Leveraging Uncertainties in Medical Prediction Systems

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Paul Fischer, M. Sc.
aus Borodulicha/Kasachstan

Tübingen
2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 21.01.2026

Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Ass. Prof. Dr. Christian F. Baumgartner
2. Berichterstatter:	Prof. Dr. Kerstin Ritter

Disclaimer: this thesis uses Felix Dangel's template which is based on Federico Marotta's `kaobook` template and on Ken Arroyo Ohori's doctoral thesis.

Acknowledgments

First and foremost, I extend my deepest gratitude to my supervisor, Christian F. Baumgartner. Your mentorship has been exceptional, far surpassing anything I could have imagined. From our very first conversation, you looked beyond my CV and listened with genuine interest, a quality that has defined your guidance throughout this journey. You consistently valued my opinion in all our discussions and always made time for me, creating an environment of mutual respect and intellectual freedom that was instrumental to my growth as a researcher. Thank you for everything.

I am grateful to Kerstin Ritter for taking the time to review this thesis. My thanks also go to Philipp Hennig and Michael Muehlebach for serving on my thesis advisory committee and for providing additional guidance and valuable perspectives throughout my doctoral studies.

A special thank you is owed to Renate Hallmayer, Sebastian Schwenk, Leila Masri, and Sara Sorce. For your tireless support in navigating all administrative matters, thank you. You are the ones who truly keep the show running, and your help has been invaluable.

This research would not have been possible without the contributions of my wonderful collaborators. My thanks go to Anna Wundram, Thomas Küstner, Daniela Thorwarth, Moritz Schneider, Jan Nikolas Morshuis, Hannah Willms, and many more. I have truly enjoyed working with each of you and have learned immensely from our joint discussions.

My PhD experience was enriched immeasurably by my friends and colleagues in the lab. To Jan Nikolas Morshuis, Stefano Woerner, Susu Sun, Anna Wundram, and Jaivardhan Kapoor, thank you for all the shared lunch breaks and wonderful discussions, even about the most random topics. During my time as a PhD student, I also had the pleasure of co supervising Leonard Siegert, Anna Wundram, and Hannah Willms. I truly enjoyed this experience and learned a great deal from it.

On a more personal note, my academic journey began long before this PhD, and I must thank Britta Dorn and Rüdiger Zeller for providing me with a fantastic introduction to the academic world. Thank you for teaching me the profound difference between mere calculation and true mathematics, and for igniting a passion for a field I did not know I had within me.

To my day ones and closest friends, Daniel Högliger and Julius Vetter: thank you for the unforgettable memories from our studies and beyond. You have taught me the meaning of true friendship.

The greatest thanks of all go to my family. To my parents, Valerij and Inna Fischer, for taking the brave and life altering decision to move to Germany at a time when we had literally nothing. To my brother, for sharing a wild and wonderful childhood. You have all supported me in ways I can never repay, often taking a step back so that I could take one forward. This achievement is as much yours as it is mine.

Finally, to my wife, Kenza Fekhari. Thank you for joining me on a wild journey that makes completing a PhD seem like a small feat in comparison. You are the funniest and most precious soul I have ever met. Your rich views and unique perspectives teach me something new every single day. I am excited for everything that lies ahead of us.

Paul Fischer
Tübingen, September 19, 2025

Table of Contents

Abstract

Machine learning can help meet rising demand for medical diagnosis and improve patient outcomes, but medicine is a high-risk domain in which uncertainty is pervasive and single-point predictions are insufficient. Despite this, the common practice in medical AI is to develop models for isolated tasks. This paradigm is fundamentally flawed, as it ignores how uncertainty from one step in a clinical workflow cascades and compounds, potentially leading to overconfident and unreliable downstream decisions. This thesis argues for a paradigm shift, moving from isolated models to a holistic, pipeline-aware framework. For machine learning to be deployed safely and effectively, uncertainty must be formally represented, propagated between tasks, made expressive through calibration, and ultimately leveraged to guide clinical decisions. To this end, we develop and validate a suite of methods that implement this vision.

First, we establish a method for propagating uncertainty from upstream data acquisition (accelerated MRI) to downstream analysis (segmentation). Second, we demonstrate how leveraging this propagated uncertainty, by marginalizing over a distribution of plausible segmentations, significantly improves the performance and robustness of a final clinical decision (glaucoma diagnosis). Third, we make uncertainties more expressive and trustworthy through a novel method for distribution-free, subgroup-specific calibration, enabling reliable error control for dose estimation in radiation therapy. Finally, we integrate these principles into a closed-loop system where calibrated uncertainty in clinical metrics dynamically guides the MRI acquisition process to optimize scan time.

Together, these contributions provide a methodological foundation for integrated, uncertainty-aware systems. By treating diagnosis as a sequence of dependent tasks, we show how expressive uncertainties can be propagated and acted upon end-to-end. Through demonstrations in accelerated MRI, glaucoma diagnosis, and radiation therapy, this work highlights a pathway towards building safer, more efficient, and clinically trustworthy ML-assisted diagnostics.

Zusammenfassung

Maschinelles Lernen kann dazu beitragen, dem steigenden Bedarf an medizinischer Diagnostik gerecht zu werden und die Behandlungsergebnisse für Patienten zu verbessern. Die Medizin ist jedoch ein Hochrisikobereich, in dem Unsicherheit allgegenwärtig ist und einzelne Punktvorhersagen unzureichend sind. Trotzdem ist es in der medizinischen KI die gängige Praxis, Modelle für isolierte Aufgaben zu entwickeln. Dieses Paradigma zeigt grundlegende Mängel auf, da es ignoriert, wie sich Unsicherheit aus einem Schritt eines klinischen Arbeitsablaufs fortpflanzt und akkumuliert, was potenziell zu übermäßig selbstsicheren und unzuverlässigen Entscheidungen in späteren Schritten führt. Diese Dissertation plädiert für einen Paradigmenwechsel: weg von isolierten Modellen hin zu einem ganzheitlichen, pipeline-orientierten Ansatz. Damit maschinelles Lernen sicher und effektiv eingesetzt werden kann, muss Unsicherheit formal repräsentiert, zwischen Aufgaben weitergegeben, durch Kalibrierung aussagekräftig gemacht und letztlich für klinische Entscheidungen genutzt werden. Zu diesem Zweck entwickeln und validieren wir eine Reihe von Methoden, die diese Vision umsetzen.

Zunächst etablieren wir eine Methode zur Weitergabe von Unsicherheit von der vorgelagerten Datenerfassung (beschleunigte MRT) zur nachgelagerten Analyse (Segmentierung). Zweitens zeigen wir, wie letztendlich die Nutzung dieser weitergegebenen Unsicherheit durch Marginalisierung über eine Verteilung plausibler Segmentierungen die Leistung und Robustheit einer klinischen Entscheidung (Glaukomdiagnose) signifikant verbessert. Drittens machen wir Unsicherheiten durch eine neuartige Methode zur verteilungsunabhängigen, untergruppenspezifischen Kalibrierung aussagekräftiger und vertrauenswürdiger. Dies ermöglicht eine zuverlässige Fehlerkontrolle bei der Dosisabschätzung in der Strahlentherapie. Schließlich integrieren wir diese Prinzipien in ein geschlossenes System, in dem kalibrierte Unsicherheit in klinischen Metriken den MRT-Akquisitionsprozess dynamisch steuert, um die Scandauer zu optimieren.

Zusammen bilden diese Beiträge eine methodische Grundlage für integrierte, unsicherheitsbewusste Systeme. Indem wir die Diagnose als eine Sequenz abhängiger Aufgaben betrachten, zeigen wir, wie aussagekräftige Unsicherheiten weitergegeben und berücksichtigt werden können. Durch Demonstrationen in der beschleunigten MRT, der Glaukomdiagnose und der Strahlentherapie zeigt diese Arbeit einen Weg zum Aufbau einer sichereren, effizienteren und klinisch vertrauenswürdigeren ML-gestützten Diagnostik auf.

List of Publications

1 Peer-Reviewed Publications

- Fischer, P., Küstner, T., Baumgartner, C.F. (2023). Uncertainty Estimation and Propagation in Accelerated MRI Reconstruction. In: Sudre, C.H., Baumgartner, C.F., Dalca, A., Mehta, R., Qin, C., Wells, W.M. (eds) Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. UNSURE 2023. Lecture Notes in Computer Science, vol 14291. Springer, Cham. https://doi.org/10.1007/978-3-031-44336-7_9

  Author Contributions: This project was led by myself in close collaboration with Prof. Dr. Thomas Küstner and under the primary supervision of Ass. Prof. Dr. Christian F. Baumgartner. The initial concept of propagating uncertainty from MRI reconstruction to a downstream segmentation task was jointly conceived by Ass. Prof. Dr. Baumgartner and me. I was responsible for the entire realization of the project, including the software implementation, conducting the experiments, and performing the initial analysis of the results. The final evaluation and interpretation of the findings were a collaborative effort among all authors. I prepared the initial draft of the manuscript and created all figures and tables. Prof. Dr. Thomas Küstner and Ass. Prof. Dr. Christian F. Baumgartner provided valuable feedback and contributed to the editing of the final manuscript.

- Wundram, A. M.*, Fischer, P.*, Wunderlich, S., Faber, H., Koch, L. M., Berens, P., Baumgartner, C. F. (2024). (2024, December). Leveraging probabilistic segmentation models for improved glaucoma diagnosis: A clinical pipeline approach. In *Medical Imaging with Deep Learning*.

  Author Contributions: Anna M. Wundram and I are listed as equal first authors, reflecting our shared lead roles in the project. This work originated as a research internship by Stefan Wunderlich and was subsequently advanced by Anna M. Wundram under my co-supervision and the primary supervision of Ass. Prof. Dr. Christian F. Baumgartner. The conceptualization of leveraging segmentation uncertainty for glaucoma diagnosis was a joint effort by Anna M. Wundram, myself, and Ass. Prof. Dr. Baumgartner. Stefan Wunderlich performed the initial implementation of the segmentation experiments. The core segmentation model was implemented by myself and Ass. Prof. Dr. Baumgartner. Anna M. Wundram conducted the comprehensive segmentation experiments, while I was responsible for the downstream classification experiments. The evaluation of all experiments was conducted jointly by Anna M. Wundram and me. The interpretation of the results was a collaborative effort involving all authors. This project was a collaboration with Dr. Hanna Faber, Ass. Prof. Dr. Lisa M. Koch, and Prof. Dr. Philipp Berens, who provided valuable feedback throughout the project. Dr. Hanna Faber, as an ophthalmologist, provided crucial clinical expertise and context. Anna M. Wundram and I jointly wrote the initial manuscript, and all co-authors contributed to subsequent revisions and editing. This publication was honored with the **Best Paper-Poster Award** at the Medical Imaging with Deep Learning (MIDL) conference in 2024.

- Fischer, P.*, Willms, H.*, Schneider, M., Thorwarth, D., Muehlebach, M., Baumgartner, C.F. (2024). Subgroup-Specific Risk-Controlled Dose Estimation in Radiotherapy. In: Linguraru, M.G., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. MICCAI 2024. Lecture Notes in Computer Science, vol 15010. Springer, Cham. https://doi.org/10.1007/978-3-031-72117-5_65

  Author Contributions: I am listed as an equal first author with Hannah Willms, reflecting our shared technical lead on different aspects of this work. This project originated from a research internship, co-supervised by myself and Ass. Prof. Dr. Christian F. Baumgartner, who was the primary supervisor. This was a collaboration with Moritz Schneider, Prof. Dr. Daniela Thorwarth, and Dr. Michael Muehlebach. The core idea of applying Risk-Controlling Prediction Sets for dose estimation in radiotherapy was conceived by me. The initial model for doese prediction was provided by Moritz Schneider. The theoretical framework and the mathematical proof of correctness were jointly developed by Dr. Michael

* These authors contributed equally.

Muehlebach and myself. Hannah Willms implemented and conducted all experiments and evaluations, with guidance from myself and Moritz Schneider. The interpretation of the results was a joint effort by all co-authors. I prepared the initial draft of the manuscript and all figures. All co-authors contributed to the editing and refinement of the final publication.

2 Manuscripts under Review

- Fischer, P., Morshuis, J. N., Küstner, T., & Baumgartner, C. (2025). CUTE-MRI: Conformalized Uncertainty-based framework for Time-adaptivE MRI. arXiv preprint arXiv:2508.14952. Under review in the journal of Medical Image Analysis by Elsevier.

  Author Contributions: I was the lead author of this project, which was conducted in collaboration with Jan Nikolas Morshuis, Prof. Dr. Thomas Küstner, and under the primary supervision of Ass. Prof. Dr. Christian F. Baumgartner. The central idea of using propagated and calibrated uncertainty from accelerated MRI scans to determine a dynamic stopping point for the acquisition was jointly conceived by myself, Prof. Dr. Thomas Küstner, and Ass. Prof. Dr. Baumgartner. I was responsible for the entire implementation of the framework, conducting all experiments, and performing the evaluation. Prof. Dr. Thomas Küstner provided the additional cardiac MRI dataset essential for the second application part of the study. Jan Nikolas Morshuis and Prof. Dr. Thomas Küstner provided regular feedback on the concepts and results. The interpretation of the results was a joint effort between myself, Prof. Dr. Küstner, and Ass. Prof. Dr. Baumgartner. I wrote the initial draft of the manuscript, and all co-authors contributed to the editing process.

1. Introduction

The landscape of modern medicine is undergoing a fundamental transformation, driven by increasing demand for medical services and the emergence of powerful computational tools. An aging global population and a growing shortage of specialists, particularly in fields like radiology, have created a critical need for technologies that can augment human expertise, streamline clinical workflows, and improve diagnostic efficiency [10, 42, 78, 117]. Machine learning (ML), and specifically deep learning, has emerged as a uniquely promising solution. In recent years, the number of Artificial Intelligence (AI) based medical applications receiving regulatory approval from bodies like the U.S. Food and Drug Administration (FDA) has grown significantly [108], marking a decisive shift from academic research to real-world clinical deployment as illustrated in Figure 1.1.

