertainty estimation using a heteroscedastic U-Net. Using such a framework marginally improves performance over the baseline and the MCD-only framework.

Uncertainty Type	↑ **Pearson CorrCoeff**	↓ **ECE-Q**	↓ **ECE-U**
Epistemic Dropout	0.45 ± 0.088	21.14	105.11
Combined	0.72 ± 0.095	10.76	20.93
Epistemic only Combined	$\mathbf{0.77 \pm 0.06}$	16.52	38.39
Aleatoric only Combined	0.66 ± 0.106	**2.35**	**6.45**

Table 4.11. Correlation between predicted uncertainty and absolute error, and Expected Calibration Error (ECE), for different uncertainty estimation strategies. The epistemic uncertainty only on the combined model shows the highest correlation to actual error. On the other hand, aleatoric-only uncertainty estimation yields the lowest ECE values of 2.35 for quantile binning and 6.45 for uniform binning.

To ensure comparability with the results obtained on the simulated dataset, calibration curves and a log-log density plot are analyzed. The calibration plots, shown in Figure 4.30, display the behavior of the predicted aleatoric uncertainty with respect to the empirical error using both quantile and uniform binning strategies. Although main results stays in agreement with simulated data, it is notable that the error range is much higher in this noisier and real scenario.

Additionally, the relationship between epistemic and aleatoric uncertainties is visualized using a log-log density plot in Figure 4.31. This plot reveals a strong linear relationship between the two uncertainty types across the dataset, with the majority of points aligning along the diagonal. The mean Pearson correlation coefficient across slices is 0.83 ± 0.042, confirming a high degree of redundancy between epistemic and aleatoric uncertainties. These findings are consistent with the simulated results and further support the notion that, in practice, epistemic uncertainty acts as a proxy to aleatoric uncertainty in the present scenario.



(a) (b)

Figure 4.30. Comparison of calibration curves using quantile and uniform binning strategies.

Qualitative Results

The qualitative assessment of the combined uncertainty framework is presented with the same slices as the other section in the real setting

In Figure 4.32, the epistemic uncertainty appears to follow the spatial distribution of the aleatoric map, confirming their overlap. The aleatoric uncertainty is moderately correlated with the error map and highlights the transition zones between bone and soft tissue. However, the map remains overly cautious in these regions and fails to capture certain high-error areas.

Figure 4.33 presents an example where the model successfully mitigates a cupping artefact present in the CBCT. The uncertainty maps show a reasonable correlation with the error, with both aleatoric and epistemic components primarily active in high-gradient regions such as bones and artefact boundaries. The strong correlation between both types of uncertainty (0.905) further confirms the high redundancy observed across the dataset.

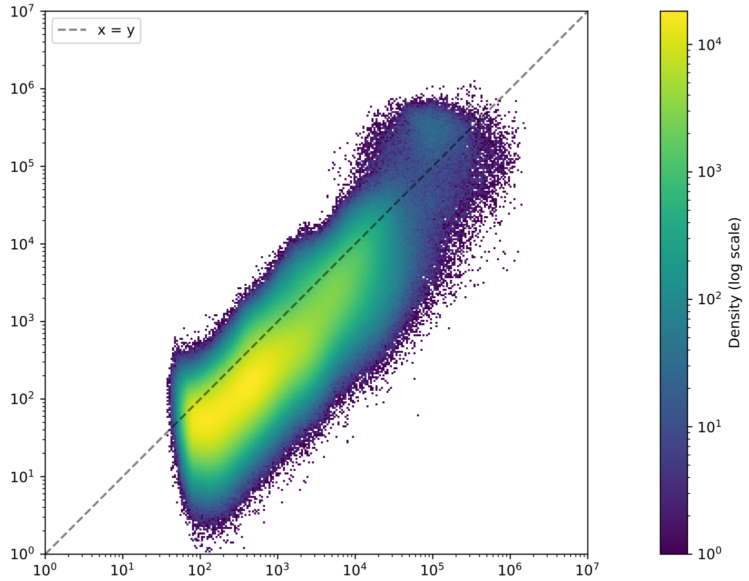


Figure 4.31. Log-log density plot comparing aleatoric uncertainty (x-axis) and epistemic uncertainty (y-axis) across the real dataset. The color scale represents the density of points on a logarithmic scale, with brighter regions indicating higher concentrations. Most of the density lies along the diagonal, suggesting that for the majority of pixels, the predicted epistemic and aleatoric uncertainties are correlated. The correlation between every slices of the dataset is of 0.83 ± 0.042, indicating a very strong correlation further showing that epistemic uncertainty is a proxy of aleatoric uncertainty, just as in the simulated case.

In Figure 4.34, the aleatoric uncertainty successfully highlights high-error regions, which were previously left undetected, especially in areas of degraded image quality. Nevertheless, while the predicted uncertainty increases in noisy regions, its magnitude still underestimates the actual error range, indicating residual overconfidence.

Out Of Distribution results

Finally, Figure 4.35 provides a real example of an OOD case, involving a hip implant unseen during training. Unlike most other examples of OOD samples seen in this work, the epistemic uncertainty clearly identifies the implant region, aligning with its theoretical purpose. This contrasts with typical behavior observed in the rest of the dataset, which pointed towards overconfidence in regards to the epistemic estimate. However, the correlation between aleatoric and epistemic components remains high (0.833), showing that even when epistemic uncertainty is informative, it remains entangled with the aleatoric signal.

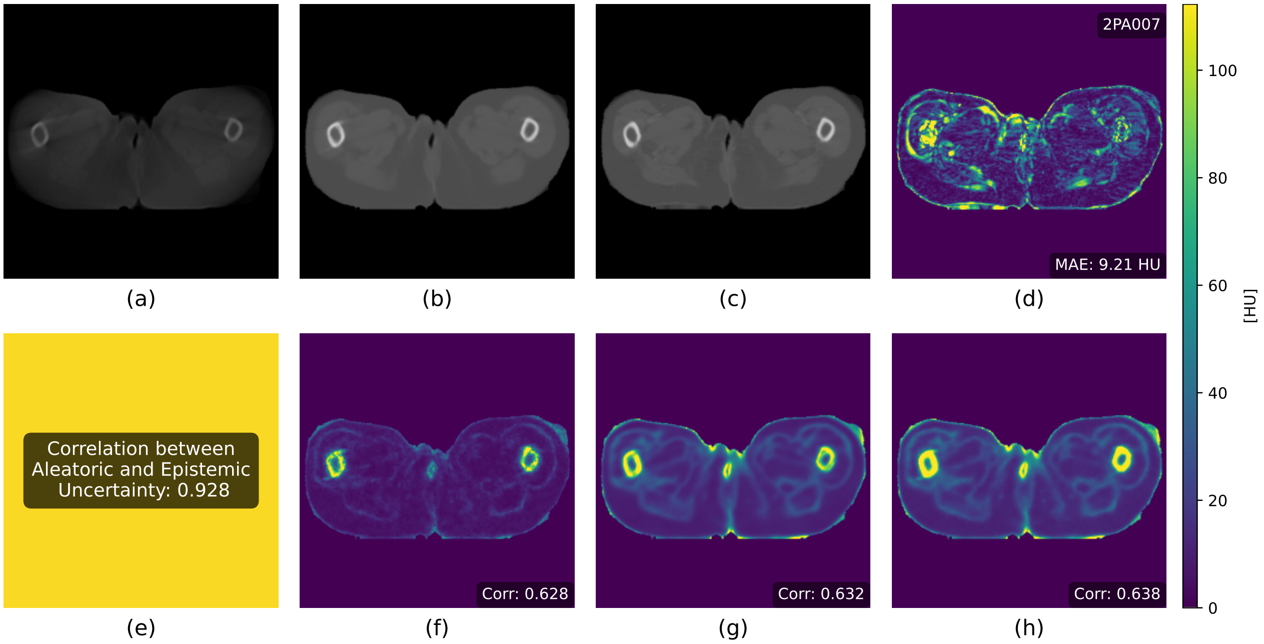


Figure 4.32. Patient 2PA007. Combined aleatoric and epistemic uncertainty. Top row: (a) CBCT input, (b) predicted sCT, (c) ground truth vCT, and (d) absolute error. Bottom row: (e) correlation map between aleatoric and epistemic uncertainty, (f) epistemic uncertainty, (g) aleatoric uncertainty, and (h) total uncertainty. MAE and correlation coefficients are indicated within the respective subplots. This example shows that epistemic uncertainty can be derived from aleatoric uncertainty and that aleatoric uncertainty seem to be visually correlated to the error map.

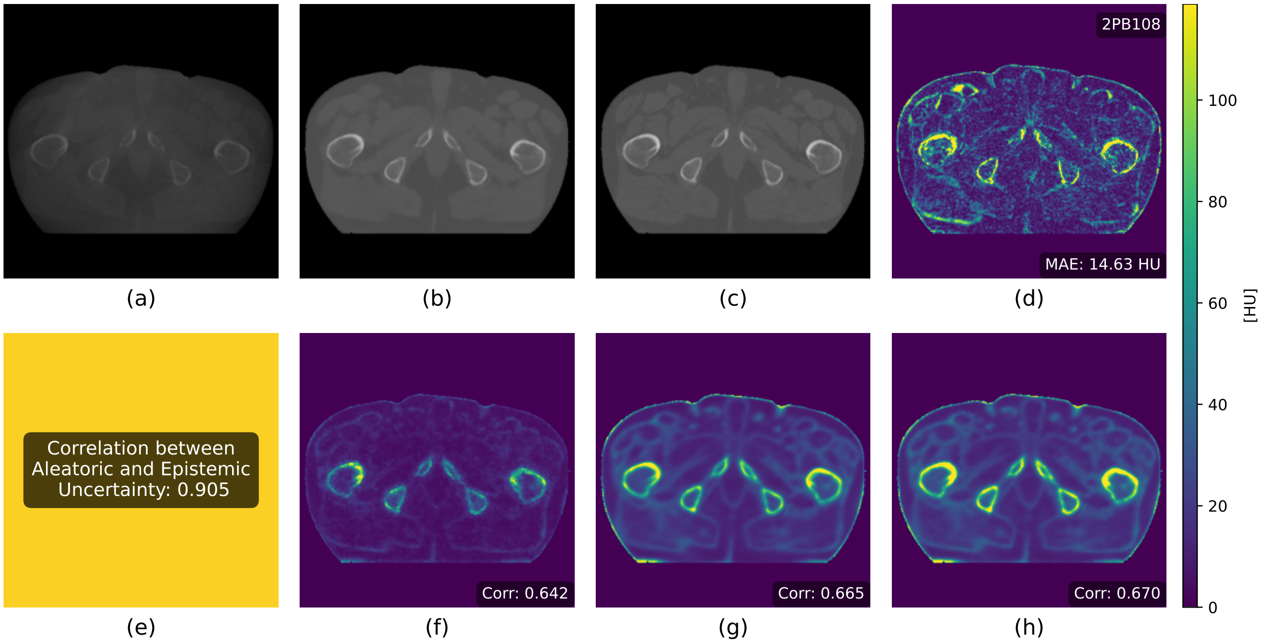


Figure 4.33. Patient 2PB108. Combined aleatoric and epistemic uncertainty. Top row: (a) CBCT input, (b) predicted sCT, (c) ground truth vCT, and (d) absolute error. Bottom row: (e) correlation map between aleatoric and epistemic uncertainty, (f) epistemic uncertainty, (g) aleatoric uncertainty, and (h) total uncertainty. MAE and correlation coefficients are indicated within the respective subplots. This example illustrates a case where the cupping artefact visible in the CBCT is mostly corrected in the sCT. The uncertainty maps show moderate correlation with the error, with uncertainty concentrated in the bony regions and at artefact boundaries. The correlation of 0.905 between aleatoric and epistemic components suggests strong redundancy, consistent with observations made across the real and simulated dataset.

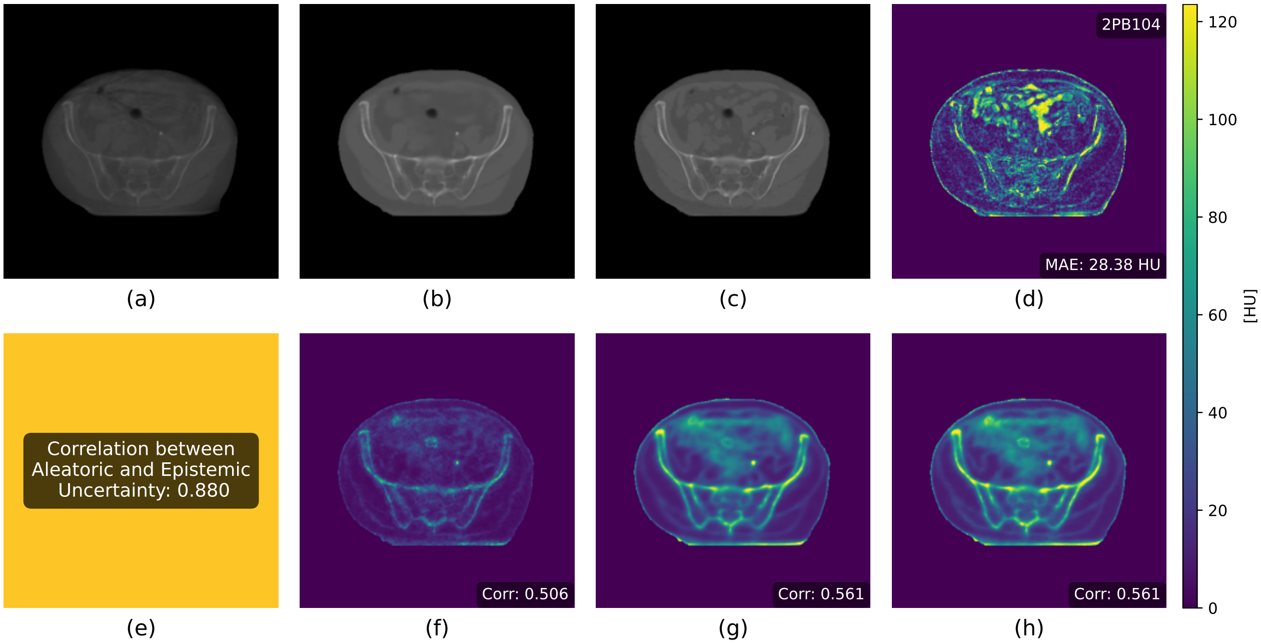


Figure 4.34. Patient 2PB104. Combined aleatoric and epistemic uncertainty. Top row: (a) CBCT input, (b) predicted sCT, (c) ground truth vCT, and (d) absolute error. Bottom row: (e) correlation map between aleatoric and epistemic uncertainty, (f) epistemic uncertainty, (g) aleatoric uncertainty, and (h) total uncertainty. MAE and correlation coefficients are indicated within the respective subplots. This examples, that showed how high error zone in the absolute error were not translated in higher uncertainty for the MCD only model, shows a much more contrasting result when taking the aleatoric uncertainty into account. In fact, it seems like the aleatoric uncertainty is indeed higher in this high noise region, highlighting the lack of confidence of the model. However, aleatoric uncertainty still seems overconfident, as the range in HU is of ≈ 40-60 HU whereas the real error ranges from 40-120 HU on visual inspection.

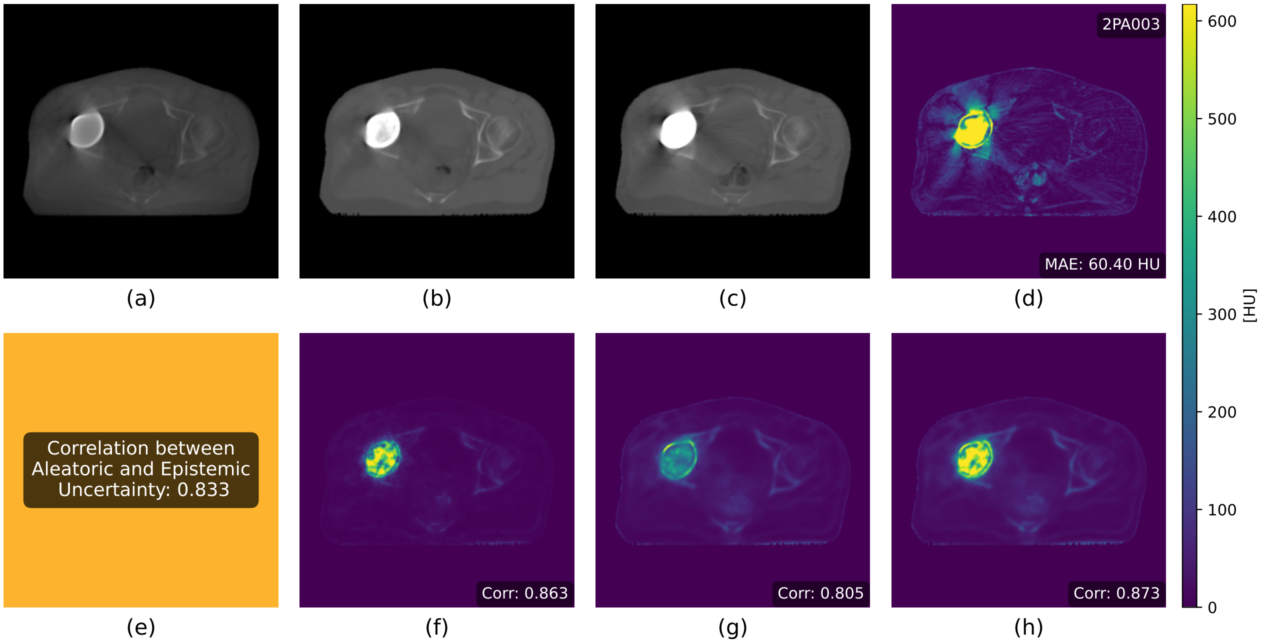


Figure 4.35. Patient 2PA003. Combined aleatoric and epistemic uncertainty. Top row: (a) CBCT input, (b) predicted sCT, (c) ground truth vCT, and (d) absolute error. Bottom row: (e) correlation map between aleatoric and epistemic uncertainty, (f) epistemic uncertainty, (g) aleatoric uncertainty, and (h) total uncertainty. MAE and correlation coefficients are indicated within the respective subplots. In contrast to the simulated dataset and previous examples, the epistemic uncertainty in this out-of-distribution case effectively highlights the implant region more distinctly than the aleatoric component, aligning with the theoretical role of epistemic uncertainty in identifying unfamiliar patterns. Nevertheless, the correlation between aleatoric and epistemic maps remains high at 0.833, suggesting that despite this divergence, redundancy between the two components persists.