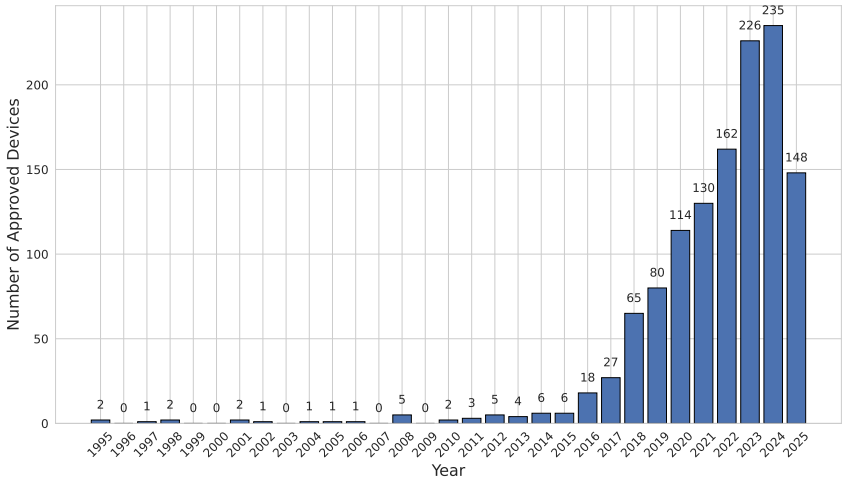

Figure 1.1: The number of FDA-approved AI/ML-enabled medical devices per year from 1995 to 2025. After a period of sparse approvals, the data shows a dramatic acceleration beginning around 2017, peaking in 2023 with 235 approved devices. Note that the data for the year 2025 is incomplete.

Despite this rapid progress, a challenge hindering the widespread translation of these tools into clinical practice is a lack of trust [36, 54]. Unlike applications in e-commerce or entertainment, medicine is a high-risk field where an incorrect prediction can have severe consequences for a patient's health. For an AI system to be a reliable partner in clinical decision making, it cannot operate as a "black box" that produces confident sounding predictions without regard for the evidence. A prerequisite for building this trust is the ability of a model to explicitly manage and communicate its uncertainty.

In fact, the uncertainty that these models must capture is not an artifact introduced by AI. It is an intrinsic property of the medical diagnostic process itself [41, 53]. Long before the rise of machine learning, clinicians have had to navigate ambiguity originating from multiple sources. This uncertainty can arise from the data, such as noise and motion artifacts in an MRI scan. It can also stem from ambiguity in human physiology, like the fuzzy border of a tumor where even expert pathologists disagree on the precise boundary [51, 84, 90]. A trustworthy AI system must therefore be able to recognize and reflect this uncertainty, signaling to a clinician when a prediction is based on ambiguous or low quality evidence.

Consequently, a great deal of research has been dedicated to develop machine learning models that incorporate uncertainty quantification (UQ), allowing them to move beyond deterministic point estimates and instead capture a distribution of possible outcomes. These approaches can model different facets of uncertainty. *Aleatoric uncertainty* captures the irreducible ambiguity in the data itself, whereas *epistemic uncertainty* reflects the model's own limitations, for instance due to a lack of training data for a rare condition [55, 57]. By

equipping models with this capability, we take the first step towards creating systems that clinicians can interrogate, understand, and ultimately trust.

Moreover, medical diagnosis is rarely a single, isolated task. It is more accurately described as a diagnostic pipeline, a sequence of interconnected and dependent steps [82]. For example, a cancer diagnosis from an MRI scan might involve (1) accelerated image acquisition, (2) computational reconstruction, (3) tumor segmentation, (4) extraction of quantitative biomarkers from the segmentation, and finally, (5) a classification of malignancy. Uncertainty introduced at any stage does not simply vanish; it propagates through every subsequent step. A slightly noisy reconstruction can lead to an uncertain segmentation, which in turn leads to unreliable biomarkers and, ultimately, a final diagnosis built on a fragile basis.

Despite the pipeline nature of clinical diagnosis, the dominant approach in the development of medical AI has been the "isolated task paradigm." In this paradigm, models are developed, trained, and evaluated for a single task, such as segmentation or classification, without formal consideration of how they fit into the larger workflow. This approach ignores the cascading effect of uncertainty. A segmentation model that reports 99% accuracy on its own metric provides a dangerously incomplete picture if it is operating on input from an upstream model that is itself highly uncertain.

While significant progress has been made in quantifying uncertainty for individual models, a critical research gap remains in understanding how to manage it holistically. This leaves the central challenges of how to formally propagate uncertainty from one pipeline stage to the next, how to ensure these uncertainties are expressive and calibrated, and how to actively leverage this information to improve the pipeline's overall performance largely unaddressed. This gap prevents the development of truly robust and reliable end-to-end systems that can be deployed with confidence. This thesis directly confronts this limitation by moving beyond isolated tasks to develop a pipeline-aware framework for uncertainty management. To substantiate this claim, this work presents a cohesive research trajectory across four primary contributions.

First, we establish the foundational capability of propagating uncertainty through a diagnostic pipeline. We develop a novel probabilistic reconstruction method, PHiRec, and provide the first comprehensive quantitative evaluation of how uncertainty from accelerated MRI reconstruction can be formally carried forward to a downstream segmentation task. This demonstrates that the ambiguity from an early processing step can be preserved and passed on, rather than being discarded.

Second, having shown that uncertainty can be propagated, we investigate whether this information can be actively leveraged to improve clinical outcomes. We present an interpretable, multi-stage pipeline for glaucoma diagnosis that explicitly models the full distribution of segmentation possibilities. By marginalizing over all potential segmentations to inform the final diagnosis, we demonstrate that utilizing upstream uncertainty leads to a significantly more accurate and robust clinical classification.

Third, to bridge the gap to real-world deployment, propagated uncertainties must not only be informative but also formally reliable. This contribution develops a method to produce more expressive and trustworthy uncertainties with mathematical guarantees. We introduce Subgroup Risk-Controlling Prediction Sets (SG-RCPS), a novel calibration algorithm that provides distribution-free, risk-controlled prediction intervals. This algorithm ensures these reliability guarantees hold even for small but clinically critical subgroups within the data, addressing a major failure point of standard calibration techniques in medical applications.

Finally, we integrate these principles of propagation, leveraging, and calibration into a single, closed-loop, uncertainty-aware system. We introduce CUTE-MRI, a framework that uses calibrated, pipeline-level uncertainty to dynamically guide and optimize the MRI acquisition process itself. By halting the scan precisely when a desired level of diagnostic confidence is reached, this work demonstrates how a holistic approach to uncertainty can make the entire clinical workflow more efficient, adaptive, and patient-specific.

The remainder of this thesis is structured to build upon these contributions. Chapter 2 provides the necessary technical background on machine learning and uncertainty quantification in medicine, substantiating the research gap identified in this introduction. Chapter 3 presents the core scientific contributions of this thesis, each corresponding to one of the published works. Section 3.1 details the method for uncertainty propagation. Section 3.2 shows how this uncertainty can be leveraged. Section 3.3 introduces the novel calibration technique. Section 3.4 demonstrates the integration of these methods in a dynamic acquisition

system. Finally, Chapter 4 concludes the thesis with a synthesis of the findings, a discussion of their broader implications, and an outlook on future research directions.

2. Technical Background

In this chapter we introduce the key concepts and frameworks relevant to the thesis. We present the main models and explains their connection to the work presented in Chapter 3.

2.1 The U-Net Architecture for Dense Prediction

A cornerstone of modern medical image analysis, particularly for tasks requiring dense, pixel-wise predictions, is the U-Net architecture [93]. Even though it has been introduced in 2015 already, its design has proven remarkably effective and versatile, serving as the foundational model for a vast range of applications including segmentation, image reconstruction, and other image-to-image translation tasks [50, 67] and still showing competitive performance up to today [45, 46]. The U-Net's success stems from its simple and powerful structure, which is specifically tailored to capture both contextual and localization information from an image.

The architecture consists of two main pathways: a *contracting path* (encoder) and an *expanding path* (decoder), which together form a characteristic U-shape. An illustration of the architecture can be seen in Figure 2.1.

1. **The Contracting Path (Encoder):** This path follows the typical structure of a convolutional neural network. It is composed of a sequence of convolutional layers, each followed by a Rectified Linear Unit (ReLU) activation function and a max pooling operation. With each step down the path, the spatial resolution of the feature maps decreases, while the number of feature channels increases. The purpose of the encoder is to extract increasingly complex and abstract features from the input image, capturing the contextual information (the "what") at the cost of spatial precision (the "where").
2. **The Expanding Path (Decoder):** This path is responsible for up-sampling the feature maps and recovering the spatial resolution required for a dense prediction. Each step in the decoder consists of an up-sampling operation (e.g., a transposed convolution or bilinear up-sampling), followed by a concatenation with the corresponding feature map from the contracting path, and then two convolutional layers with ReLU activations.

The most critical innovation of the U-Net is the use of **skip connections**, which concatenate the high-resolution feature maps from the encoder with the up-sampled outputs of the decoder. These connections provide a direct pathway for fine-grained, local information to be passed to the higher levels of the network. This allows the decoder to combine high-level semantic information learned deep in the network with precise, low-level localization details from the earlier layers. This fusion is essential for producing accurate, high-resolution segmentation masks or reconstructions, and it is the primary reason for the U-Net's enduring prevalence in the field [119], including as a backbone architecture in the works presented in Sections 3.1, 3.2, 3.3 and 3.4 of this thesis.

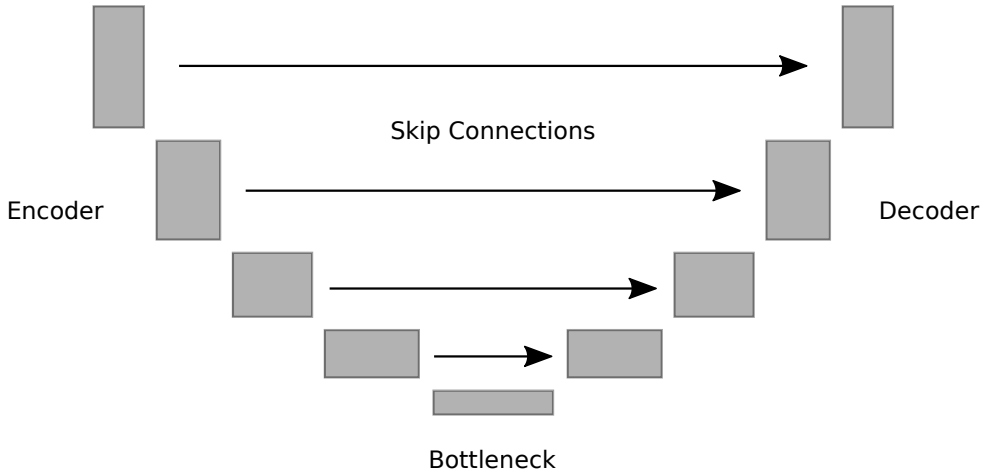

Figure 2.1: A schematic of the U-Net architecture, illustrating the contracting path (encoder), expanding path (decoder), and the crucial skip connections that fuse high-resolution features with upsampled features.

2.2 Deep Learning for Inverse Problems

Many critical tasks in medical imaging can be formulated as *inverse problems*, where the goal is to infer a hidden or unobserved cause from a set of observed effects [12, 72]. Mathematically, given a physical forward process A, we observe a measurement $\boldsymbol{y}$ that is related to an underlying signal of interest $\boldsymbol{x}$ by the equation:

$$\boldsymbol{y} = A(\boldsymbol{x}) + \boldsymbol{\varepsilon} \tag{2.1}$$

where $\boldsymbol{\varepsilon}$ represents measurement noise. The inverse problem is to recover $\boldsymbol{x}$ given the measurement $\boldsymbol{y}$. This problem becomes particularly challenging when it is *ill-posed*, meaning that a unique, stable solution does not exist. This is often the case when the forward operator A is non-invertible, leading to a situation where a single measurement $\boldsymbol{y}$ could have been generated by an infinite number of possible signals $\boldsymbol{x}$.

A canonical example of an ill-posed inverse problem in medicine, and one central to this thesis, is **accelerated Magnetic Resonance Imaging (MRI)** [66]. In MRI, the raw data, known as k-space, is acquired in the frequency domain and is related to the anatomical image $\boldsymbol{x}$ via the Fourier transform, $\mathcal{F}$. Modern MRI scanners employ multiple receiver coils, each with a spatially varying sensitivity map, to acquire data in parallel [37, 88]. The sensitivity encoding operator $\boldsymbol{S}$ modulates the underlying image by each coil's sensitivity before it is transformed into k-space. A fully-sampled acquisition can be prohibitively slow. To accelerate the process, only a fraction of the k-space is acquired. This can be modeled by applying a binary undersampling mask $\boldsymbol{M}$ to the full k-space data. The forward process for accelerated MRI is thus $A = \boldsymbol{M}\mathcal{F}\boldsymbol{S}$, and the measured undersampled k-space is $\boldsymbol{y}_u = \boldsymbol{M}\mathcal{F}\boldsymbol{S}(\boldsymbol{x}_{fs})$, where $\boldsymbol{x}_{fs}$ is the desired fully-sampled image.

The inverse problem is to reconstruct $\boldsymbol{x}_{fs}$ from $\boldsymbol{y}_u$. Since the mask $\boldsymbol{M}$ discards information, the operator A is non-invertible, and the problem is severely ill-posed. A naive reconstruction via the inverse Fourier transform, $\boldsymbol{x}_u = \mathcal{F}^{-1}(\boldsymbol{y}_u)$, results in a highly aliased image.