Chapter 5

Discussion

This section aims to address the research question: *Can state-of-the-art deep learning-based uncertainty quantification methods provide trustworthy and interpretable outputs when applied to synthetic CT generation from CBCT images?*. To answer this, the two essential properties of uncertainty estimates tied to the research question, interpretability and trustworthiness, are examined across the different uncertainty modeling strategies explored in this work.

In this context, interpretability refers to the degree to which the uncertainty estimates produced by the model can be meaningfully understood by the clinicians leveraging these maps. An interpretable uncertainty map should highlight regions that are less reliable, such as those affected by image noise, metal artifacts, or anatomical ambiguity. Furthermore, interpretability entails that different sources of uncertainty, such as model ignorance and data noise, are distinguishable in order to adapt the clinical workflow.

Trustworthiness, on the other hand, concerns the reliability and consistency of uncertainty estimates in reflecting actual prediction error. A trustworthy uncertainty quantification method should not only correlate well with prediction errors but also be well-calibrated: low predicted uncertainty should correspond to low actual error, and high uncertainty should signal regions where the model's output is potentially unreliable. Furthermore, a trustworthy model should be able to highlight out of distribution samples. In practical terms, trustworthiness ensures that uncertainty can be used as a safety measure, guiding clinicians to make informed decisions about whether to rely on, verify, or reject a synthetic CT prediction.

The discussion is organized into four main parts. The first evaluates the use of Monte Carlo Dropout for epistemic uncertainty estimation, with a focus on its performance across both simulated and real clinical datasets. The second part investigates a combined framework that models both epistemic and aleatoric uncertainty using heteroscedastic modeling with MCD, assessing its ability to improve calibration, enhance interpretability, and provide more comprehensive uncertainty decomposition. The third part reflects on the impact of dataset quality and modeling choices, outlining key methodological limitations that may affect both model performance and the evaluation of uncertainty. The final part synthesizes the findings to answer the research question. While the analysis is based on quantitative and qualitative indicators such as calibration curves, correlation coefficients, and OOD detection behavior, its central aim remains to assess whether the uncertainty maps produced are interpretable and trustworthy enough to support clinical decision-making, particularly in regards to these indicators.

5.1 Monte Carlo Dropout

This section evaluates the performance of Monte Carlo Dropout as a method for estimating epistemic uncertainty in the context of synthetic CT generation from CBCT images. The goal is to assess whether MCD provides interpretable and trustworthy uncertainty estimates that can support clinical decision-making. The results show that while MCD offers a sound theoretical foundation and marginal improvements in prediction accuracy, it suffers from key limitations. These include limited interpretability of the uncertainty maps, systematic overconfidence, and a general lack of robustness. Together, these findings raise concerns about the suitability of MCD as a standalone solution for uncertainty estimation in this setting as well as a broader concern about the quality of the dataset commonly currently used in research.

5.1.1 Regression Performance

A first aspect, that supports trust in the proposed approach, lies in the improved regression performance observed when using dropout during training (Table 4.3). Although the improvement observed in this work is marginal and the dropout rate was not thoroughly optimized, the gain still suggests a modest increase in model reliability. The sCT output is slightly closer to the ground truth compared to the baseline model without uncertainty estimation, suggesting that incorporating epistemic uncertainty contributes positively to the predictive quality. This result is however not surprising as Dropout, the main mechanism used in the Monte-Carlo Dropout framework, was originally created for regularization purposes (Srivastava et al. 2014).

Correlation and Calibration

One key limitation to interpretability comes from relying only on epistemic uncertainty. As explained by Kendall and Gal (2017), epistemic uncertainty does not capture the full range of uncertainty present in real-world data. It only reflects uncertainty due to a lack of knowledge or limited training data. This means that the model cannot separate uncertainty into two parts: one that can be reduced with more data (epistemic) and one that is inherent in the data itself (aleatoric). Without this distinction, it is difficult to interpret what the uncertainty values actually mean.

Besides the conceptual limitation of using only epistemic uncertainty, the uncertainty maps produced by the model also show a reduced dynamic range. The calibration curve (Figure 4.5) and the moderate correlation between uncertainty and error (Table 4.4, ≈ 0.54) suggest that the predicted uncertainty behaves like a compressed or scaled-down version of the actual error map. This compression effect intensifies as the prediction error increases, limiting the ability of the uncertainty map to reflect the true variability of the model's performance. As a result, the maps are harder to interpret, particularly in clinical contexts where understanding the model's confidence is crucial.

This compression is also reflected in the model's miscalibration. As seen in the same calibration curve, the model becomes increasingly overconfident as errors grow larger. In a well-calibrated system, high prediction errors should correspond to high predicted uncertainty. However, in this study, the model does not present the same amount of overconfidence in high-error regimes as it does in a more moderate error regime. This non-uniform miscalibration significantly undermines the trustworthiness of the uncertainty estimates, especially in high-risk clinical situations. Because the overconfidence is not linearly related to the error and tends to worsen with larger mistakes, it raises serious concerns about the safe deployment of the model in practice.

Another limitation in terms of trust, is the relationship between predicted uncertainty and actual error that shows important weaknesses as the correlation between the two is moderate. The value of 0.54, found in this work, is much lower than the correlation reported by Galapon Jr et al. (2024) (about 0.9) for a similar method applied to MRI-to-sCT translation. One possible reason for this difference is the absence of post-hoc calibration in the present study, although Galapon Jr et al. (2024) do not clearly report such calibration either in their methodology. An alternative explanation could be that the MRI-to-CT task leads to more systematic or predictable errors, which makes the uncertainty estimates easier to calibrate and better correlated with actual errors, which may not be the case for CBCT-to-CT.

A critical limitation towards trustworthiness, is that some regions with high prediction error are not reflected in the uncertainty map as can be seen in Figure 4.7. This disconnection between actual error and predicted uncertainty directly undermines the trustworthiness of the model. This omission severely undermines trust in the model, as users cannot rely on the uncertainty signal to identify all areas of failure. Such a failure mode is unacceptable in high-stakes clinical contexts, where undetected errors may have significant consequences. This issue might be partially addressed through the addition of aleatoric uncertainty modeling, especially if the missed high-error zones are caused by inherent data noise.

To mitigate these calibration issues, post-hoc solutions have been explored. Laves et al. (2021) demonstrate the effectiveness of such recalibration methods in the context of medical imaging, improving the alignment between predicted uncertainty and actual model performance.

Lastly, the disconnection between model confidence and real performance has also been observed in other studies. Laves et al. (2021) report overconfidence in MCD-based regression, while Mucsányi et al. (2024) rank MCD as a moderately to poorly calibrated method. Moreover, Kendall and Gal (2017) reported that the epistemic uncertainty, reducible with more observations, could be explained away quicker than expected in some scenario, although this assumption sound less reasonable in the present context due to the complexity of the task. These limitations impact both the interpretability and the trustworthiness of the produced uncertainty maps.

5.1.2 Out Of Distribution detection

Monte Carlo Dropout (MCD) offers a theoretically grounded framework for estimating epistemic uncertainty, as it approximates Bayesian inference by performing multiple stochastic forward passes through the network (Gal et al. 2016). This theoretical foundation contributes to its interpretability. Since the posterior distribution over the model's weights reflects the model's familiarity with the data, it should ideally signal high uncertainty in the presence of previously unseen or out-of-distribution inputs. This is supported by Mucsányi et al. (2024), who report that epistemic methods are particularly effective in detecting OOD samples.

In the present case, MCD's epistemic uncertainty maps successfully identified the main OOD object introduced in the simulated dataset as can be seen in Figures 4.9 and 4.10. However, several additional artefacts, that can be seen as by-products of the data corruption introduced, present in the resulting sCT in Figure 4.10 were not flagged. This indicates that MCD detects prominent OOD structures but may fail to highlight subtler or more complex anomalies. These limitations align with the findings of Ovadia et al. (2019), who observed that Bayesian approaches such as dropout are effective primarily when OOD inputs are substantially different from the training distribution.

While the correct identification of prominent OOD structures supports the interpretability of MCD-based uncertainty maps, this benefit is significantly undermined by the inconsistent detection of other corrupted regions. In clinical settings, it is not sufficient for a model to highlight only the most obvious abnormalities, reliable uncertainty estimation requires that all relevant anomalies be consistently flagged. The fact that several artefacts remain undetected weakens trust in the model's uncertainty output. Users cannot confidently rely on these maps to alert them to potential failures, especially when subtle or localized changes may go unnoticed. Consequently, even though MCD provides interpretable signals in ideal cases, its practical trustworthiness is compromised by this lack of robustness.

5.1.3 Real world setting

Finally, the situation becomes even more problematic when considering real clinical data. Despite extensive data curation, the dataset still presents intrinsic issues such as misregistration and time delays between modalities, which can artificially increase the measured error. In this context, some of the apparent prediction failures might not be due to the model itself but rather to inconsistencies in the reference data. However, even if the true error is sometimes inflated, the fact remains that the model continues to express high confidence in these problematic regions. If uncertainty estimates cannot be reliably evaluated due to such inconsistencies, it becomes difficult to assess whether the model is genuinely robust. It appears that in such conditions, the ability to develop and validate trustworthy uncertainty-aware models is fundamentally compromised.

The calibration curves derived from this setting (Figure 4.25) show degraded performance and increased overconfidence. The same can be said about the correlation measures, which drop from 0.54 to 0.45 (Tables 4.4 and 4.9). This deterioration may partly reflect the dataset issues mentioned above. Nevertheless, this mismatch between predicted uncertainty and actual error further reduces trust in the model's outputs when applied in realistic clinical scenarios. It also raises important questions about the use of such problematic datasets for training robust and trustworthy models.