To overcome this, classical methods solve the inverse problem using constrained optimization, balancing data consistency with a manually chosen regularization prior. The most influential framework has been Compressed Sensing (CS), which assumes that medical images are sparse in a transform domain (e.g., wavelets) [19, 68]. A CS reconstruction is found by solving an optimization problem of the form:

$$\underset{\boldsymbol{x}}{\operatorname{argmin}} \quad \frac{1}{2}||\boldsymbol{M}\mathcal{F}\boldsymbol{S}(\boldsymbol{x}) - \boldsymbol{y}_u||_2^2 + \lambda||\Psi(\boldsymbol{x})||_1 \tag{2.2}$$

where the first term enforces consistency with the acquired k-space data, and the second term is a regularization prior that promotes sparsity in a transform domain Ψ, weighted by a parameter λ. While groundbreaking, these iterative methods suffer from significant computational drawbacks. The optimization problem must

be solved using iterative algorithms that can take minutes to hours to converge [30], which is often too slow for clinical workflows. This computational burden scales poorly to higher-dimensional imaging (e.g., 3D or dynamic MRI) and the performance is highly dependent on the hand-tuning of parameters like λ and the choice of the sparsifying transform Ψ for each specific application [58]. These limitations, the slow reconstruction time and the reliance on hand-crafted, non-adaptive priors, created a strong need for faster and more powerful data-driven approaches.

Deep learning offers such an approach for accelerated MRI. [40, 50]. A network $f_{\boldsymbol{\theta}}$ is trained to learn the mapping from the aliased input $\boldsymbol{x}_u$ to the ground truth fully-sampled image $\boldsymbol{x}_{fs}$. The model's parameters $\boldsymbol{\theta}$ are optimized by minimizing a loss function, such as the Mean Squared Error (MSE), over a large dataset of paired images:

$$\mathcal{L}_{\text{MSE}}(\boldsymbol{\theta}) = \mathbb{E}(\boldsymbol{x}_u, \boldsymbol{x}_{fs}) \sim p_{\text{data}} \left[||f_{\boldsymbol{\theta}}(\boldsymbol{x}_u) - \boldsymbol{x}_{fs}||_2^2 \right] \tag{2.3}$$

While highly effective at producing high-quality reconstructions in milliseconds, this deterministic approach yields only a single point estimate and fundamentally ignores the inherent uncertainty stemming from the ill-posed nature of the problem [6, 64]. This motivates the need for probabilistic models that can capture the entire distribution of possible reconstructions.

2.3 Probabilistic Modeling

To address the limitations of deterministic models, probabilistic modeling aims to capture the uncertainty in predictions. This is particularly important in high-risk medical applications, where a model that is "confidently wrong" can be more dangerous than one that expresses its uncertainty, allowing a clinician to intervene [3, 9]. The goal is to move beyond learning a single mapping $f : \boldsymbol{x} \rightarrow \boldsymbol{y}$ and instead learn a full conditional probability distribution $p(\boldsymbol{y}|\boldsymbol{x})$. This distribution inherently captures both aleatoric uncertainty (the irreducible ambiguity in the data, reflected in the variance of $p(\boldsymbol{y}|\boldsymbol{x})$ itself) and epistemic uncertainty (the model's own uncertainty, reflected in its confidence about the parameters of that distribution) [55, 57]. From this distribution, one can derive not only a point estimate (e.g., the mean) but also a measure of uncertainty (e.g., the variance). In this thesis, we focus on two primary families of methods to approximate this distribution: Bayesian approximation methods and explicit generative models.

2.3.1 Bayesian Neural Networks for Epistemic Uncertainty

A principled way to capture model uncertainty is through the lens of Bayesian inference. Instead of learning a single point estimate for the model's weights $\boldsymbol{\theta}$, a Bayesian Neural Network (BNN) aims to infer a full posterior distribution over the weights, $p(\boldsymbol{\theta}|\mathcal{D})$, given the training data $\mathcal{D}$ [69, 80]. Predictions for a new input $\boldsymbol{x}_*$ are then made by marginalizing over this posterior:

$$p(\boldsymbol{y}_*|\boldsymbol{x}_*, \mathcal{D}) = \int p(\boldsymbol{y}_*|\boldsymbol{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{2.4}$$

The variance of this posterior predictive distribution, $p(\boldsymbol{y}_*|\boldsymbol{x}_*, \mathcal{D})$, represents the model's *epistemic uncertainty*. However, the integral is intractable for deep neural networks. Therefore, practical approximation methods are required. Common approaches include Variational Inference (VI) [15, 16] and Laplace approximations [22, 70], which make simplifying assumptions about the posterior's form. In this thesis, we focus on two pragmatic and widely-adopted approaches: Monte Carlo Dropout and Deep Ensembles.

Monte Carlo (MC) Dropout. Originally introduced as a regularization technique to prevent overfitting, dropout works by randomly setting a fraction of neuron activations to zero during each forward pass of the training phase [104]. This prevents neurons from co-adapting too much and forces the network to learn more robust features. Proposed by Gal & Ghahramani [35], MC Dropout provides a scalable method to approximate Bayesian inference. By keeping dropout active at *inference time* and performing T stochastic

forward passes for the same input $\boldsymbol{x}$, one obtains a set of T different predictions, $\{\hat{\boldsymbol{y}}^{(t)}\}_{t=1}^{T}$. These predictions can be treated as samples from the approximate posterior predictive distribution. The predictive mean and variance are then estimated via the sample statistics:

$$\hat{\boldsymbol{\mu}}(\boldsymbol{x}) \approx \frac{1}{T}\sum_{t=1}^{T} f_{\hat{\theta}^{(t)}}(\boldsymbol{x}) \tag{2.5}$$

$$\hat{\sigma}^2(\boldsymbol{x}) \approx \frac{1}{T}\sum_{t=1}^{T}\left(f_{\hat{\theta}^{(t)}}(\boldsymbol{x}) - \hat{\boldsymbol{\mu}}(\boldsymbol{x})\right)^2 \tag{2.6}$$

where $f_{\hat{\theta}^{(t)}}$ represents the network with a specific dropout mask applied at inference pass t.

Deep Ensembles. An alternative and powerful baseline for uncertainty estimation is the Deep Ensemble method [63]. This approach involves training N identical networks from scratch using the same dataset but usually with different random weight initializations. At inference time, the predictions from all N models are aggregated. The predictive mean is the average of the individual model predictions, and the predictive variance across the ensemble members serves as a measure of epistemic uncertainty. While computationally expensive, ensembles are remarkably effective. Although not explicitly derived from Bayesian principles, they have been shown to be a high-performing method for approximate Bayesian marginalization. The different network initializations encourage the models to converge to different modes in the loss landscape, and averaging their outputs effectively samples from a diverse, implicit posterior over the weights [34].

2.3.2 Conditional VAEs for Aleatoric Uncertainty

While Bayesian Neural Networks excel at capturing epistemic uncertainty, they are less suited for modeling the inherent, irreducible ambiguity in the data, known as *aleatoric uncertainty*. Generative models, such as the Variational Autoencoder (VAE) [26, 56, 91], are designed for this purpose. The Conditional VAE (cVAE) [101] extends this framework to supervised, conditional settings, making it ideal for tasks like probabilistic image reconstruction or segmentation [59].

The goal of a cVAE is to model the full conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ by introducing a latent variable $\boldsymbol{z}$ that captures the variability in $\boldsymbol{y}$ not explained by $\boldsymbol{x}$. A visualization of the corresponding graphical model can be found in Figure 2.2. The generative process defines the joint probability over the output and the latent variable as:

$$p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x}) = p_\theta(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z})p_\psi(\boldsymbol{z}|\boldsymbol{x}) \tag{2.7}$$

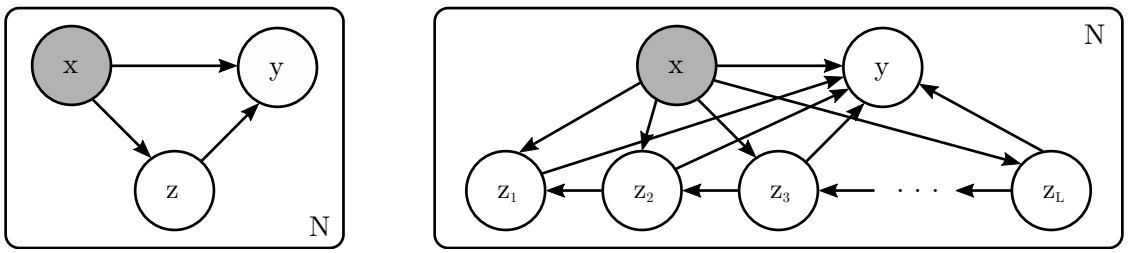

Figure 2.2: A visualization of the probability distribution for the generative process for a cVAE (left) and for a hierarchical cVAE (right). Figure adapted from [8].

The full conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ is then recovered by marginalizing out the latent variable:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \int p_\theta(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z})p_\psi(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z} \tag{2.8}$$

Since this integral and the true posterior $p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$ are intractable for deep neural networks, the model is trained by maximizing the Evidence Lower Bound (ELBO) on the conditional log likelihood:

$$\log p(\boldsymbol{y}|\boldsymbol{x}) \geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}[\log p_\theta(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z})] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})||p_\psi(\boldsymbol{z}|\boldsymbol{x})) =: \mathcal{L}_{\mathrm{cELBO}} \tag{2.9}$$

The cVAE framework consists of three neural networks trained jointly:

1. **The Prior Network** $p_\psi(\boldsymbol{z}|\boldsymbol{x})$: At inference time, this network takes the input $\boldsymbol{x}$ and outputs the parameters of a distribution (typically Gaussian) over the latent space, from which a latent code $\boldsymbol{z}$ can be sampled.
2. **The Likelihood Network (Decoder)** $p_\theta(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z})$: This network takes both the input $\boldsymbol{x}$ and a sampled latent code $\boldsymbol{z}$ to generate a plausible output $\boldsymbol{y}$. By sampling different $\boldsymbol{z}$ values, it can produce a diverse set of outputs, thereby modeling the aleatoric uncertainty.
3. **The Posterior Network (Encoder)** $q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$: Used only during training, this network approximates the true posterior by learning to encode the ground truth pair $(\boldsymbol{x}, \boldsymbol{y})$ into the latent space.

The ELBO consists of a reconstruction term, which encourages the decoder to produce realistic outputs, and a Kullback Leibler (KL) divergence term, which acts as a regularizer, forcing the approximate posterior from the encoder to stay close to the distribution produced by the prior network.

2.3.3 Hierarchical Conditional VAEs

For complex, high-dimensional data like medical images, a single low-dimensional latent variable $\boldsymbol{z}$ can act as an information bottleneck, failing to capture the full complexity of the data distribution. It struggles to model both high-level global structure and fine-grained local details simultaneously. To address this, Hierarchical cVAEs extend the standard cVAE by introducing a sequence of latent variables, typically arranged across different spatial scales [8, 102, 109].

A hierarchical model with L levels defines a generative process that factors the joint distribution over the output $\boldsymbol{y}$ and the latent variables $\boldsymbol{z}_{1:L} = \boldsymbol{z}_1, \ldots, \boldsymbol{z}_L$ as follows:

$$p(\boldsymbol{y}, \boldsymbol{z}_{1:L}|\boldsymbol{x}) = p_\theta(\boldsymbol{y}|\boldsymbol{z}_{1:L}, \boldsymbol{x})p_\psi(\boldsymbol{z}_1|\boldsymbol{z}_2, \boldsymbol{x}) \ldots p_\psi(\boldsymbol{z}_{L-1}|\boldsymbol{z}_L, \boldsymbol{x})p_\psi(\boldsymbol{z}_L|\boldsymbol{x}) \tag{2.10}$$

In this structure, the top-level latent $\boldsymbol{z}_L$ (often at the coarsest spatial resolution) can capture abstract, global variations, while lower-level latents $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{L-1}$ (at finer resolutions) can model more detailed, local variations conditioned on the information from the levels above. A comparison to the standard cVAE approach is visualized as a graphical model in Figure 2.2.

Training is again performed by maximizing a corresponding hierarchical ELBO, which involves an approximate posterior $q_\phi(\boldsymbol{z}_{1:L}|\boldsymbol{x}, \boldsymbol{y})$ that is also factorized. The resulting objective function includes a reconstruction term and a sum of KL divergence terms, one for each level of the hierarchy, regularizing the posterior of each latent variable towards its prior. This more expressive architecture allows the model to learn a much richer and more structured representation of the data distribution, which is critical for generating high-fidelity and diverse samples of medical images. This hierarchical framework forms the basis of the probabilistic model developed in Section 3.1 of this thesis.

2.3.4 Uncertainty Propagation in Pipelines

A central theme of this thesis is the analysis of multi-stage diagnostic pipelines, which can be represented as a composition of models, $f = f_N \circ \cdots \circ f_1$, where the output of one model serves as the input to the next. In such a setup, it is crucial to understand how uncertainty from an early stage, such as image reconstruction (f_1), propagates to and influences the final output, such as a clinical diagnosis (f_N). This requires robust methods for uncertainty propagation.

Several strategies exist for this task, broadly categorized as analytical, deterministic approximation, and sampling-based methods.

Analytical propagation attempts to derive a closed-form expression for the output distribution's parameters (e.g., mean and variance) based on the input distribution. This is often called moment matching. The core challenge lies in solving integraks which is only tractable for very simple, often linear, models and specific distribution families (e.g., propagating a Gaussian distribution through a linear transformation) [14]. For the complex, non-linear functions defined by deep neural networks, this approach is infeasible.

Deterministic approximation methods avoid the limitations of direct analytical solutions. For instance, Taylor series approximations (e.g., the Delta method) linearize the model around the input mean, allowing for straightforward propagation of moments through the resulting linear function. However, this approach requires computing Jacobians and its accuracy degrades significantly for highly non-linear models like deep neural networks. A more advanced technique is the Unscented Transform (UT), which is a deterministic sampling method. The UT propagates a carefully chosen set of 'sigma points' through the exact non-linear model and then reconstructs the output statistics from the transformed points [52, 112]. While often more accurate than linearization, the UT can be challenging to apply in the high-dimensional spaces typical of medical imaging and deep learning.