Moreover, although one salient OOD structure was correctly highlighted in the real dataset (Figure 4.21), this result should be interpreted cautiously. The anomaly in question is visually prominent, and it remains unclear whether less pronounced OOD structures would be equally well detected. This cautiousness is supported by Ovadia et al. (2019), who found that the sample has to be far out of distribution to be detected by Bayesian methods, as highlighted in the precedent section. Cases were observed where high prediction errors occurred without corresponding uncertainty increases (Figure 4.28), further illustrating the misalignment between uncertainty estimates and true model performance. This may result from both MCD overconfidence and the incompleteness of epistemic-only modeling, just as in the simulated setting.

5.1.4 Conclusion

Monte Carlo Dropout offers a theoretically grounded and widely recognized approach to estimating epistemic uncertainty, leveraging approximate Bayesian inference through stochastic forward passes. In controlled or simulated conditions, it demonstrates the ability, although limited, to highlight major out-of-distribution features and provides a marginal improvement in predictive accuracy, which supports its baseline utility in uncertainty-aware modeling.

However, when applied to synthetic CT generation from real data, significant limitations emerge. From an interpretability standpoint, the uncertainty maps exhibit a compressed dynamic range and often fail to align with error-prone regions, particularly in real clinical settings. While some salient OOD structures are flagged, subtler artefacts and complex anatomical noise frequently go undetected. The calibration curves further reveal that the model becomes increasingly overconfident as error increases, which is especially problematic given that trust in uncertainty estimates is most critical in these regions.

In terms of trustworthiness, the moderate correlation between uncertainty and prediction error, coupled with the failure to reflect major errors in the uncertainty maps, undermines confidence in the method. These issues persist across simulated and real datasets. Although part of the observed mismatch may stem from the real dataset limitations, the overall behavior of the model remains inconsistent and unreliable.

These findings raise broader questions about the use of clinically imperfect datasets for developing robust uncertainty aware models. If the uncertainty cannot be meaningfully evaluated due to calibration breakdown or data noise, then both interpretability and validation of robustness become fundamentally compromised.

Overall, despite its theoretical appeal, Monte Carlo Dropout alone does not provide sufficiently reliable or interpretable uncertainty estimates for clinical use in this setting. To overcome these limitations, the following section investigates a combined framework that integrates heteroscedastic aleatoric modeling with Monte Carlo Dropout, aiming to achieve more comprehensive and trustworthy uncertainty estimation.

5.2 Monte Carlo Dropout and Heteroscedasticity

This section evaluates the performance of the combined heteroscedastic and Monte Carlo Dropout framework for uncertainty estimation in synthetic CT generation from CBCT images. The goal is to determine whether jointly modeling aleatoric and epistemic uncertainty leads to more interpretable and trustworthy uncertainty maps compared to epistemic modeling alone. While the approach shows theoretical promise and yields improved predictive performance, several limitations undermine its effectiveness. The persistent entanglement between the two uncertainty components, the poor calibration of the epistemic branch, and the occasional failure to signal high-error regions are the main reasons. These issues limit the framework's interpretability, trustworthiness, and practical suitability for clinical deployment.

5.2.1 Regression Performance

When compared to both the baseline model without uncertainty estimation and the model using only Monte Carlo Dropout, the combined heteroscedastic and Monte Carlo Dropout framework achieves better regression performance (Table 4.10). This observation aligns with the findings of Kendall and Gal (2017), who reported similar behavior across various benchmarking datasets. In their analysis, the improved performance is attributed to two factors: the regularization effect introduced by Dropout during training, and the heteroscedastic loss function, which allows the model to adjust the weighting of errors based on predicted aleatoric uncertainty. Together, these mechanisms encourage better generalization and more accurate predictions. From a trustworthiness standpoint, this improvement supports the reliability of the model's outputs, as it suggests that the inclusion of uncertainty modeling not only provides additional information but also enhances the overall predictive accuracy.

5.2.2 Entanglement

The combined modeling of aleatoric and epistemic uncertainty, as proposed by Kendall and Gal (2017), offers a theoretically complete framework for uncertainty estimation. This approach relies on a heteroscedastic formulation, where aleatoric uncertainty is modeled as a data-dependent variance term directly incorporated into the loss function. Specifically, the model learns to predict a per-voxel variance that reflects the expected noise in each region of the input, allowing it to weight the regression loss accordingly. By combining this with Monte Carlo Dropout to estimate epistemic uncertainty, the framework aims to decompose uncertainty into two interpretable components: one that captures inherent data noise (aleatoric) and one that reflects the model's ignorance (epistemic). In principle, this separation should result in more informative and clinically interpretable uncertainty maps, enabling users to distinguish between uncertainty due to acquisition limitations and that due to unfamiliar or underrepresented anatomical features.

This theoretical framework finds partial support in the results obtained on simulated data. According to the model, aleatoric uncertainty should increase with data noise. As the level of scatter increases, the mean aleatoric uncertainty also increases, from 20.01 to 20.40 HU (Figure 4.14). However, this trend is small in scale and not readily visible in the maps. Moreover, statistical significance was only reached when comparing scatter levels of 30% to

50%, and 40% to 50%, limiting the strength of the conclusion. Thus, while the results point in the expected direction, the practical interpretability gain remains limited.

A more concerning issue is the strong overlap between epistemic and aleatoric uncertainty maps. In both simulated and real settings, the two maps highlight nearly identical regions, including artefacts and OOD samples. This is confirmed by the correlation between both estimates at the voxel level is really strong at 0.9 (Figure 4.13). This behavior suggests that the model fails to disentangle the two sources of uncertainty, which contradicts the theoretical expectation. A possible explanation, as discussed by Mucsányi et al. (2024), is the architectural choice of using a shared U-Net backbone for both branches. Shared feature extraction tends to promote coupling between outputs, undermining the goal of achieving distinct, interpretable representations of aleatoric and epistemic uncertainty. However, this observation also raises some questions about how well the theory actually works in practice. While the framework proposed by Kendall and Gal (2017) is widely accepted, the lack of clear separation seen in this work could also suggest that the way uncertainty is modeled or interpreted is still not fully understood. Even if this explanation seems less likely than the architectural hypothesis, it may still play a role in the entanglement observed.

Nevertheless, this observed entanglement poses a serious challenge for trustworthiness. Whether this problem arises from architectural constraints such as shared feature extraction or from an incompleteness of the theoretical framework, the implications are the same, the resulting uncertainty maps deviate from expected behavior. Without this theoretical coherence, even a well calibrated model cannot be considered fully trustworthy.

5.2.3 Correlation and Calibration

The problem of entanglement becomes even more pronounced when examining the calibration results. Only the aleatoric component is well-calibrated, while the epistemic uncertainty shows poor alignment with actual prediction error and remains overconfident as can be seen in Figure 4.11. This adds a new layer to the interpretability issue. In fact, since the epistemic map is both poorly calibrated and visually similar to a scaled-down version of the aleatoric map, it suggests that the model has not learned to produce two independent forms of uncertainty, joining the observation on overlap.

When looking solely at the aleatoric map, its good calibration and its ability to highlight anatomical edges support the findings of Kendall and Gal (2017), that noted the aleatoric maps to highlight edges of objects where fast change of values in ground truth was noticeable (Figure 4.16). However, this also creates a new risk, namely if only this type of uncertainty is reliable, users might wrongly assume that the combined or epistemic uncertainty maps are just as interpretable. This can make interpretation more difficult and increase the chance of misunderstanding the results, especially in clinical situations where accuracy and clarity are very important. Without a proper discrimination between the two uncertainty types, the overall interpretability of the method becomes much weaker.

Another notable observation concerns the correlation between uncertainty estimates and actual prediction error. In the combined framework, the epistemic uncertainty map exhibits a slightly stronger correlation with reconstruction error compared to its standalone Monte Carlo

Dropout counterpart (Tables 4.4 and 4.6). This improvement may result from the guiding influence of the heteroscedastic loss, which enables the model to focus epistemic uncertainty more specifically on model-driven errors, while allowing aleatoric uncertainty to account for data-related noise ((Kendall and Gal 2017)). This might help the epistemic branch focus more on the model's lack of knowledge instead of confusing it with noisy or unclear data. However, this idea is still uncertain, because, as explained earlier, the overlap is still present. So, even though the stronger correlation with error looks promising, it should be viewed with caution, since it probably doesn't mean that the model truly separates the two types of uncertainty.

Building on the previous observation that the Monte Carlo Dropout only model often exhibits high confidence in regions with substantial prediction error, similar limitation emerges in the combined framework. Although the aleatoric component is now well-calibrated, it still fails to flag certain high-error regions, as shown in Figure 4.15. This mismatch introduces a critical trust issue: the improved calibration may give clinicians a false sense of security, encouraging trust in predictions that remain unreliable in key areas. This mismatch represents one of the most critical limitations towards trust in the framework, just like the MCD-only model. Although in the present framework, the better calibration could let clinician trust more easily a model that fails in the same way.

Recent works such as Stirn, Wessels, et al. (2022) and Seitzer et al. (2022) have proposed alternative loss functions and training strategies to help models disentangle and calibrate these uncertainty types more effectively. While not evaluated in the present work, such approaches may offer promising directions for mitigating the entanglement and miscalibration issues observed here.

5.2.4 Out Of Distribution detection

Compared to the MCD-only model, OOD detection appears to improve within the combined heteroscedastic and MCD framework, as all introduced anomalies are successfully flagged in the uncertainty maps (Figures 4.18 and 4.17). This marks a notable gain in interpretability: the model produces more comprehensive uncertainty responses, including subtler artefacts that were missed in the epistemic only setting. However, a closer examination reveals that this improved detection is primarily driven by the aleatoric component, which responds most visibly to OOD structures, particularly by outlining their borders. This observation mitigate once again the gain in interpretability.

This outcome contrasts with theoretical expectations, which states that epistemic uncertainty should serve as the principal signal for unfamiliar inputs. A likely explanation for this discrepancy lies in the poor calibration of the epistemic component. As shown earlier, epistemic uncertainty remains systematically underestimated, which may visually suppress its contribution in the combined map and lead to biased interpretation. Although the model now detects OOD content more reliably, this response is predominantly attributed to the aleatoric component, which challenges the theoretical assumptions underlying the intended decomposition of uncertainty.

As a result, trustworthiness remains partially compromised. The lack of clear separation between uncertainty types not only undermines interpretability by blurring the meaning of the signals but also challenges trust, since users may misattribute the source of uncertainty. If epistemic uncertainty cannot be relied upon to flag unfamiliar inputs, and if aleatoric uncertainty absorbs that role without clear justification, then the theoretical guarantees of the framework are not met. Therefore, while the combined model shows encouraging progress in OOD detection coverage, it still falls short of delivering the type of uncertainty decomposition necessary for trustworthy clinical deployment.

5.2.5 Real world setting

Finally, the results obtained on the real processed dataset confirm the same issues observed in the simulated setting. The calibration curves show a similar pattern, where only the aleatoric component is well calibrated, while the epistemic estimate continues to display overconfidence and poor alignment with prediction error as can be seen in Figure 4.11. This consistency across both settings reinforces the concern that the model has not learned to separate the two uncertainty types. As a result, the same interpretability risks apply.

It also however introduces some notable observations. The uncertainty maps often provide visually meaningful insights, with stronger correlation coefficients observed between predicted uncertainty and actual error (Table 4.11). Although this improvement may be partially influenced by the known problems of the dataset, the increase in correlation remains encouraging and could indicate improved alignment with prediction error. Most notably, the aleatoric component appears to flag certain high-error regions as can be seen in Figure 4.34, consistent with theoretical expectations regarding inherent uncertainty. However, this detection remains overconfident despite its calibration, limiting its practical trustworthiness.

Furthermore, the epistemic component successfully identifies the OOD structure present in the real data and appears more visually calibrated than in the simulated setting (Figure 4.35). This finding aligns with theoretical predictions and echoes the observations made by Ovadia et al. (2019). Nevertheless, caution remains necessary, the OOD anomaly in the real dataset is particularly evident, and it remains unclear whether more subtle anomalies would be similarly flagged, similarly as in the MCD-only experiment. Although the real dataset has some well-known limitations, it still seems to partly support the theory behind separating uncertainty types. Because of this difference, the real dataset might be even more interesting to study, as it could help understand when and why the theory works.

5.2.6 Conclusion

The integration of heteroscedastic aleatoric modeling with Monte Carlo Dropout aims to provide a more comprehensive and theoretically grounded framework for uncertainty estimation, building on the formulation proposed by Kendall and Gal (2017). By modeling aleatoric uncertainty through a data-dependent variance term and epistemic uncertainty via stochastic sampling, the approach aspires to disentangle distinct sources of uncertainty: inherent data noise and model ignorance. This dual modeling strategy holds strong theoretical appeal, particularly in clinical applications where distinguishing between these sources is critical for risk assessment and decision-making.

Experimental results offer partial support for this framework. On simulated data, aleatoric uncertainty increases with added noise as expected, and the model shows a slight improvement in error correlation and out of distribution detection compared to the MCD-only baseline. Moreover, the aleatoric maps are well-calibrated and delineate anatomical structures clearly, providing interpretable and visually intuitive outputs in both synthetic and real data settings.

However, the framework interpretability remains limited by the persistent entanglement between aleatoric and epistemic estimates that violates theory. The two uncertainty components exhibit a strong correlation, lowering the fact that each should capture a distinct type of uncertainty. Epistemic uncertainty, in particular, remains poorly calibrated and resembles a scaled down version of the aleatoric map. This behavior suggests that the model struggles to meaningfully separate uncertainty sources, possibly due to architectural choices such as shared feature extraction or due to some training choice as the loss function. As a result, the interpretability of the uncertainty maps is compromised.

From a trustworthiness perspective, the framework does yield a measurable improvement in predictive performance, supporting its utility in enhancing output quality. The heteroscedastic loss function and Dropout regularization appear to promote better generalization, which aligns with prior findings in the literature. However, the failure to detect certain high-error regions despite otherwise good calibration for the aleatoric estimate poses a serious limitation. In particular, clinicians may be falsely reassured by low uncertainty estimates in areas where the model performs poorly, which directly contradicts the intended role of uncertainty as a risk signal. This represents a critical trust failure.

The detection of out of distribution structures is another area where the framework demonstrates partial success. While all OOD inputs are flagged, the dominant contribution stems from the aleatoric component, contrary to theoretical expectations. This discrepancy may result from the underestimation of epistemic uncertainty, which reduces its visibility in the uncertainty maps. Although this behavior still provides useful information, it raises concerns about whether the model's uncertainty decomposition is functioning as intended.

Within the real setting with preprocessed real data, the framework replicates most patterns seen in the simulated setting. Uncertainty maps generally align with anatomical variability and error-prone regions, and both types of uncertainty exhibit better correlation with actual error. Notably, the aleatoric component occasionally succeeds in highlighting high-error zones, and the epistemic component shows improved OOD detection behavior. While these findings are encouraging, they should be interpreted with caution, given the limitations of the dataset and the obviousness of the chosen OOD anomaly. Nonetheless, the fact that the real dataset partially validates the theoretical framework unlike the MCD-only model suggests it could serve as a valuable basis for future refinement of uncertainty aware methods.

In conclusion, the combined Heteroscedastic and Monte Carlo Dropout framework offers a promising direction for uncertainty modeling, with clear improvements over purely epistemic approaches. However, its failure to disentangle uncertainty sources and its occasional inability to reflect true prediction errors makes it impossible to recommend it for clinical deployment at this stage. Until these core limitations are addressed, the framework cannot be considered an interpretable and trustworthy solution for the CBCT to CT task.

5.3 Methodology Limitations

Although this study benefits from the use of two complementary datasets and a carefully designed preprocessing pipeline, several limitations remain. The real dataset, SynthRad2023, is limited in anatomical and demographic diversity. It focuses solely on pelvis regions, and within the pelvis subset, there is a strong gender imbalance due to the predominance of prostate cancer cases. The lack of extended metadata beyond age and sex also restricts the ability to analyze potential biases or stratify performance across relevant subgroups. Despite extensive preprocessing efforts, including deformable registration and intensity correction, certain imperfections such as residual misalignments, anatomical inconsistencies between CT and CBCT scans, and implant-related artefacts persist. These issues introduce noise that may interfere with both training stability and the interpretability of uncertainty estimates, especially in regions where differences in metrics can be inflated by these problems.

In parallel, while the simulated dataset provides a controlled and idealized environment free from misregistration, anatomical mismatches, or acquisition noise, it inevitably simplifies the complexity of real CBCT scans. For example, some elements such as the patient couch have been omitted, and other subtle scanner related artefacts may not be fully reproduced. These omissions represent potential blind spots in the current analysis. Nonetheless, the use of this dataset is justified as a necessary first step. If the proposed uncertainty framework fails to produce disentangled and calibrated estimates in this simplified setting, it is unlikely to succeed under real-world conditions.

Additionally, the use of Monte Carlo Dropout for modeling epistemic uncertainty is subject to architectural and hyperparameter choices. While dropout was applied following existing literature, its placement and rate were not exhaustively optimized. This means that the uncertainty estimates may not be fully optimal. In fact, The Bayesian SegNet (Kendall, Badrinarayanan, et al. 2015), the selected architecture, was developed for segmentation. Hence, the dropout placement may not be optimal in regards to regression application as the task at hand.

In addition, the use of the segmentation_models_pytorch library imposes constraints on architectural customization. While the library provides a streamlined interface for deploying standard encoder-decoder models like U-Net, it offers limited flexibility in modifying the internal topology of the network. This restricts the ability to introduce architectural innovations or adjustments, such as altering skip connections. As a result, the model architecture remains a relatively standard U-Net.

Finally, the architecture proposed may not be the most adapted for the task at hand. Other work such as the one of Thummerer, Zaffino, et al. (2020), not only compare a U-Net but also other architectures such as a cycle-GAN or a vision transformer (Hu et al. 2024). This can be a problem as the proposed architecture might not be the right fit for the uncertainty estimation task at hand.

5.4 Overall Discussion Conclusion

This thesis set out to answer the following research question: *Can state-of-the-art deep learning-based uncertainty quantification methods provide trustworthy and interpretable outputs when applied to synthetic CT generation from CBCT images?* To explore this, two modeling approaches were investigated: Monte Carlo Dropout alone, and the combination of MCD with heteroscedastic modeling. Both were assessed with respect to two central criteria from the research question: interpretability and trustworthiness of the resulting uncertainty maps and every observation made is compiled in Table 5.1.