Given these limitations, this thesis focuses on Monte Carlo (MC) sampling for uncertainty propagation [14, 29, 92]. This method is powerful due to its simplicity, generality, and scalability. The process is straightforward:

1. An upstream probabilistic model $p(\boldsymbol{y}_1|\boldsymbol{x})$ is used to generate M independent samples of its output: $\boldsymbol{y}_1^{(m)}$, $m = 1, \ldots, M$.
2. Each sample $\boldsymbol{y}_1^{(m)}$ is then passed independently through the subsequent deterministic downstream model f_2 to obtain a corresponding set of downstream outputs: $\boldsymbol{y}_2^{(m)} = f_2(\boldsymbol{y}_1^{(m)})$, $m = 1, \ldots, M$.
3. This resulting set of samples, $\boldsymbol{y}_2^{(m)}$, forms an empirical estimate of the propagated posterior distribution $p(\boldsymbol{y}_2|\boldsymbol{x})$.

The key advantage of MC sampling is that it is model-agnostic; it makes no assumptions about the functional form of the downstream model f_2, which can be any complex, non-linear function like a deep neural network. Furthermore, the process is highly parallelizable, as each sample can be processed independently, making it computationally efficient on modern hardware. By analyzing the statistics of the resulting samples (e.g., their mean and variance), one can robustly estimate the impact of upstream uncertainty on any subsequent stage in a complex diagnostic pipeline. This technique is therefore a cornerstone of the pipeline-aware analyses presented in Sections 3.1, 3.2 and 3.4.

2.4 From Heuristics to Guarantees: Evaluating and Calibrating Uncertainty

Producing an uncertainty estimate is only the first step. For an AI system to be integrated into clinical workflows, its uncertainty estimates must be reliable. A clinician needs to know if a model's stated 90% confidence level truly corresponds to a 90% chance of being correct. Therefore, ensuring that this estimate is reliable, well-calibrated, and trustworthy is absolutely necessary for clinical adoption. This section reviews methods for evaluating uncertainty and frameworks for providing formal calibration guarantees.

2.4.1 Metrics for Uncertainty Quantification

Calibration. An uncertainty estimate is considered *calibrated* if the confidence of a prediction aligns with its empirical accuracy. For classification, a common way to visualize this is with a *reliability diagram*, which plots the expected accuracy as a function of the predicted confidence. For a perfectly calibrated model, this plot would be the identity line. A widely used summary metric is the *Expected Calibration Error (ECE)*, which computes the weighted average of the gap between confidence and accuracy across different bins [38]. For regression, calibration is often assessed via the *coverage probability* of its prediction intervals. For a model producing $(1 - \alpha) \times 100\%$ prediction intervals, the empirical coverage, i.e. the fraction of true values

that actually fall within their respective intervals, should be close to the nominal level $1 - \alpha$. A significant deviation indicates miscalibration [114].

Spatial Correlation. In imaging tasks, it is also crucial that the uncertainty is spatially correlated with the model's error. A good uncertainty map should have high values in regions where the prediction is erroneous and low values where it is accurate. The **Normalized Cross-Correlation (NCC)** between the model's squared error map and its predicted variance map is an effective metric for this [65]. For an error map $\boldsymbol{E}$ and a variance map $\boldsymbol{V}$, the NCC is given by:

$$\text{NCC}(\boldsymbol{E}, \boldsymbol{V}) = \frac{1}{N} \sum_{i=1}^{N} \frac{(E_i - \mu_E)(V_i - \mu_V)}{\sigma_E \sigma_V} \tag{2.11}$$

where i indexes the pixels, N is the total number of pixels, and μ and σ are the respective means and standard deviations. An NCC value close to 1 indicates a strong positive correlation, signifying a useful, well-localized uncertainty estimate.

2.4.2 Calibration with Formal Guarantees

While the metrics above can diagnose miscalibration on a test set, they do not correct the model or provide guarantees for future data. Several post-hoc calibration methods exist, such as Platt Scaling [38, 87] or Isotonic Regression [118], which learn a mapping from a model's raw scores to calibrated probabilities. However, these methods provide no formal guarantees of calibration for unseen data points. To achieve this, more rigorous, distribution-free frameworks are needed.

Conformal Prediction. A model-agnostic framework that converts heuristic outputs of any machine learning model into prediction sets with formal, distribution-free guarantees on their coverage probability is **Conformal Prediction** [98, 110].

The *split conformal* or *inductive* approach operates as follows [4, 85]:

1. The available data is split into a proper training set $\mathcal{D}_{\text{train}}$ and a dedicated calibration set $\mathcal{D}_{\text{cal}} = (\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^{n}$ where n denotes the number of elements in the calibration set.
2. A model f is trained on $\mathcal{D}_{\text{train}}$.
3. A *non-conformity score* $S(\boldsymbol{x}, \boldsymbol{y})$ is defined, which measures how "strange" a true label $\boldsymbol{y}$ is, given the model's prediction for $\boldsymbol{x}$. For regression, a common choice is the absolute error: $S(\boldsymbol{x}, \boldsymbol{y}) = |\boldsymbol{y} - f(\boldsymbol{x})|/u(x)$ where $u(x)$ is some measure of uncertainty.
4. The non-conformity scores are computed for all samples in the calibration set: $s_i = S(\boldsymbol{x}_i, \boldsymbol{y}_i)$.
5. For a desired miscoverage level α, the $(1-\alpha)(1+1/n)$-th quantile of the calibration scores s_i is computed. Let this value be $\hat{q}$.
6. For a new test sample $\boldsymbol{x}_{\text{new}}$, the prediction interval is constructed as"

$$\mathcal{C}(\boldsymbol{x}_{\text{new}}) = [f(\boldsymbol{x}_{\text{new}}) - u(x)\hat{q}, f(\boldsymbol{x}_{\text{new}}) + u(x)\hat{q}]$$

By construction, this procedure guarantees that for a new data point drawn from the same distribution, the true label will be contained in the prediction interval with a probability of at least $1 - \alpha$:

$$P(\boldsymbol{y}_{\text{new}} \in \mathcal{C}(\boldsymbol{x}_{\text{new}})) \geq 1 - \alpha$$

This provides a rigorous way to ensure reliability, as explored in Section 3.4.

Risk-Controlling Prediction Sets (RCPS). While Conformal Prediction guarantees control over the coverage probability, it is a special case of a more general goal: controlling statistical risk. The framework of **Risk-Controlling Prediction Sets (RCPS)** generalizes this idea to control the expected value of a user-defined loss function $L(\boldsymbol{y}, \mathcal{C}(\boldsymbol{x}))$ to be below a certain level δ [7]. This allows for the control of more complex, task-specific

risks beyond simple miscoverage, such as controlling the size of the prediction set or other clinical cost functions.

The standard RCPS algorithm operates similarly to conformal prediction but is more flexible. It begins with a heuristic, set-valued function $\mathcal{T}_\lambda(\boldsymbol{x})$, which is parameterized by a scalar $\lambda \geq 0$. This function is designed to be monotonic, such that larger values of λ produce larger (more conservative) prediction sets. A common choice is to scale a heuristic interval: $\mathcal{T}_\lambda(\boldsymbol{x}) = [f(\boldsymbol{x}) - \lambda \cdot \hat{\sigma}(\boldsymbol{x}), f(\boldsymbol{x}) + \lambda \cdot \hat{\sigma}(\boldsymbol{x})]$, where $\hat{\sigma}(\boldsymbol{x})$ is a heuristic uncertainty estimate. The goal is to find the smallest λ that guarantees the risk is controlled. The procedure is as follows:

1. A model that produces a heuristic prediction set $\mathcal{T}\lambda(\boldsymbol{x})$ is trained on $\mathcal{D}_{\text{train}}$.
2. For each sample $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ in the calibration set $\mathcal{D}_{\text{cal}}$, the loss is evaluated for a range of possible λ values: $l_i(\lambda) = L(\boldsymbol{y}_i, \mathcal{T}_\lambda(\boldsymbol{x}_i))$.
3. For a desired risk level δ, the algorithm finds the smallest $\hat{\lambda}$ such that an upper confidence bound on the true risk is less than or equal to δ. A common choice is the Hoeffding-Bentkus upper bound on the empirical risk [7, 11]:
$$\hat{R}_{\text{upper}}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} l_i(\lambda) + \sqrt{\frac{\log(1/\beta)}{2n}} \tag{2.12}$$
where β is a small confidence parameter.
4. The final prediction rule for a new test point $\boldsymbol{x}_{\text{new}}$ is $\mathcal{C}(\boldsymbol{x}_{\text{new}}) = \mathcal{T}_{\hat{\lambda}}(\boldsymbol{x}_{\text{new}})$.

This procedure guarantees that with probability at least $1 - \beta$, the true risk of the final prediction sets is controlled below the target level δ. This powerful framework for achieving formal, task-specific reliability is the foundation of the work presented in Section 3.3.

3. Publications

In this section we provide informations on the publications that this thesis is based on. We start with a short overview and continue presenting each paper including a short motivation, methodological summary and presentation of the results. All papers are provided in the appendix. A complete list of publications along with a describtion of the author contributions can be found in the beginning of the document.

Overview

The central contribution of this thesis is the development and validation of a holistic, pipeline-aware framework for managing uncertainty in medical diagnostics. The overarching goal of modeling uncertainty in diagnostic pipelines was systematically addressed by tackling three fundamental sub-problems that emerged during the research: (1) how to formally propagate uncertainty between consecutive tasks in a pipeline; (2) how to actively leverage this propagated uncertainty to improve clinical outcomes; and (3) how to ensure these uncertainty estimates are robust and formally calibrated to be trustworthy for critical decisions.

The following four sections present the publications that form the core of this work. Each chapter addresses one or more of these questions, building progressively to form a cohesive narrative.

The first publication lays the foundational groundwork by addressing the propagation problem [32]. We investigate the ill-posed problem of accelerated MRI reconstruction and introduce PHiRec, a probabilistic model that yields well-calibrated uncertainty estimates. We then demonstrate how Monte Carlo sampling can be used to formally pass this uncertainty onward to a downstream segmentation task.

The second paper builds on this by tackling the leveraging problem [116]. In the context of glaucoma diagnosis, a process reliant on the segmentation of highly ambiguous structures, we show how propagating a full distribution of possible segmentations, rather than a single estimate, significantly improves the accuracy of the final classification.

The third publication addresses the critical calibration problem [33]. Recognizing that propagated uncertainties must be trustworthy to be useful, we focus on the high-risk application of radiotherapy dose estimation. We develop SG-RCPS, a novel algorithm that provides formal, subgroup-specific risk guarantees, ensuring that uncertainty estimates are reliable especially in the most clinically critical regions.

Finally, the last paper integrates all these principles into a single, closed-loop system [31]. We introduce CUTE-MRI, a framework that propagates, calibrates, and leverages end-to-end pipeline uncertainty to dynamically guide the MRI acquisition process itself, optimizing scan time on a patient-specific basis.

Together, these publications chart a clear path from foundational mechanisms for uncertainty management to a fully integrated, adaptive, and trustworthy clinical system.

3.1 Uncertainty Estimation and Propagation in Accelerated MRI Reconstruction

Motivation

Magnetic Resonance Imaging (MRI) is a vital tool in clinical practice, but its long acquisition times can cause patient discomfort and limit throughput. Fast MRI techniques are crucial for addressing these issues and are key to enabling novel applications such as real-time MR-guided radiation therapy and the automated, rapid estimation of clinical parameters from scans [18, 95, 106, 111]. In recent years, reconstruction techniques based

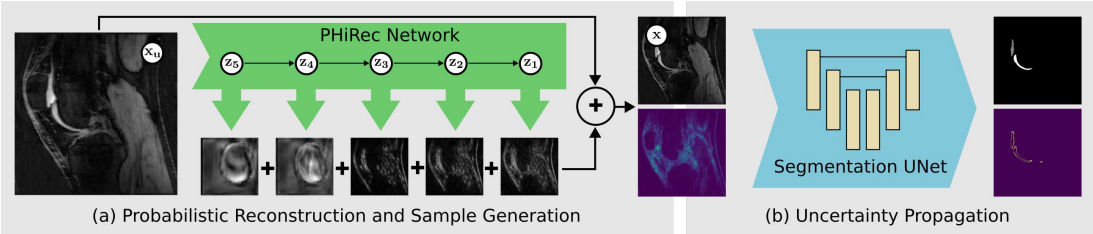

Figure 3.1: In our proposed Probabilistic Hierarchical Reconstruction (PHiRec) model, five latent variables z_l generate residual changes at different resolution scales. These changes are added to the undersampled input x to generate the final output y. The model can then be used to sample likely reconstructions and propagate them to a subsequent segmentation network. The Figure is from [32].

on deep learning (DL) have become prominent due to their excellent performance at very high acceleration rates and their ability to provide reconstructions in real-time [47, 50, 83].

Despite their impressive performance, the use of DL models for this task is complicated by a fundamental challenge: accelerated MRI reconstruction is an ill-posed problem. Because a single undersampled k-space measurement can correspond to an infinite number of plausible images, a deterministic model is forced to select only one possibility, creating a significant risk of "hallucinating" anatomical details [75]. This introduces profound uncertainty, which is dangerous because reconstructed images are rarely the final diagnostic product. They are almost always an intermediate step for downstream tasks like tumor segmentation or biomarker extraction. Consequently, the uncertainty from the reconstruction stage does not simply vanish. Instead, it propagates, directly impacting the reliability of any subsequent analysis. A segmentation performed on a hallucinated structure is inherently wrong, yet a standard deterministic pipeline would report it with full confidence. This highlights a critical gap that this work aims to address: the literature has focused on estimating and visualizing reconstruction uncertainty in isolation, while the crucial challenge of how to formally propagate this uncertainty to downstream tasks and quantitatively evaluate its impact remains largely unexplored.

Several DL-based approaches have been proposed to model this uncertainty in accelerated MRI [5, 43, 79, 105]. However, common methods show drawbacks in computational efficiency, long sampling times, lack of quantitative evaluation or comparisons with strong baselines; studies have relied almost exclusively on qualitative interpretation of the uncertainty maps. This work aims to address these gaps by proposing a novel, efficient probabilistic reconstruction technique and providing the first comprehensive quantitative evaluation of uncertainty quantification for MRI reconstruction and its propagation to a downstream segmentation task.