When using Monte Carlo Dropout alone, the results show that while the method improves predictive accuracy slightly and can detect clear OOD anomalies in the simulated setting, it suffers from serious limitations. The uncertainty maps are only moderately correlated with prediction error and remain poorly calibrated, particularly in regions with high error where uncertainty should be greatest. In the real dataset, performance degrades further, with visible overconfidence and reduced alignment between uncertainty and error. Importantly, the epistemic uncertainty fails to flag subtle artefacts. These limitations suggest that MCD alone, despite its theoretical grounding, does not meet the standards of interpretability or trustworthiness required for clinical use.

The second approach combined MCD with heteroscedastic modeling to estimate both epistemic and aleatoric uncertainty. This framework was designed to disentangle uncertainty due to model ignorance from uncertainty arising from noisy or variable data. On paper, this formulation should produce more informative and interpretable maps. In practice, it provides better predictive performance, improved detection of OOD regions, and well-calibrated aleatoric uncertainty. However, it still fails to achieve the intended separation between the two uncertainty types. The epistemic and aleatoric maps are highly correlated and visually similar, and the epistemic branch remains poorly calibrated. This entanglement reduces the interpretability of the model and limits its ability to signal when predictions should not be trusted. In some cases, regions with high error are still assigned low uncertainty, undermining the model's reliability. Although this framework outperforms MCD alone, it remains insufficiently robust for clinical deployment in its current form.

Following this model analysis, several methodological limitations have been acknowledged. The real dataset used is restricted in anatomical variety and demographic balance. It focuses only on pelvic anatomy, with an over representation of male patients. Despite careful preprocessing, residual artefacts such as misalignments and imaging noise persist, possibly distorting both training and evaluation of uncertainty maps. These imperfections make it difficult to interpret whether observed uncertainty behaviors truly reflect model capability or data quality issues.

The simulated dataset, while useful for controlled experiments, lacks the full complexity of clinical CBCT imaging. It omits certain realistic components such as the patient couch and may not fully replicate subtle anatomical or acquisition artefacts. Insights gained in this controlled setting may not generalize to real-world applications.

Additionally, the choice of architecture, based on U-Net and constrained by external library dependencies, limits architectural flexibility. Dropout placement and rate were not extensively optimized, and the use of segmentation-derived architectures for regression tasks may not be optimal. These constraints likely contribute to the model's limited ability to separate and calibrate uncertainty components.

In conclusion and based on the results, the answer to the research question is "no". The tested approaches, Monte Carlo Dropout and heteroscedastic modeling, do show some benefits, such as slight improvements in prediction accuracy and the ability to highlight certain out of distribution regions. However, these methods do not consistently produce uncertainty maps that are reliable or easy to interpret. They often fail to signal subtle artefacts and tend to be overconfident in the regions where uncertainty should be highest. As a result, they cannot yet be considered suitable for clinical use. Even so, the findings highlight promising directions for future research. Improvements in model architecture could lead to better separation of uncertainty types, and more advanced calibration techniques may help make uncertainty estimates more accurate and clinically meaningful.

Criterion	MCD	MCD & Heteroscedasticity
Modelled Uncertainty	Epistemic	Epistemic And Aleatoric
Trustworthiness	+ Slight improvement in regression performance + Detects salient OOD regions + Possible post-hoc recalibration − Poor calibration, especially in high-error zones − Moderate correlation with prediction error (≈0.45–0.54) − Fails to flag subtle errors or artefacts	+ Theoretical Bayesian foundation + Well-calibrated aleatoric uncertainty + Improved OOD coverage − Entangled uncertainty components − Epistemic branch remains poorly calibrated − Some high-error regions not flagged despite calibration − False reassurance risk due to aleatoric dominance
Interpretability	+ Detects large OOD anomalies + Provides some visual uncertainty cues − Uncertainty map behaves like a compressed error map − theoretical separation between uncertainty types overlooked (epistemic only modeling)	+ Aleatoric maps clearly delineate anatomical structures + Better visual clarity in some clinical scenarios + Slight improvement in uncertainty–error correlation − Strong correlation between epistemic and aleatoric estimates (≈0.9) − Epistemic map visually mimics aleatoric map − Theoretical uncertainty disentanglement not observed in practice

Table 5.1. Summary of the comparative evaluation of Monte Carlo Dropout (MCD) and the combined Heteroscedastic and MCD framework with respect to the interpretability and trustworthiness of their uncertainty estimates in the context of synthetic CT generation from CBCT. While MCD provides marginal improvements in prediction accuracy and detects salient out of distribution anomalies in controlled settings, it suffers from poor calibration, moderate error correlation, and unreliable uncertainty estimates in high-error regions. The combined framework improves regression performance and aleatoric calibration, but fails to disentangle uncertainty types and remains limited by entanglement and epistemic overconfidence. Both methods ultimately fall short of providing robust and clinically interpretable uncertainty maps, though the combined approach highlights promising directions for future research in uncertainty aware modeling.

Chapter 6

Conclusion and perspectives

This thesis addresses the robustness and reliability of deep learning-based synthetic CT generation from cone-beam CT images, with a particular focus on the evaluation of uncertainty estimation methods. The work investigates whether state of the art Bayesian uncertainty quantification strategies can provide interpretable and trustworthy outputs in the context of proton therapy. To this end, two complementary uncertainty modeling frameworks are explored: epistemic uncertainty via Monte Carlo Dropout and combined uncertainty through the integration of MCD with a heteroscedastic loss function.

The study relies on two datasets designed to support complementary aspects of the analysis. The first is a simulated dataset specifically constructed to closely mimic the characteristics of IBA's clinical CBCT acquisitions. While it enables controlled experimentation, it remains an approximation and includes several simplifications. Notably, features such as the treatment couch are omitted, and the data generation process requires extensive hyperparameter tuning, which may limit its representativeness. The second dataset consists of real clinical data from the SynthRad2023 Task 2 challenge, offering a more authentic but substantially noisier environment. This dataset is affected by significant misalignment and variations in acquisition timing between CBCT and CT scans. Despite efforts to correct this, residual registration artefacts and inconsistencies persist.

6.1 Conclusion

Firstly, the uncertainty estimation methods were tested on a simulated dataset, allowing for controlled analysis. The baseline U-Net, without uncertainty modeling, significantly outperformed both the water baseline and raw CBCT. Adding Monte Carlo Dropout improved performance slightly at a 30% dropout rate and produced uncertainty maps that roughly matched the prediction error and preserved anatomical details. However, these maps were often overconfident and poorly calibrated, in such a way that large prediction errors are not translated to similarly large uncertainty values. Adding aleatoric uncertainty through a heteroscedastic model led to small improvements in accuracy and significantly better calibration, particularly for the aleatoric component. These maps were more informative and better at highlighting unusual areas like metal artefacts or artificial distortions. Still, the epistemic and aleatoric uncertainties were highly correlated, suggesting redundancy and poor disentanglement between the two types.

Secondly, the models were tested on a processed version of the SynthRad2023 dataset, which includes real patient data. The basic U-Net, without uncertainty modeling, performed worse than on the simulated data due to real-world artefacts and inconsistencies between CBCT and ground truth. Adding Monte Carlo Dropout gave only slight improvements in accuracy and showed limited ability to highlight high-error areas, often producing overconfident uncertainty maps. However, it was still able to detect some unusual features, like metal implants. The heteroscedastic U-Net, which models both epistemic and aleatoric uncertainties, improved the correlation between predicted uncertainty and actual error, especially for the aleatoric part, which was also better calibrated. Still, the epistemic and aleatoric maps were very similar, meaning they didn't provide clearly separate information.

The discussion is built upon on these results in order to assess whether the uncertainty quantification strategies explored in this thesis satisfy the criteria of interpretability and trustworthiness, as defined in the research question. Across both datasets, the evaluation showed that Monte Carlo Dropout provided limited practical benefits: although it offered some correlation with prediction error and could flag certain artefacts or outliers, its uncertainty maps were overconfident and poorly calibrated, particularly in challenging conditions. These problems were more pronounced in the real dataset, where noise and anatomical inconsistencies reduced both predictive accuracy and the reliability of uncertainty estimates. The extended framework combining MCD with heteroscedastic modeling improved overall calibration and better highlighted regions of potential error, especially through the aleatoric component. Nevertheless, the lack of clear separation between epistemic and aleatoric uncertainties persisted, limiting the interpretability of the uncertainty maps. These findings suggest that while progress has been made toward generating uncertainty aware synthetic CTs, current methods do not have the trustworthiness and interpretability needed for clinical integration without complementary measures.

6.2 Perspectives

Based on the findings of this thesis, several directions can enhance uncertainty quantification for synthetic CT generation. Optimizing Monte Carlo Dropout, particularly its rate and placement, could improve epistemic calibration, as current configurations may be suboptimal for regression tasks. The observed overlap between aleatoric and epistemic estimates highlights the need for architectures that explicitly separate these sources, such as dual-branch designs, combined with specialized loss functions designed to enhance uncertainty calibration and disentanglement.

Another path for future research is to investigate the specific role and limitations of each dataset type in relation to clinical applications. The synthetic dataset, while useful for controlled experimentation, may be overly simplistic, and its relevance to real world scenarios remains to be validated. Additionally, the real dataset, despite its clinical origin, includes imperfections such as misregistration and acquisition inconsistencies, which distort both training and evaluation. To which extent is that a problem for clinical applications remains an open question.

Another important direction would be to move beyond image-level error correlation and assess the clinical impact of uncertainty through its relationship with dose prediction. For example, by investigating the correlation between predicted uncertainty and dose error, rather than voxel-wise image error alone. This approach aligns more directly with the goals of proton therapy, where accurate dose delivery is the primary concern. Since proton dose deposition is highly sensitive to tissue stopping powers, which in turn can be derived from the Water Equivalent Thickness (WET), evaluating how uncertainty correlates with WET error could provide a meaningful surrogate for assessing its impact on treatment accuracy. Thus, future work could explore whether uncertainty estimates reliably track WET variations, offering a more clinically relevant validation of their utility in proton therapy planning.

Declaration on AI-assisted technologies in the writing process

While this thesis represents original research and contributions, the author acknowledge having used ChatGPT, a large language model developed by OpenAI, to refine and improve the English language throughout the manuscript. As English is not the author native language, this tool was employed to enhance clarity, grammar, and readability.

All content generated with the assistance of this tool was carefully reviewed, revised, and edited by the authors. Full responsibility for the ideas, structure, and final content of this thesis is acknowledged by the author.

Bibliography

Abdar, Mohammad et al. (Dec. 2021). "A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges". In: *Information Fusion* 76, pp. 243–297. DOI: 10.1016/j.inffus.2021.05.008.

Aljaafari, Nura (Feb. 2018). "Ichthyoplankton Classification Tool using Generative Adversarial Networks and Transfer Learning". PhD thesis. DOI: 10.25781/KAUST-K902H.

Amini, Amir et al. (2020). "Deep Evidential Regression". In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., pp. 14927–14937. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/aab085461de182608ee9f607f3f7d18f-Paper.pdf.

Ayhan, Murat Seckin and Philipp Berens (2018). "Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks". In: *Medical Imaging with Deep Learning*. URL: https://openreview.net/forum?id=rJZz-knjz.

Barber, David and Christopher Bishop (1997). "Ensemble Learning for Multi-Layer Networks". In: *Advances in Neural Information Processing Systems*. Ed. by M. Jordan, M. Kearns, and S. Solla. Vol. 10. MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/1997/file/e816c635cad85a60fabd6b97b03cbcc9-Paper.pdf.

BBC News (July 2015). "Google apologises for Photos app's racist blunder". Accessed: 2025-05-10. URL: https://www.bbc.com/news/technology-33347866.

– (Oct. 2018). "Amazon scrapped 'sexist AI' tool". Accessed: 2025-05-10. URL: https://www.bbc.com/news/technology-45809919.

Biewald, Lukas (2020). "Experiment Tracking with Weights and Biases". Software available from wandb.com. URL: https://www.wandb.com/.

Bishop, Christopher M (1994). "Mixture density networks". Technical Report NCRG-94-004. Birmingham, UK: Aston University.

Blundell, Charles et al. (July 2015). "Weight Uncertainty in Neural Network". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 1613–1622. URL: https://proceedings.mlr.press/v37/blundell15.html.

Chen, Tianqi, Emily Fox, and Carlos Guestrin (June 2014). "Stochastic Gradient Hamiltonian Monte Carlo". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research

2. Bejing, China: PMLR, pp. 1683–1691. URL: https://proceedings.mlr.press/v32/cheni14.html.

Delporte, Guillaume (2025). "Master Thesis Code Repository". GitHub repository, accessed May 2025. URL: https://github.com/bahboul42/master-thesis.

Depeweg, Stefan et al. (July 2018). "Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1184–1193. URL: https://proceedings.mlr.press/v80/depeweg18a.html.

Endoblog, The (Feb. 2011). "Cone Beam Computed Tomography (CBCT)". URL: http://www.theendoblog.com/2011/02/cone-beam-computed-tomography-cbct.html.

Fabibombo (2023). "FLARACC_sCT_Pipeline: CBCT to CT Translation". https://github.com/fabibombo/FLARACC_sCT_Pipeline/tree/main/cbct2ct_translation. Accessed: 2025-05-28.

Gal, Yarin and Zoubin Ghahramani (June 2016). "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, pp. 1050–1059. URL: https://proceedings.mlr.press/v48/gal16.html.

Galapon Jr, Arthur Villanueva et al. (2024). "Feasibility of Monte Carlo dropout-based uncertainty maps to evaluate deep learning-based synthetic CTs for adaptive proton therapy". In: *Medical Physics* 51.4, pp. 2499–2509.

Gao, Junyu et al. (2024). "A Comprehensive Survey on Evidential Deep Learning and Its Applications". arXiv: 2409.04720 [cs.LG]. URL: https://arxiv.org/abs/2409.04720.

Ghaznavi, Hamid et al. (2025). "Quantitative use of cone-beam computed tomography in proton therapy: challenges and opportunities". In: *Physics in Medicine & Biology* 70.9, 09TR01.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy (2015). "Explaining and Harnessing Adversarial Examples". arXiv: 1412.6572 [stat.ML]. URL: https://arxiv.org/abs/1412.6572.

Graves, Alex (2011). "Practical Variational Inference for Neural Networks". In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2011/file/7eb3c8be3d411e8ebfab08eba5f49632-Paper.pdf.

Greenway, Knipe, Gaillard, et al. (2025). "Hounsfield unit". Reference article, Radiopaedia.org. Accessed on 14 May 2025. URL: https://doi.org/10.53347/rID-38181.

Hansen, Lars Kai and Peter Salamon (1990). "Neural network ensembles". In: *IEEE transactions on pattern analysis and machine intelligence* 12.10, pp. 993–1001.

He, Bobby, Balaji Lakshminarayanan, and Yee Whye Teh (2020). "Bayesian Deep Ensembles via the Neural Tangent Kernel". In: *Advances in Neural Information Process-*

ing Systems. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 1010–1022. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/0b1ec366924b26fc98fa7b71a9c249cf-Paper.pdf.

He, Kaiming et al. (2016). “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

Hsieh, Jiang (2015). “Computed tomography: principles, design, artifacts, and recent advances”. Bellingham, Washington: SPIE.

Hu, Yuxin et al. (2024). “Synthetic CT generation based on CBCT using improved vision transformer CycleGAN”. In: *Scientific Reports* 14.1, p. 11455.

Huijben, Emma M. C. et al. (Oct. 2024). “Generating synthetic computed tomography for radiotherapy: SynthRAD2023 challenge report”. In: *Medical Image Analysis* 97, p. 103276. DOI: 10.1016/j.media.2024.103276.

Hüllermeier, Eyke and Willem Waegeman (2021). “Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods”. In: *Machine learning* 110.3, pp. 457–506.

Immer, Alexander et al. (2023). “Effective Bayesian Heteroscedastic Regression with Deep Neural Networks”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=A6EquH0enk.

Janssens, Guillaume et al. (2011). “Diffeomorphic Registration of Images with Variable Contrast Enhancement”. In: *International Journal of Biomedical Imaging* 2011.1, p. 891585. DOI: https://doi.org/10.1155/2011/891585. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2011/891585. URL: https://onlinelibrary.wiley.com/doi/abs/10.1155/2011/891585.

Jsang, Audun (2018). “Subjective Logic: A formalism for reasoning under uncertainty”. Springer Publishing Company, Incorporated.

Kendall, Alex, Vijay Badrinarayanan, and Roberto Cipolla (2015). “Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding”. In: *arXiv preprint arXiv:1511.02680*.

Kendall, Alex and Yarin Gal (2017). “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?” In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., pp. 5574–5584. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.

Kingma, Diederik P and Jimmy Ba (2014). “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980*.

Kiureghian, A. D. and O. Ditlevsen (Mar. 2009). “Aleatory or Epistemic? Does It Matter?” In: *Structural Safety* 31.2, pp. 105–112. DOI: 10.1016/j.strusafe.2008.06.020.

Klein, S. et al. (Jan. 2010). “elastix: A Toolbox for Intensity-Based Medical Image Registration”. In: *IEEE Trans. Med. Imaging* 29.1, pp. 196–205. DOI: 10.1109/TMI.2009.2035616.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell (2017). “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles”. In: *Advances in Neural Information Processing Systems* 30.

Lambrou, A., H. Papadopoulos, and A. Gammerman (Jan. 2011). “Reliable Confidence Measures for Medical Diagnosis With Evolutionary Algorithms”. In: *IEEE Transactions on Information Technology in Biomedicine* 15.1, pp. 93–99. DOI: 10.1109/TITB.2010.2091144.

Landry, Guillaume and Chia-ho Hua (2018). “Current state and future applications of radiological image guidance for particle therapy”. In: *Medical Physics* 45.11, e1086–e1095. DOI: https://doi.org/10.1002/mp.12744. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.12744. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.12744.

Lapen, Kaitlyn and Yoshiya Yamada (Apr. 2023). “The Development of Modern Radiation Therapy”. In: *Current Physical Medicine and Rehabilitation Reports* 11, pp. 1–8. DOI: 10.1007/s40141-023-00395-6.

Laves, Max-Heinrich et al. (2021). “Recalibration of Aleatoric and Epistemic Regression Uncertainty in Medical Imaging”. arXiv: 2104.12376 [eess.IV]. URL: https://arxiv.org/abs/2104.12376.

LeCun, Yann et al. (1998). “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Leibig, Christian et al. (Dec. 2017). “Leveraging Uncertainty Information from Deep Neural Networks for Disease Detection”. In: *Sci Rep* 7.1, p. 17816. DOI: 10.1038/s41598-017-17876-z.

Liu, J. et al. (2020). “Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness”. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., pp. 7498–7512. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/543e83748234f7cbab21aa0ade66565f-Paper.pdf.

Louppe, Gilles (2024). “INFO8010 - Deep Learning”. Course materials, University of Liege. Available at: https://github.com/glouppe/info8010-deep-learning.