Methods

This paper proposes a novel probabilistic reconstruction technique called **PHiRec** (Probabilistic Hierarchical Reconstruction), which is based on a hierarchical conditional variational autoencoder (cVAE) [8]. The goal is to model the full posterior distribution of a fully-sampled MR image $\boldsymbol{y}$ given an undersampled image $\boldsymbol{x}$, denoted as $p(\boldsymbol{y}|\boldsymbol{x})$. The undersampled image $\boldsymbol{x}$ is obtained by applying the inverse Fourier operator to the zero-filled k-space measurement data.

The core of PHiRec is its **hierarchical structure**, which models the distribution using $L = 5$ separate latent variables, $z_1, ..., z_L$, where each latent variable operates on a different resolution scale. For instance, z_1 operates at the original image resolution, while z_5 operates at a resolution that has been down-sampled four times. Each level is responsible for probabilistically generating residual changes that are progressively added to the input image $\boldsymbol{x}$ to remove undersampling artifacts and produce the final reconstruction $\boldsymbol{y}$. This hierarchical approach is a very expressive model for capturing complex, high-dimensional probability distributions. A key architectural feature of PHiRec is a **skip connection** from the input $\boldsymbol{x}$ to the final output, a modification that was found to facilitate the de-aliasing task. The overall generative model can be written

as:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \int p(\boldsymbol{y}|\boldsymbol{z}_{1:L}, \boldsymbol{x}) p(\boldsymbol{z}_1|\boldsymbol{z}_2, \boldsymbol{x}) ... p(\boldsymbol{z}_{L-1}|\boldsymbol{z}_L, \boldsymbol{x}) p(\boldsymbol{z}_L|\boldsymbol{x}) d\boldsymbol{z}_{1:L} \tag{3.1}$$

The model is trained end-to-end by maximizing the Evidence Lower Bound (ELBO) on the true log-likelihood, using pairs of undersampled images $\boldsymbol{x}$ and their corresponding ground truth reconstructions $\boldsymbol{y}$. After training, the prior network can be used to generate an arbitrary number of latent variable samples, which are then decoded by the likelihood network to produce a set of final reconstruction samples. The mean of these samples serves as the final prediction, while their spread (e.g., variance) provides the uncertainty estimate.

A key contribution of this work is demonstrating how these reconstruction uncertainties can be propagated to a downstream task. Given a separately trained, deterministic segmentation network, the distribution of possible segmentations $p(\boldsymbol{s}|\boldsymbol{x})$ can be estimated using a Monte Carlo approach. Specifically, each of the reconstruction samples generated by PHiRec is passed through the segmentation network, and the resulting distribution of segmentations is analyzed empirically to estimate the segmentation uncertainty.

Results

The proposed PHiRec method was comprehensively evaluated on the Stanford Knee MRI Multi-Task Evaluation (SKM-TEA) dataset [25]. We compared our proposed method against several other common uncertainty quantification models. Experiments were conducted in two distinct settings: an **in-domain (ID)** setting to test for aleatoric uncertainty, where models were trained and tested on the same acceleration rates (4x, 8x, 16x), and an **out-of-domain (OOD)** setting, where models were trained only on 4x accelerated data but tested on 4x, 8x, and 16x data. This OOD setting introduces additional epistemic uncertainty, as the model encounters test data that differs from its training distribution. PHiRec was compared against several strong baselines for uncertainty quantification, including MC Dropout, a heteroscedastic variance approach, a combination of the two, an ensemble-based method, and the Probabilistic U-Net.

In terms of pure reconstruction quality, all methods performed similarly, with performance degrading as acceleration rates increased, as expected. PHiRec slightly underperformed in terms of peak signal-to-noise ratio (PSNR) but performed similarly in terms of structural similarity index measure (SSIM).

The crucial evaluation focused on the **quality of the uncertainty quantification**. A key metric for this is calibration, which measures the correlation between the predicted uncertainty and the actual model error. To assess the spatial correlation, the Normalized Cross-Correlation (NCC) between the model's variance map and the squared error map was computed. The results clearly showed that **PHiRec is substantially better calibrated than the baseline methods** in all settings. Furthermore, the study measured whether uncertainty correctly increased with higher acceleration rates. Surprisingly, PHiRec was the **only model for which the uncertainty consistently increased** as the acceleration rate became higher; for most other models, the uncertainty counter-intuitively decreased in some cases. These quantitative findings show that the model is particularly suitable for uncertainty propagation.

Finally, in the uncertainty propagation experiment, PHiRec also demonstrated superior performance. The uncertainty of the downstream segmentation task was estimated and compared to the actual segmentation error. Again, **PHiRec clearly outperformed all baseline methods** in terms of the calibration of its propagated uncertainty, showing a high correlation between the predicted segmentation variance and the true segmentation error. This was, to the authors' knowledge, the first work to explore and quantitatively evaluate the propagation of uncertainties from DL-based MRI reconstruction to a downstream.

The preceding work successfully established a foundational capability: a formal mechanism for propagating uncertainty through a diagnostic pipeline using Monte Carlo sampling. This demonstration, however, naturally leads to the next critical question. It is not enough to simply pass uncertainty forward; for this information to be clinically relevant, it must be actively used. Therefore, we now shift our focus from the how of propagation to the why: can we leverage this sampling-based uncertainty to build more robust and accurate diagnostic models? The following chapter directly investigates this by exploring whether a full distribution of possibilities can lead to a better clinical outcome than a single, deterministic prediction.

3.2 Leveraging probabilistic segmentation models for improved glaucoma diagnosis: A clinical pipeline approach

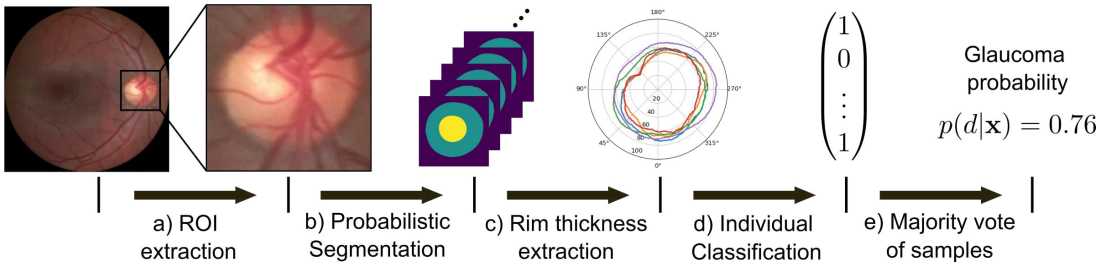

Figure 3.2: The proposed five-step pipeline: a) ROI extraction, b) Probabilistic segmentation to generate samples, c) Rim thickness curve (RTC) extraction from each sample, d) Individual classification of each RTC, and e) Majority vote over samples to obtain the final glaucoma probability. Figure from [116].

Motivation

Glaucoma is a chronic eye disease characterized by the gradual degeneration of nerve fibers, which leads to damage in the optic nerve head. As the second leading cause of blindness worldwide, affecting a significant portion of the elderly population, its early and accurate diagnosis is of paramount clinical importance [96]. The diagnostic process heavily relies on the visual assessment of retinal fundus images, specifically the precise localization and measurement of the optic disc and optic cup [73]. The shape, size, and spatial relationship of these structures, often summarized by metrics like the cup-to-disc ratio (CDR), are crucial disease markers [49]. Consequently, the automated segmentation of the cup and disc is a vital first step for any computer-aided diagnosis framework for glaucoma as we can infer these metrics directly from the segmentation.

However, this segmentation task is exceptionally challenging due to the inherent ambiguity in delineating these structures, particularly the optic cup. This ambiguity is not just a technical challenge for algorithms but reflects a clinical reality, with significant disagreements observed even among highly trained human experts [2, 61]. Most existing automated segmentation methods are deterministic, providing only a single "best-guess" segmentation and failing to account for this critical uncertainty [89, 97, 107]. This can lead to misleadingly confident predictions and potential misdiagnosis.

Furthermore, the majority of recent automated glaucoma diagnosis systems have been developed as end-to-end "black box" models [24, 28, 74, 99]. While these models can achieve high classification performance, they often lack interpretability, which makes them difficult to debug and less likely to gain acceptance in clinical practice, where understanding the reasoning behind a diagnosis is crucial. An alternative and more clinically aligned approach is to construct a multi-stage diagnostic pipeline that mirrors the individual steps of a clinician's reasoning process. Such pipelines, which first segment the relevant structures and then extract diagnostic markers, have shown great promise [23]. However, these pipelines have often relied on the traditional CDR as the primary diagnostic marker [1, 13]. Clinical literature suggests that the rim thickness, which is a measure of the distance between the disc and cup borders at every angle, may be a more informative and robust marker for glaucoma [62, 103].

Finally, while probabilistic segmentation models can expose the ambiguity in delineating retinal structures, the critical question remains: how can this uncertainty be actively leveraged to improve the final diagnosis? The prevailing pipeline approach is to take a single "best-guess" segmentation, compute biomarkers, and make a classification, effectively discarding all the rich information contained within the uncertainty. To our knowledge, few studies have explored whether embracing the full distribution of segmentation possibilities, and marginalizing over this uncertainty, can lead to a more robust and accurate downstream classification. This work addresses these interconnected gaps by proposing an interpretable, multi-stage pipeline for glaucoma diagnosis that explicitly leverages propagated segmentation uncertainty, using modern probabilistic methods and introducing the clinically motivated Rim Thickness Curve as a superior diagnostic feature.

Methods

The paper proposes a five-step, clinically inspired pipeline for glaucoma diagnosis that explicitly incorporates and leverages segmentation uncertainty. In the following, we describe the key components of the pipeline.

1. Automated Region of Interest (ROI) Detection: The process begins by deterministically extracting the relevant ROI from the full fundus image. Since the optic nerve head containing the cup and disc occupies only a small portion of the image, this two-step approach improves the subsequent segmentation accuracy. A standard U-Net is first used to perform a coarse segmentation of the cup and disc on the full-view image. A padded, square bounding box is then placed around the predicted segmentation, and this region is cropped and resized to a standard 320×320 pixel image, which serves as the input $\mathbf{x}$ for the rest of the pipeline.

2. Probabilistic Optic Cup and Disc Segmentation: Recognizing the large uncertainty in cup and disc delineation, this core step uses probabilistic segmentation models to approximate the full distribution of possible segmentations, $p(\mathbf{s}|\mathbf{x})$, rather than producing a single, deterministic output. The paper evaluates and compares four different widely used techniques for this task:

- **Probabilistic U-Net** [59] and **PHiSeg** [8]: These are generative models based on conditional variational autoencoders (cVAEs) that are designed to capture the *aleatoric* uncertainty inherent in the data and task ambiguity.
- **MC Dropout** [35] and **Ensembles** [63]: These methods are used to estimate the *epistemic* uncertainty. Two types of ensembles were created: one by training ten U-Nets on the same data but with different random initializations, and another by training a separate U-Net for each of the 11 expert annotators available in the data.

All these methods allow for the generation of multiple segmentation samples, $\mathbf{s}_i$, from the estimated distribution $p(\mathbf{s}|\mathbf{x})$.

3. Rim Thickness Curve (RTC) Extraction: For each segmentation sample $\mathbf{s}_i$, a corresponding RTC, $\mathbf{r}_i$, is extracted. The RTC measures the width of the neuroretinal rim, which is defined as the distance between the optic cup and disc borders. This is computed geometrically by rotating a beam centered at the optic cup through 360 degrees and, at every half-degree interval, calculating the Euclidean distance between the points where the beam intersects the cup and disc borders. This process yields a 720-dimensional data vector for each segmentation sample. This deterministic extraction procedure is denoted as a function $g : \mathbf{s}_i \mapsto \mathbf{r}_i$.

4. Glaucoma Classification: Each extracted RTC sample $\mathbf{r}_i$ is then fed into a classifier to obtain a binary diagnosis, $d_i \in 0, 1$, of "glaucoma suspect" or "not glaucoma suspect". A logistic regression classifier was used, as more complex models did not show improvements in preliminary experiments. To prevent overfitting, the 720-dimensional RTC data is first reduced to 72 dimensions by averaging values within 72 bins. The classification function, $f : \mathbf{r}_i \mapsto d_i$, is trained separately for the RTCs generated by each of the different probabilistic segmentation methods.

5. Uncertainty-Aware Robust Classification: The final step leverages the propagated uncertainty to produce a robust final diagnosis. The probability of the image being "glaucoma suspect" is obtained by marginalizing over all possible segmentations, which is approximated using the Monte Carlo method with the generated samples: $p(d|\mathbf{x}) = \int p(\mathbf{s}|\mathbf{x}) f(g(\mathbf{s})) d\mathbf{s} \approx \frac{1}{N} \sum_{i=1}^{N} f(g(\mathbf{s}_i))$. This final probability, derived from the agreement or disagreement across the multiple segmentation possibilities, serves as a natural and interpretable measure of diagnostic uncertainty. The final binary prediction is made by thresholding this probability at 0.5.

Results

The proposed pipeline and its components were rigorously evaluated using two public datasets: the Cháksu dataset [61], which contains fundus images with annotations and diagnoses from five experts. Additionally, we used the RIGA dataset [2] to augment the training data for the segmentation models. The key findings are summarized below.

Uncertainty Quantification Improves Downstream Predictions: A central finding was that explicitly modeling and propagating uncertainty leads to superior diagnostic performance. We observed that **all pipelines using probabilistic segmentation techniques consistently achieved better downstream classification performance** (measured by AUROC) compared to the baseline pipeline that used a deterministic U-Net for segmentation. The 'Ensemble_seeds' approach achieved the highest overall AUROC score of 0.899. This result demonstrates that simply achieving a high Dice score in segmentation is insufficient; considering the entire probability distribution of possible segmentations provides crucial information that improves the final diagnosis. Notably, the best-performing pipeline also slightly outperformed a standard "black box" ResNet50 classifier trained directly on the image ROIs, suggesting that the interpretable pipeline approach can be highly competitive.

Rim Thickness Curves (RTC) Outperformed CDR: The study empirically verified the hypothesis that the **RTC is a more informative biomarker** for glaucoma than the commonly used cup-to-disc ratio (CDR). When the pipeline was run using the area CDR instead of the RTC, the resulting AUROC scores were consistently lower. The improvement was particularly large for the best-performing 'Ensemble_seeds' model, confirming that the choice of a clinically motivated feature like the RTC is critical for building a high-performing diagnostic pipeline.

Propagated Uncertainty Correlates with Expert Disagreement: To assess whether the model's uncertainty was clinically meaningful, the pipeline's final probability output $p(d|x)$ was compared to the mean expert diagnosis (i.e., the proportion of experts who labeled a case as "glaucoma suspect") in terms of correlation. **The PHiSeg model achieved the highest correlation** coefficient, indicating that the uncertainty propagated through the pipeline effectively captures the same ambiguities and disagreements that are present among human experts.

Qualitative Analysis Confirms Findings: Qualitative results further support these conclusions. The entropy maps generated by the probabilistic models (e.g., PHiSeg) show high uncertainty in the same regions where the expert annotators disagreed. Similarly, the predicted RTCs show greater variance (wider shaded bands) in cases that were clinically ambiguous or where structures like blood vessels obscured the rim, matching the variance seen in the experts' RTCs. This visual evidence confirms that the model learns to be uncertain in a **clinically meaningful way**.

We demonstrated how propagated, sampling based uncertainty can be successfully leveraged within a diagnostic pipeline to improve clinical classification accuracy. However, a persistent challenge in this approach is that while the heuristic uncertainty estimates showed a strong correlation with model error, they came with no formal guarantees of reliability. For any method to be trusted in a safety critical application, this is a significant limitation. This realization prompted us to take a step back and address a more fundamental question: how can we ensure that the uncertainties we intend to propagate are themselves trustworthy and well calibrated, perhaps even with mathematical guarantees? The following chapter tackles this problem directly, introducing a novel method to achieve formal risk control for different regions in an image regression task, even when those regions are unknown at test time.

3.3 Subgroup-Specific Risk-Controlled Dose Estimation in Radiotherapy

Motivation

Radiotherapy (RT) is a cornerstone of cancer treatment. Recent technological advancements, such as the magnetic resonance-guided linear accelerator (MR-Linac), integrate MR imaging directly with RT, allowing treatment plans to be adapted based on the patient's anatomy at the time of treatment [39]. This capability for adjustments holds the potential for more precise tumor targeting and improved treatment outcomes. However, a major bottleneck in realizing this potential is the need for fast and accurate dose deposition calculations. While standard Monte Carlo simulations, the standard method for dose estimation, provide highly accurate results, they are computationally intensive, often taking minutes to hours to complete, which is too slow for on-the-fly adjustments [17].

Deep learning (DL) frameworks have emerged as a promising solution, offering the ability to perform dose estimation with both high speed and accuracy [60, 71, 81]. However, despite the high-risk nature of RT, where prediction errors can have severe clinical consequences, prior work in DL-based dose estimation has not adequately addressed risk assessment and uncertainty quantification (UQ). While various methods for UQ exist, such as deep ensembles or variational autoencoders, they lack formal guarantees about the correctness or reliability of their uncertainty estimates which is a critical limitation in a safety-critical field like RT.

A more rigorous, model-agnostic strategy for UQ is Risk-Controlling Prediction Sets (RCPS) [7]. As introduced in Section 2.4, the RCPS framework allows one to construct a prediction interval with a formal mathematical guarantee that the true value is contained within that interval with a user-defined probability. These intervals hold the potential to reliably indicate poor model performance by becoming excessively large, signaling that the model's prediction may not be trustworthy for a high-risk decision.

However, RCPS has a significant limitation: it can only provide these guarantees on a global level, for the entire data distribution. In many applications, especially in medical imaging, it is crucial to have guarantees for specific subgroups within the data. For instance, in RT dose estimation, one would want the dose prediction to be well-calibrated both in the background regions and, more importantly, in the high-dose region along the radiation beam. If there is a large imbalance between these subgroups (e.g., many more background voxels than high-dose regions), a naive application of RCPS will be dominated by the majority group (the background) and may fail to meet the required guarantees for the minority, but clinically critical, subgroup (the high-dose area). Simply calibrating each subgroup separately is not a solution if the subgroup membership is unknown at test time, which is the case in dose prediction. This paper addresses this critical gap by proposing a novel calibration algorithm that extends RCPS to provide risk guarantees for multiple subgroups, even when subgroup membership is not known at inference.

Methods

This work proposes an extension of the Risk-Controlling Prediction Sets (RCPS) framework, named **Subgroup-RCPS (SG-RCPS)**, which allows for controlling prediction risk across multiple subgroups even when their membership is unknown at test time. The framework is built upon a DL-based dose estimation model, DeepDose [60], which is a 3D U-Net architecture that takes a personalized RT treatment plan as input and outputs a voxel-wise dose prediction. The input consists of five channels: the CT image, beam shape, center beam line distance, source distance, and radiological depth.

To enable uncertainty quantification, the DeepDose network is first extended to provide heuristic prediction intervals using quantile regression [5]. Two additional output channels, $\tilde{l}(X)$ and $\tilde{u}(X)$, are added to the network to estimate the lower and upper bounds of a prediction interval, respectively. These heads are trained using a pinball loss, which is designed to estimate a specific quantile of the data distribution. The total training objective combines the standard Mean Squared Error (MSE) loss for the main dose prediction with the pinball losses for the upper and lower quantile estimators. This yields a model that outputs not just a point prediction for the dose, but also a heuristic interval for each voxel.

The core contribution is the **SG-RCPS calibration procedure**, which adjusts these heuristic intervals to provide formal risk guarantees. The risk is defined as the probability that the ground truth dose value Y falls outside the predicted interval $\mathcal{T}(X)$. The goal is to find a single non-negative scaling factor, $\hat{\lambda}$, that is applied to the heuristic interval widths such that the risk is controlled below a user-specified level α for *each subgroup simultaneously*.

The SG-RCPS algorithm, summarised in the paper's Algorithm 1, achieves this by evaluating an upper confidence bound (UCB) on the risk for each of the M subgroups separately, using a dedicated calibration set for each. The algorithm starts with a large λ (a very wide interval) and iteratively reduces it, checking at each step if the UCB for all M subgroups remains below the target risk level α. The process stops as soon as the risk criterion is violated for the *first* subgroup, and the previous, valid λ is chosen as the final scaling factor $\hat{\lambda}$. This ensures that the final prediction intervals are as narrow as possible while still jointly satisfying the risk guarantees for all subgroups, including under-represented ones.

Results

The SG-RCPS method was evaluated on a clinical dataset containing CT scans and RT treatment plans from 125 patients, covering five anatomical entities: prostate, liver, mamma (breast), head and neck (HN), and lymph nodes. The lymph node data was treated as an out-of-distribution (OOD) entity to assess generalization. The proposed SG-RCPS method was compared directly with the standard RCPS algorithm. Three subgroups were considered for calibration: the **foreground** (voxels within the high-dose beam), the **background** (voxels outside the beam), and the **total image**. The target risk level was set to $\alpha = 0.1$, meaning the goal was for the prediction interval to contain the true dose value in at least 90% of cases for each subgroup.

The results demonstrated a clear failure of the standard RCPS method. The standard RCPS calibration was dominated by the background class, which contains the vast majority of voxels. Consequently, while it sometimes controlled the risk for the total image or the background, it **failed to control the risk in the critical foreground (beam) subgroup for any of the anatomical entities**. For several entities, the empirical risk for the foreground was nearly 1.0, meaning the predicted intervals almost never contained the true dose value inside the beam.

In stark contrast, the proposed **SG-RCPS algorithm was able to control the risk substantially better across all areas**. For the total image and background subgroups, the empirical risk was 0.0, indicating the intervals were actually more conservative than required. Most importantly, for the critical **foreground subgroup**, the risk was successfully controlled to the desired levels for all entities except for a slight miss in the head & neck cases. Notably, the risk was also well-controlled for the OOD lymph node entity, demonstrating the method's ability to generalize.

A qualitative example for a liver tumor (Figure 3 in the presented paper [33]) visually confirms these findings, showing that intervals from SG-RCPS are visibly wider, particularly in the beam region, reflecting a more reliable quantification of uncertainty. By ensuring that guarantees hold for the most clinically relevant subgroup, the SG-RCPS method significantly increases the safety and trustworthiness of DL-based dose prediction for high-risk applications.

The previous work demonstrated the critical importance of calibration and introduced a novel method for achieving distribution free, subgroup specific guarantees on uncertainty estimates at test time. This provides the final necessary component for our framework: a source of formally trustworthy uncertainty. With this in place, our final project seeks to unify the core themes of this thesis. We now aim to construct a single, integrated system that puts everything together, demonstrating how the principles of propagation, calibration, and leveraging can be combined to build systems that are self aware of their own limitations, a foundational requirement for their trustworthy integration into clinical workflows.

3.4 CUTE-MRI: Conformalized Uncertainty-based framework for Time-adaptivE MRI

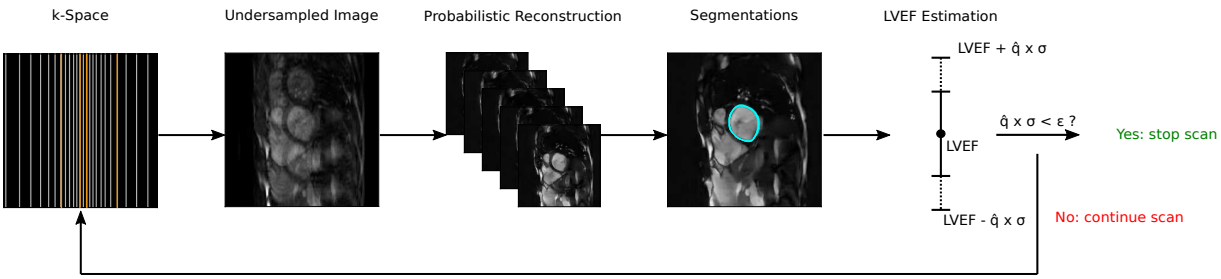

Figure 3.3: Overview of the proposed dynamic and iterative CUTE-MRI framework. At each time step, k-space data is acquired, a set of reconstructions and segmentations is generated, and a clinical metric (e.g., left ventricular ejection fraction (LVEF)) is estimated along with its calibrated uncertainty. The scan is stopped if the uncertainty is below a threshold ε. Figure from the original publication [31].

Motivation

As established in the first project of this thesis, the reconstruction of medical images from accelerated MRI scans is a severely ill posed inverse problem, introducing significant uncertainty that can propagate through and corrupt any downstream analysis. This final chapter investigates how principled uncertainty management can be integrated into the very fabric of the clinical workflow, moving beyond post hoc analysis to active, real time guidance. Successfully leveraging uncertainty in such a dynamic framework requires the integration of three essential components that have been the focus of investigation throughout this thesis: first, a mechanism to formally propagate uncertainty from reconstruction through any downstream tasks is necessary. Second, this propagated uncertainty must be rigorously calibrated to provide trustworthy, mathematical guarantees on its reliability. Finally, a clear strategy is needed to leverage this calibrated information to make a tangible, operational decision.

With these components in place, we can address a central, unaddressed limitation of current accelerated MRI methods: their reliance on static acquisition strategies with fixed, predetermined undersampling rates [48, 58, 68]. The acquisition process remains agnostic to the content of the image being acquired. This "one size fits all" approach is inherently inefficient. It may waste valuable scanner time on "easy" tasks that could have been resolved with fewer measurements, or conversely, it may terminate prematurely for "hard" cases, resulting in diagnostically inadequate images.

While significant research has focused on estimating and propagating uncertainty for post hoc analysis, its potential to actively guide and optimize the MRI acquisition process itself remains largely unexplored [27, 32, 76, 77, 94]. Prior work has typically focused on precalculating sampling trajectories or using simple image level metrics like signal-to-noise ratio as stopping criteria, without considering model confidence along the entire diagnostic pipeline from reconstruction to clinical endpoint [21, 44, 86, 113, 115]. This paper hypothesizes that by monitoring the uncertainty of a full diagnostic pipeline, one can create a patient specific, adaptive stopping rule for the scan. The core idea is to halt the acquisition precisely when the system reaches a predefined level of diagnostic confidence, rather than adhering to a fixed sampling budget, thereby optimizing scan duration for each individual without sacrificing diagnostic integrity.

Methods

This work introduces **CUTE-MRI**, a Conformalized Uncertainty-based framework for Time-adaptivE MRI, which dynamically adjusts scan time on a per-subject basis. The framework operates as an iterative loop that is executed after each new segment of k-space data is acquired. The scan is automatically terminated once the uncertainty of a downstream clinical metric falls below a user-defined precision threshold.

The CUTE-MRI pipeline proceeds as follows at each acquisition step:

1. **Probabilistic Reconstruction**: First, a set of M plausible image reconstructions, $\hat{\mathbf{x}}^{(m)}$, is generated from the currently undersampled k-space data. This is done using **PHiRec**, introduced in Section 3.1 of this thesis, a probabilistic reconstruction model chosen for its high sampling speed, which makes it suitable for real-time applications.
2. **Uncertainty Propagation**: Each of the M candidate reconstructions is passed through a full, deterministic downstream analysis pipeline. This involves applying a trained **U-Net** to obtain a set of M segmentation masks. From these segmentations, a final clinical metric of interest, $\mathbf{w}$, is computed. In the experiments, these metrics were patellar cartilage volume and left ventricular ejection fraction (LVEF) for knee MRI data and cardiac CINE MRI data respectively. This process results in a set of M metric samples, $\{\mathbf{w}^{(m)}\}$, which form an empirical estimate of the metric's posterior distribution.
3. **Uncertainty Calibration**: While the standard deviation of the metric samples, $\sigma_{\mathbf{w}}$, provides a heuristic measure of uncertainty, it lacks formal guarantees and was found to be poorly calibrated. To address this, the framework employs **split conformal prediction**, introduced in Section 2.4 to transform the heuristic uncertainty into a rigorous, calibrated confidence interval. A nonconformity score, defined as the normalized absolute error $sc_i = |\mathbf{w}_i - \hat{\mathbf{w}}_i|/\sigma_{\mathbf{w},i}$, is calculated for each sample in a dedicated calibration set. As described in the technical background section, $\hat{q}$ is being computed based on these scores.
4. **Stopping Criterion**: The final conformal prediction interval is constructed as $[\hat{\mathbf{w}} - \hat{q}\sigma_{\mathbf{w}}, \hat{\mathbf{w}} + \hat{q}\sigma_{\mathbf{w}}]$. By construction, this interval is guaranteed to contain the true metric value with a user-specified probability (e.g., 90%). The acquisition is terminated if the width of this interval falls below a predefined precision threshold, ε. If not, the scan continues to the next acquisition step.