MacKay, David J. C. (May 1992). “A Practical Bayesian Framework for Backpropagation Networks”. In: *Neural Computation* 4.3, pp. 448–472. DOI: 10.1162/neco.1992.4.3.448.

Malinin, Andrey and Mark Gales (2018). “Predictive uncertainty estimation via prior networks”. In: *Advances in neural information processing systems* 31.

McCulloch, Warren S and Walter Pitts (1943). “A logical calculus of the ideas immanent in nervous activity”. In: *The bulletin of mathematical biophysics* 5, pp. 115–133.

McGowan, S. E., N. G. Burnet, and A. J. Lomax (Jan. 2013). “Treatment Planning Optimisation in Proton Therapy”. In: *Br J Radiol* 86.1021, p. 20120288. DOI: 10.1259/bjr.20120288.

Miyato, Takeru et al. (2018). “Spectral Normalization for Generative Adversarial Networks”. arXiv: 1802.05957 [cs.LG]. URL: https://arxiv.org/abs/1802.05957.

Mucsányi, Bálint, Michael Kirchhof, and Seong Joon Oh (2024). “Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks”. In: *Advances in neural information processing systems* 37, pp. 50972–51038.

Mukhoti, Jishnu et al. (2021). “Deterministic Neural Networks with Appropriate Inductive Biases Capture Epistemic and Aleatoric Uncertainty”. In: *CoRR* abs/2102.11582. arXiv: 2102.11582. URL: https://arxiv.org/abs/2102.11582.

Neal, Radford M et al. (2011). “MCMC using Hamiltonian dynamics”. In: *Handbook of markov chain monte carlo* 2.11, p. 2.

Neal, Radford M (2012). “Bayesian learning for neural networks”. Vol. 118. Springer Science & Business Media.

Nix, David A and Andreas S Weigend (1994). “Estimating the mean and variance of the target probability distribution”. In: *Proceedings of 1994 ieee international conference on neural networks (ICNN’94)*. Vol. 1. IEEE, pp. 55–60.

Ovadia, Yaniv et al. (2019). “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift”. arXiv: 1906.02530 [stat.ML]. URL: https://arxiv.org/abs/1906.02530.

Paganetti, Harald et al. (2021). “Proton Therapy Physics”. In: *Phys. Med. Biol.* 66.22, 22TR01. DOI: 10.1088/1361-6560/ac344f.

Pan, Jingshan et al. (Apr. 2022). “Enhanced FCN for farmland extraction from remote sensing image”. In: *Multimedia Tools and Applications* 81, pp. 1–28. DOI: 10.1007/s11042-022-12141-6.

Papadopoulos, Harris (2008). “Inductive conformal prediction: Theory and application to neural networks”. In: *Tools in artificial intelligence*. Citeseer.

Pearce, Tim, Felix Leibfried, and Alexandra Brintrup (Aug. 2020). “Uncertainty in Neural Networks: Approximately Bayesian Ensembling”. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 234–244. URL: https://proceedings.mlr.press/v108/pearce20a.html.

Prince, Jerry and Jonathan Links (2006). “Medical imaging signals and systems”. Vol. 37. Pearson Prentice Hall Upper Saddle River.

Quiñonero-Candela, Joaquin (2009). “Dataset shift in machine learning”. The MIT Press.

Rasmussen, Carl Edward (2003). “Gaussian processes in machine learning”. In: *Summer school on machine learning*. Springer, pp. 63–71.

Ritter, Hippolyt, Aleksandar Botev, and David Barber (2018). “A scalable laplace approximation for neural networks”. In: *6th international conference on learning representations, ICLR 2018-conference track proceedings*. Vol. 6. International Conference on Representation Learning.

Romano, Yaniv, Evan Patterson, and Emmanuel Candes (2019). “Conformalized quantile regression”. In: *Advances in neural information processing systems* 32.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, pp. 234–241.

Rosenblatt, Frank (1958). “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6, p. 386.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). “Learning representations by back-propagating errors”. In: *nature* 323.6088, pp. 533–536.

Rupprecht, Christian et al. (2017). “Learning in an Uncertain World: Representing Ambiguity Through Multiple Hypotheses”. arXiv: 1612.00197 [cs.CV]. URL: https://arxiv.org/abs/1612.00197.

Schulze, R. et al. (July 2011). “Artefacts in CBCT: a review”. In: *Dentomaxillofacial Radiology* 40.5, pp. 265–273. DOI: 10.1259/dmfr/30642039.

Seitzer, Maximilian et al. (2022). “On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks”. arXiv: 2203.09168 [cs.LG]. URL: https://arxiv.org/abs/2203.09168.

Sensoy, Murat, Lance Kaplan, and Melih Kandemir (2018). “Evidential Deep Learning to Quantify Classification Uncertainty”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/a981f2b708044d6fb4a71a1463242520-Paper.pdf.

Shafer, Glenn and Vladimir Vovk (2008). “A tutorial on conformal prediction.” In: *Journal of Machine Learning Research* 9.3.

Shanmugam, Divya et al. (Oct. 2021). “Better Aggregation in Test-Time Augmentation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1214–1223.

Skafte, Nicki, Martin Jørgensen, and Søren Hauberg (2019). “Reliable training and estimation of variance networks”. In: *Advances in Neural Information Processing Systems* 32.

Srivastava, Nitish et al. (2014). “Dropout: a simple way to prevent neural networks from overfitting”. In: *The journal of machine learning research* 15.1, pp. 1929–1958.

Stirn, Andrew and David A. Knowles (2020). “Variational Variance: Simple and Reliable Predictive Variance Parameterization”. In: *CoRR* abs/2006.04910. arXiv: 2006.04910. URL: https://arxiv.org/abs/2006.04910.

Stirn, Andrew, Hans-Hermann Wessels, et al. (2022). “Faithful Heteroscedastic Regression with Neural Networks”. arXiv: 2212.09184 [cs.LG]. URL: https://arxiv.org/abs/2212.09184.

Thummerer, Adrian, Erik van der Bijl, Arthur Galapon, et al. (Apr. 2023). “SynthRAD2023 Challenge design: Synthesizing computed tomography for radiotherapy”. DOI: 10.5281/zenodo.7781049. URL: https://doi.org/10.5281/zenodo.7781049.

Thummerer, Adrian, Erik van der Bijl, Arthur Galapon Jr, et al. (June 2023). “SynthRAD2023 Grand Challenge dataset: Generating synthetic CT for radiotherapy”. In: *Medical Physics* 50.7, pp. 4664–4674. ISSN: 2473-4209. DOI: 10.1002/mp.16529. URL: http://dx.doi.org/10.1002/mp.16529.

Thummerer, Adrian, Paolo Zaffino, et al. (2020). “Comparison of CBCT based synthetic CT methods suitable for proton dose calculations in adaptive proton therapy”. In: *Physics in Medicine & Biology* 65.9, p. 095002.

Timm (2020). “Pytorch Image Models (timm) | timmdocs”. URL: https://timm.fast.ai/.

Tuna, Ömer, Ferhat Ozgur Catak, and Taner Eskil (Apr. 2022). “Uncertainty as a Swiss army knife: new adversarial attack and defense ideas based on epistemic uncertainty”. In: *Complex & Intelligent Systems* 9. DOI: 10.1007/s40747-022-00701-0.

U.S. Department of Transportation, National Highway Traffic Safety Administration (Jan. 2017). “PE 16-007”. Technical report. NHTSA.

Ulmer, Dennis, Christian Hardmeier, and Jes Frellsen (2023). “Prior and Posterior Networks: A Survey on Evidential Deep Learning Methods For Uncertainty Estimation”. arXiv: 2110.03051 [cs.LG]. URL: https://arxiv.org/abs/2110.03051.

Van Amersfoort, Joost et al. (July 2020). “Uncertainty Estimation Using a Single Deep Deterministic Neural Network”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 9690–9700. URL: https://proceedings.mlr.press/v119/van-amersfoort20a.html.

Veiga, Catarina et al. (2016). “First clinical investigation of cone beam computed tomography and deformable registration for adaptive proton therapy for lung cancer”. In: *International Journal of Radiation Oncology* Biology* Physics* 95.1, pp. 549–559.

Wang, Zhou et al. (2004). “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing* 13.4, pp. 600–612.

Welling, Max and Yee W Teh (2011). “Bayesian learning via stochastic gradient Langevin dynamics”. In: *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer, pp. 681–688.

Yakubovskiy, Pavel (2020). “Segmentation Models Pytorch”. https://github.com/qubvel/segmentation_models.pytorch.

Yang, F. et al. (Jan. 2009). “Using Random Forest for Reliable Classification and Cost-Sensitive Learning for Medical Diagnosis”. In: *BMC Bioinformatics* 10.1, S22. DOI: 10.1186/1471-2105-10-S1-S22.

Zhang, Yiwen et al. (2023). “Transfer Learning and 2.5D UNet++ for CBCT-CT Synthesis in SynthRAD2023”. In: *SynthRAD2023 - MICCAI Grand Challenge*. School of Biomedical Engineering, Southern Medical University. URL: https://grand-challenge.org/challenges/synthrad2023/.

Zhou, Zongwei, Md Mahfuzur Rahman Siddiquee, et al. (2019). “Unet++: Redesigning skip connections to exploit multiscale features in image segmentation”. In: *IEEE transactions on medical imaging* 39.6, pp. 1856–1867.

Zhou, Zongwei, Vatsal Sodha, et al. (2019). “Models genesis: Generic autodidactic models for 3d medical image analysis”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer, pp. 384–393.