Results

The CUTE-MRI framework was validated on two distinct clinical applications: estimating patellar cartilage volume from the public SKM-TEA knee MRI dataset [25] and computing Left Ventricular Ejection Fraction (LVEF) from an in-house cardiac CINE MRI dataset. The acquisition was simulated by retrospectively undersampling the k-space data across a range of acceleration factors from 4x to 32x.

The central finding of the study was that it is indeed **possible to create a patient specific, adaptive stopping rule** for MRI acquisition by monitoring the uncertainty of a downstream clinical metric. The CUTE-MRI framework successfully terminated scans at varied acceleration rates depending on the subject, demonstrating that an automated, content aware acquisition strategy is feasible. However, the reliability of this adaptive mechanism was found to be critically dependent on the formal calibration of the underlying uncertainty estimates.

A direct comparison of the system's behavior with and without conformal prediction revealed a stark difference. Without calibration, the system was systematically overconfident and consistently **terminated scans prematurely**. For the SKM-TEA dataset, the uncalibrated model stopped every single scan at the earliest possible opportunity (the 32x acceleration factor). This led to extremely poor statistical reliability; the uncalibrated intervals at the point of stopping achieved an empirical coverage of only 17.6% for the knee data and 20.0% for the cardiac data, far below the target of 90%. This demonstrates that while the adaptive framework is sound in principle, relying on uncalibrated, heuristic uncertainty poses a significant clinical risk by producing misleadingly narrow confidence intervals.

In contrast, the calibrated setting led to **significantly longer, more varied, and more appropriate scan durations**. The calibrated model reliably terminated scans at lower acceleration rates where the prediction error was safely below the task specific threshold. Calibration substantially improved the reliability of the intervals, increasing the empirical coverage to 61.1% for SKM-TEA and 85.7% for CINE. While still slightly below the 90% target, this represents a major improvement in safety and trustworthiness.

A qualitative analysis of the results further supported these findings. It was observed that cases for which the framework automatically chose **longer scan times corresponded to images with more challenging anatomy**

or visible reconstruction artifacts. Conversely, scans that were **terminated early were typically those with more clear reconstructions and accurate segmentations**. This provides strong evidence that the pipeline's uncertainty metric is clinically meaningful and correctly identifies cases that require more data. The entire pipeline was also shown to be computationally efficient, with an overhead of approximately 28 ms per slice, making the approach practical for real time clinical implementation.

This final contribution, therefore, closes the loop on the research presented in this thesis. It serves as a comprehensive proof of concept, synthesizing the principles of uncertainty propagation, leveraging, and calibration into a single, functional system that charts a course toward more adaptive, efficient, and ultimately trustworthy AI in clinical practice.

4. Conclusion

This thesis has advanced a framework centered on pipelines for managing uncertainty in medical machine learning. The overarching goal was to move beyond optimizing models for isolated tasks and instead develop systems that explicitly model, propagate, and leverage uncertainty across entire clinical workflows. The preceding chapters detailed the specific contributions that systematically built towards this goal, each addressing a fundamental component of the approach aware of the pipeline. This chapter now steps back to synthesize these individual contributions into a cohesive whole, interpret their collective findings, situate them within the broader scientific discourse, acknowledge their limitations, and chart a course for future inquiry.

To ground this discussion, we first summarize how each core publication contributed to the central thesis argument. Our research began in Section 3.1 by establishing the foundational capability of **uncertainty propagation**. We developed a probabilistic MRI reconstruction model, PHiRec, that can be used to pass uncertainty from an ill-posed problem to a downstream segmentation task. This work provided the essential proof of concept that uncertainty from a low level processing step can be preserved. Building on this, in Section 3.2 we demonstrated how propagated uncertainty can be **leveraged** for tangible clinical improvement. In our glaucoma diagnosis pipeline, we showed that embracing the full distribution of ambiguous segmentations led to a significant improvement in the accuracy of the final clinical classification. The subsequent work in Section 3.3 addressed the critical requirement of **trustworthiness** through formal **calibration**. We developed SG-RCPS, a novel algorithm that provides distribution free risk guarantees for specific, clinically critical subgroups that are unknown at test time, transforming heuristic uncertainty estimates into mathematically robust ones. Finally, Section 3.4 synthesized all these principles into a **closed loop, integrated system**. The CUTE-MRI framework showed how propagated, leveraged, and calibrated uncertainty could actively guide the clinical workflow itself, creating a dynamic MRI acquisition process specific to the patient.

Building on this integrated body of work, we will now discuss key themes that have emerged from this research.

4.1 The Case for Pipelines: Aligning AI with Clinical Reality

A central design choice underpinning all contributions in this thesis is the focus on multi-step diagnostic pipelines rather than isolated approaches. One might argue that a direct mapping from image acquisition to diagnosis could, in theory, minimize information loss. However, this perspective overlooks a crucial factor for real-world deployment: clinical trust and interpretability. The diagnostic process in medicine is inherently modular. Clinicians reason through a sequence of steps: assessing image quality, identifying anatomical structures and measuring biomarkers in order to arrive at a conclusion.

By structuring our ML systems to mirror these workflows, we open the "black box." Each component can be individually inspected, understood, and validated. Our work was deliberately set in these pipeline structures not as a concession, but as a principled decision to build systems that are more transparent and align with how clinicians already work. While our primary focus was on the behavior of uncertainty, the pipeline context was chosen to ensure our findings remain relevant to eventual clinical application. The field of medical image analysis has historically focused on optimizing isolated task performance. We argue that future work must increasingly focus on the interactions between tasks within these pipelines. Exploring how to leverage uncertainty to improve not just a single component but the holistic system's utility is a vital direction for building AI that clinicians can understand, trust, and confidently integrate into their practice.

4.2 The Blurring Lines of Uncertainty: Beyond Aleatoric and Epistemic

In our projects, we employed and developed various models for uncertainty quantification (UQ). A common entry point when selecting a UQ method is the classic distinction between aleatoric uncertainty (inherent data ambiguity) and epistemic uncertainty (model ambiguity). Theoretically, one might choose an aleatoric model for segmenting blurry boundaries and an epistemic one to detect out-of-distribution data from a new scanner.

However, recent findings in the literature suggest this distinction, while conceptually useful, may be less clear in practice. Many studies report that models designed for one type of uncertainty often perform surprisingly well at capturing the other, and a clear performance advantage for one class of models is rarely observed [32, 100, 116]. Our own research, as presented in Sections 3.1 and 3.2 corroborates this experience. Across our experiments, we did not find a consistent pattern where a specific class of UQ models clearly outperformed others for tasks notionally dominated by one type of uncertainty.

This raises an important question about the practical utility of this traditional dichotomy. More importantly, it underscores a critical point: the choice of UQ model is less important than the rigorous evaluation of the uncertainty it produces. Works that claim to model uncertainty must go beyond reporting predictive performance and provide direct measures of uncertainty quality. Metrics such as calibration error and prediction interval coverage should become standard practice. The focus must shift from simply choosing a UQ model to empirically validating its expressive power.

4.3 Model Performance and Uncertainty: Two Sides of the Same Coin

Ultimately, the goal of UQ in medicine is to inform clinical action. Expressive uncertainties are not merely an academic exercise. They are a direct reflection of a model's competence. A model that is "unsure" and produces large, well-calibrated uncertainty regions is not a failure of the UQ method. On the contrary, it is a successful diagnosis of an underlying base model that is not yet fit for clinical use.

In our work, we repeatedly observed that generating expressive uncertainties revealed that a model's confidence was often too low to be clinically acceptable [31, 33]. A large, calibrated uncertainty interval implies a high probability of error, rightfully leading a clinician to distrust the prediction. This suggests that the underlying predictive models in many medical AI applications may simply not be performant enough. This leads to an essential conclusion: model performance and uncertainty quantification should not be treated as separate research problems. Large uncertainties often mean large error propagation in a pipeline, with potentially drastic effects on the final diagnosis. Future work should therefore focus on the co-design of models and UQ frameworks. The objective should not be to simply get a higher score on a single metric, but to enhance robust model performance in a way that simultaneously reduces well-calibrated uncertainty, making both the prediction and its associated confidence more clinically useful.

4.4 The Fragile Guarantees of Calibration

In our pursuit of expressive uncertainties, we investigated calibration methods like Conformal Prediction and Risk-Controlling Prediction Sets (RCPS), which are highly attractive due to their model-agnosticism and ability to provide formal, mathematical guarantees on error rates [4, 7]. Such guarantees are immensely appealing for high-risk medical applications.

However, these methods rely on a critical assumption: exchangeability between the calibration and test datasets. In the heterogeneous and dynamic world of clinical data, this assumption is fragile and difficult to verify. Patient populations shift, imaging protocols evolve, and hardware is upgraded. All this can introduce distribution shifts that violate this core requirement. When the assumption is not met, the promised guarantees may not hold, potentially giving users a false sense of security.

Across several of our experiments, we observed this pattern: while calibration methods often improved performance on the test set, they frequently failed to meet the precise theoretical guarantees they held on the calibration data [31, 33]. This is a critical real-world limitation. To leverage the full potential of these powerful methods, future work must proceed on two fronts. First, on the data side, a greater emphasis on data curation by design is needed to create datasets that better approximate the exchangeability assumption. Second, and equally important, is clear communication. Clinicians must be educated on the statistical nature of these guarantees. They are probabilistic and contingent on assumptions. We need to ensure that clinicians can interpret them correctly and understand both their power and their limitations.

4.5 The Challenge of Interacting Uncertainties

This thesis successfully demonstrated the propagation of uncertainty from a single upstream source through downstream tasks. However, this represents a simplified view of the complex reality within a diagnostic pipeline. Uncertainty does not typically arise at a single point; it can be introduced and compounded at every stage. For instance, in our MRI workflow, uncertainty arises from the reconstruction of an undersampled image, but a single, perfectly clear reconstruction can still contain anatomical structures with inherently ambiguous boundaries, introducing new uncertainty at the segmentation stage.

We did not explicitly model these complex interactions of uncertainties. Doing so poses a significant computational challenge. Our Monte Carlo sampling approach, while effective for serial propagation, scales exponentially. A pipeline with five probabilistic stages, each requiring M samples to characterize its uncertainty, would necessitate M^5 samples to model the full joint distribution for a single prediction, quickly becoming intractable.

Overcoming this computational barrier is a critical frontier for future research. While our sampling-based approach is theoretically sound, practical systems will require more efficient methods. Promising directions include exploring techniques from using surrogate models to estimate the outcome distribution, thereby avoiding the exponential cost of explicit sampling [20].

4.6 Summary and Outlook

This thesis sought to move the conversation in medical AI from the optimization of isolated tasks to the holistic management of uncertainty within integrated diagnostic pipelines. We argued that the medical diagnostic process is a cascade of dependent tasks, rich with inherent ambiguity. To build trustworthy ML systems, we must explicitly model, propagate, and leverage this uncertainty. Through our contributions, we have shown how uncertainties can be propagated from acquisition to analysis, made more expressive through calibration, and leveraged to actively improve and guide the diagnostic process.

This work offers a conceptual and methodological shift away from a narrow focus on performance metrics. By treating diagnosis as an interconnected system, this thesis has aimed to model uncertainty in the bigger picture, identifying both solutions and remaining gaps in creating reliable clinical decision-support tools. It is our hope that this pipeline-aware perspective encourages the research community to look beyond the isolated task and continue building the foundations for a more robust and trustworthy generation of medical AI.

Bibliography

[1] B. Al-Bander, B. M. Williams, W. Al-Nuaimy, M. A. Al-Taee, H. Pratt, and Y. Zheng. "Dense Fully Convolutional Segmentation of the Optic Disc and Cup in Colour Fundus for Glaucoma Diagnosis". *Symmetry* 10.4 (2018).

[2] A. Almazroa, S. Alodhayb, E. Osman, E. Ramadan, M. Hummadi, M. Dlaim, M. Alkatee, K. Raahemifar, and V. Lakshminarayanan. "Retinal fundus images for glaucoma analysis: the RIGA dataset". *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*. Edited by J. Zhang and P.-H. Chen. Volume 10579. International Society for Optics and Photonics. *SPIE*, 2018, 105790B.

[3] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. "Concrete Problems in AI Safety". 2016. arXiv: 1606.06565[cs]. URL: http://arxiv.org/abs/1606.06565 (visited on 09/01/2025).

[4] A. N. Angelopoulos and S. Bates. "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification". 2022. arXiv: 2107.07511[cs]. URL: http://arxiv.org/abs/2107.07511 (visited on 09/01/2025).

[5] A. N. Angelopoulos, A. P. Kohli, S. Bates, M. I. Jordan, J. Malik, T. Alshaabi, S. Upadhyayula, and Y. Romano. "Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging". arXiv:2202.05265 [cs, eess, q-bio, stat]. 2022. URL: http://arxiv.org/abs/2202.05265 (visited on 02/05/2023).

[6] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen. "On instabilities of deep learning in image reconstruction and the potential costs of AI". *Proceedings of the National Academy of Sciences* 117.48 (2020), pages 30088–30095. (Visited on 08/29/2025).

[7] S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan. "Distribution-Free, Risk-Controlling Prediction Sets". 2021. arXiv: 2101.02703[cs]. URL: http://arxiv.org/abs/2101.02703 (visited on 09/01/2025).

[8] C. F. Baumgartner, K. C. Tezcan, K. Chaitanya, A. M. Hötker, U. J. Muehlematter, K. Schawkat, A. S. Becker, O. Donati, and E. Konukoglu. "PHiSeg: Capturing Uncertainty in Medical Image Segmentation". *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Edited by D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan. Volume 11765. Series Title: Lecture Notes in Computer Science. Cham: *Springer International Publishing*, 2019, pages 119–127. (Visited on 09/01/2025).

[9] E. Begoli, T. Bhattacharya, and D. Kusnezov. "The need for uncertainty quantification in machine-assisted medical decision making". *Nature Machine Intelligence* 1.1 (2019), pages 20–23. (Visited on 09/01/2025).

[10] C. E. Bender, S. Bansal, D. Wolfman, and J. R. Parikh. "2018 ACR Commission on Human Resources Workforce Survey". *Journal of the American College of Radiology* 16.4 (2019), pages 508–512.

[11] V. Bentkus. "On Hoeffding's inequalities". *The Annals of Probability* 32.2 (2004). (Visited on 09/01/2025).

[12] M. Bertero, P. Boccacci, and C. De Mol. "Introduction to Inverse Problems in Imaging". Second edition. Boca Raton London New York: *CRC Press*, 2022. 1 page.

[13] L. Bi, Y. Guo, Q. Wang, D. Feng, M. Fulham, and J. Kim. "Automated Segmentation of the Optic Disk and Cup using Dual-Stage Fully Convolutional Networks". 2019. arXiv: 1902.04713 [cs.CV].

[14] C. M. Bishop. "Pattern Recognition and Machine Learning". Information science and statistics. New York: *Springer*, 2006. 738 pages.

[15] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. "Variational Inference: A Review for Statisticians". *Journal of the American Statistical Association* 112.518 (2017), pages 859–877. (Visited on 09/01/2025).

[16] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. "Weight Uncertainty in Neural Network". *Proceedings of the 32nd International Conference on Machine Learning*. Edited by F. Bach and D. Blei. Volume 37. Proceedings of Machine Learning Research. Lille, France: *PMLR*, 2015, pages 1613–1622.

[17] G. Bol, S Hissoiny, J. Lagendijk, and B. Raaymakers. "Fast online Monte Carlo-based IMRT planning for the MRI linear accelerator". *Physics in Medicine & Biology* 57.5 (2012), page 1375.

[18] F. Calivá, A. P. Leynes, R. Shah, U. U. Bharadwaj, S. Majumdar, P. E. Larson, and V. Pedoia. "Breaking Speed Limits with Simultaneous Ultra-Fast MRI Reconstruction and Tissue Segmentation". *Medical Imaging with Deep Learning*. PMLR. 2020, pages 94–110.

[19] E. Candes, J. Romberg, and T. Tao. "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information". *IEEE Transactions on Information Theory* 52.2 (2006), pages 489–509. (Visited on 08/29/2025).

[20] K. Cranmer, J. Brehmer, and G. Louppe. "The frontier of simulation-based inference". *Proceedings of the National Academy of Sciences* 117.48 (2020), pages 30055–30062. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1912789117.

[21] P. Daudé, R. Ramasawmy, A. Javed, R. J. Lederman, K. Chow, and A. E. Campbell-Washburn. "Inline automatic quality control of 2D phase-contrast flow MRI for subject-specific scan time adaptation". *Magnetic Resonance in Medicine* 92.2 (2024), pages 751–760.

[22] E. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer, and P. Hennig. "Laplace Redux – Effortless Bayesian Deep Learning". Version Number: 3. 2021. URL: https://arxiv.org/abs/2106.14806 (visited on 09/01/2025).

[23] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". *Nature medicine* 24.9 (2018), pages 1342–1350.

[24] C. De Vente, K. A. Vermeer, N. Jaccard, H. Wang, H. Sun, F. Khader, D. Truhn, T. Aimyshev, Y. Zhanibekuly, T.-D. Le, et al. "AIROGS: Artificial Intelligence for Robust Glaucoma Screening Challenge". *IEEE transactions on medical imaging* (2023).

[25] A. D. Desai, A. M. Schmidt, E. B. Rubin, C. M. Sandino, M. S. Black, V. Mazzoli, K. J. Stevens, R. Boutin, C. Ré, G. E. Gold, B. A. Hargreaves, and A. S. Chaudhari. "SKM-TEA: A Dataset for Accelerated MRI Reconstruction with Dense Image Labels for Quantitative Clinical Evaluation". 2022. arXiv: 2203.06823 [eess.IV].

[26] C. Doersch. "Tutorial on Variational Autoencoders". 2021. arXiv: 1606.05908[stat]. URL: http://arxiv.org/abs/1606.05908 (visited on 09/01/2025).

[27] V. Edupuganti, M. Mardani, S. Vasanawala, and J. Pauly. "Uncertainty Quantification in Deep MRI Reconstruction". *IEEE Transactions on Medical Imaging* 40.1 (2020), pages 239–250.

[28] R. Fan, K. Alipour, C. Bowd, M. Christopher, N. Brye, J. A. Proudfoot, M. H. Goldbaum, A. Belghith, C. A. Girkin, M. A. Fazio, J. M. Liebmann, R. N. Weinreb, M. Pazzani, D. Kriegman, and L. M. Zangwill. "Detecting Glaucoma from Fundus Photographs Using Deep Learning without Convolutions: Transformer for Improved Generalization". *Ophthalmology Science* 3.1 (2023), page 100233.

[29] L. F. Feiner, M. J. Menten, K. Hammernik, P. Hager, W. Huang, D. Rueckert, R. F. Braren, and G. Kaissis. "Propagation and Attribution of Uncertainty in Medical Imaging Pipelines". *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*. Springer. 2023, pages 1–11.

[30] J. A. Fessler. "Optimization Methods for Magnetic Resonance Image Reconstruction: Key Models and Optimization Algorithms". *IEEE Signal Processing Magazine* 37.1 (2020), pages 33–40. (Visited on 08/29/2025).

[31] P. Fischer, J. N. Morshuis, T. Küstner, and C. Baumgartner. "CUTE-MRI: Conformalized Uncertainty-based framework for Time-adaptivE MRI". 2025. arXiv: 2508.14952 [eess.IV]. URL: https://arxiv.org/abs/2508.14952.

[32] P. Fischer, K Thomas, and C. F. Baumgartner. "Uncertainty Estimation and Propagation in Accelerated MRI Reconstruction". *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*. Springer. 2023, pages 84–94.

[33] P. Fischer, H. Willms, M. Schneider, D. Thorwarth, M. Muehlebach, and C. F. Baumgartner. "Subgroup-Specific Risk-Controlled Dose Estimation in Radiotherapy". *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pages 696–706.

[34] S. Fort, H. Hu, and B. Lakshminarayanan. "Deep Ensembles: A Loss Landscape Perspective". 2020. arXiv: 1912.02757[stat]. URL: http://arxiv.org/abs/1912.02757 (visited on 09/01/2025).

[35] Y. Gal and Z. Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". *Proceedings of The 33rd International Conference on Machine Learning*. Edited by M. F. Balcan and K. Q. Weinberger. Volume 48. Proceedings of Machine Learning Research. New York, New York, USA: *PMLR*, 2016, pages 1050–1059.

[36] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam. "The false hope of current approaches to explainable artificial intelligence in health care". *The Lancet Digital Health* 3.11 (2021), e745–e750. (Visited on 08/29/2025).

[37] M. A. Griswold, P. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, and A. Haase. "Generalized autocalibrating partially parallel acquisitions (GRAPPA)". *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 47.6 (2002), pages 1202–1210.

[38] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. "On Calibration of Modern Neural Networks". *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: *JMLR.org*, 2017, 1321–1330.

[39] W. Hall, E. Paulson, U. van der Heide, C. Fuller, B. Raaymakers, J. Lagendijk, X. Li, D. Jaffray, L. Dawson, B Erickson, M Verheij, K. Harrington, A Sahgal, P Lee, P. Parikh, M. Bassetti, C. Robinson, B. Minsky, A Choudhury, R. Tersteeg, and C. Schultz. "MR Linac Atlantic Consortium and the ViewRay C2T2 Research Consortium. The transformation of radiation oncology using real-time magnetic resonance guidance: A review". *European Journal of Cancer* 122 (2019), pages 42–52.

[40] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll. "Learning a variational network for reconstruction of accelerated MRI data". *Magnetic Resonance in Medicine* 79.6 (2018), pages 3055–3071. (Visited on 08/29/2025).

[41] P. K. J. Han, W. M. P. Klein, and N. K. Arora. "Varieties of Uncertainty in Health Care: A Conceptual Taxonomy". *Medical Decision Making* 31.6 (2011), pages 828–838. (Visited on 08/29/2025).

[42] J. A. Harolds, J. R. Parikh, E. I. Bluth, S. C. Dutton, and M. P. Recht. "Burnout of Radiologists: Frequency, Risk Factors, and Remedies: A Report of the ACR Commission on Human Resources". *Journal of the American College of Radiology* 13.4 (2016), pages 411–416. (Visited on 08/29/2025).

[43] T. Hepp, S. Gatidis, K. Hammernik, and T. Küstner. "Uncertainty estimation via ensembling for deep learning-based MR image reconstruction". *ISMRM*. Volume 685. 2022.

[44] Z. Huang, J. Duan, Y. Xie, and Y. Liu. "UDNet: Unified Deep Network based on Transformer and Multi-stage Fusion for brain tumor classification from undersampled MRI". *Neurocomputing* 619 (2025), page 129109.

[45] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". *Nature Methods* 18.2 (2021), pages 203–211. (Visited on 08/29/2025).

[46] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, and P. F. Jäger. "nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation". *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Edited by M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel. Cham: *Springer Nature Switzerland*, 2024, pages 488–498.

[47] A. Jalal, M. Arvinte, G. Daras, E. Price, A. G. Dimakis, and J. I. Tamir. "Robust Compressed Sensing MRI with Deep Generative Priors". arXiv:2108.01368 [cs, math, stat]. 2021. URL: http://arxiv.org/abs/2108.01368 (visited on 02/05/2023).

[48] O. N. Jaspan, R. Fleysher, and M. L. Lipton. "Compressed sensing MRI: a review of the clinical literature". *The British journal of radiology* 88.1056 (2015), page 20150487.

[49] H. Jayaram, M. Kolko, D. S. Friedman, and G. Gazzard. "Glaucoma: now and beyond". *The Lancet* 402.10414 (2023), pages 1788–1801.

[50] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. "Deep Convolutional Neural Network for Inverse Problems in Imaging". *IEEE Transactions on Image Processing* 26.9 (2017), pages 4509–4522. (Visited on 08/29/2025).

[51] L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna. "Inter-observer variability of manual contour delineation of structures in CT". *European Radiology* 29.3 (2019), pages 1391–1399. (Visited on 08/29/2025).

[52] S. J. Julier and J. K. Uhlmann. "New extension of the Kalman filter to nonlinear systems". AeroSense '97. Edited by I. Kadar. Orlando, FL, USA, 1997, page 182. (Visited on 09/01/2025).

[53] J. P. Kassirer. "Our Stubborn Quest for Diagnostic Certainty". *New England Journal of Medicine* 320.22 (1989), pages 1489–1491. (Visited on 08/29/2025).

[54] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. "Key challenges for delivering clinical impact with artificial intelligence". *BMC Medicine* 17.1 (2019), page 195. (Visited on 08/29/2025).

[55] A. Kendall and Y. Gal. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" *Advances in Neural Information Processing Systems*. Edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Volume 30. *Curran Associates, Inc.*, 2017.

[56] D. P. Kingma and M. Welling. "Auto-Encoding Variational Bayes". 2013. arXiv: 1312.6114[stat]. URL: http://arxiv.org/abs/1312.6114 (visited on 09/01/2025).

[57] A. D. Kiureghian and O. Ditlevsen. "Aleatory or epistemic? Does it matter?" *Structural Safety* 31.2 (2009), pages 105–112. (Visited on 08/29/2025).

[58] F. Knoll, K. Hammernik, C. Zhang, S. Moeller, T. Pock, D. K. Sodickson, and M. Akcakaya. "Deep-Learning Methods for Parallel Magnetic Resonance Imaging Reconstruction: A Survey of the Current Approaches, Trends, and Issues". *IEEE Signal Processing Magazine* 37.1 (2020), pages 128–140. (Visited on 08/29/2025).

[59] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. M. A. Eslami, D. Jimenez Rezende, and O. Ronneberger. "A Probabilistic U-Net for Segmentation of Ambiguous Images". *Advances in Neural Information Processing Systems*. Edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Volume 31. *Curran Associates, Inc.*, 2018.

[60] C Kontaxis, G. Bol, J. Lagendijk, and B. Raaymakers. "DeepDose: towards a fast dose calculation engine for radiation therapy using deep learning". *Physics in Medicine & Biology* 65.7 (2020), page 075013.

[61] J. H. Kumar, C. S. Seelamantula, J. Gagan, Y. S. Kamath, N. I. Kuzhuppilly, U Vivekanand, P. Gupta, and S. Patil. "Chákṣu: A glaucoma specific fundus image database". *Scientific data* 10.1 (2023), page 70.

[62] J. H. Kumar, C. S. Seelamantula, Y. S. Kamath, and R. Jampala. "Rim-to-Disc Ratio Outperforms Cup-to-Disc Ratio for Glaucoma Prescreening". *Scientific reports* 9.1 (2019), page 7099.

[63] B. Lakshminarayanan, A. Pritzel, and C. Blundell. "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles". *Advances in Neural Information Processing Systems*. Edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Volume 30. *Curran Associates, Inc.*, 2017.

[64] M.-H. Laves, S. Ihler, J. F. Fast, L. A. Kahrs, and T. Ortmaier. "Well-Calibrated Regression Uncertainty in Medical Imaging with Deep Learning". *Proceedings of the Third Conference on Medical Imaging with Deep Learning*. Edited by T. Arbel, I. Ben Ayed, M. de Bruijne, M. Descoteaux, H. Lombaert, and C. Pal. Volume 121. Proceedings of Machine Learning Research. *PMLR*, 2020, pages 393–412.

[65] J. Lewis. "Fast Normalized Cross-Correlation". *Ind. Light Magic* 10 (2001).

[66] Z.-P. Liang and P. C. Lauterbur. "Principles of Magnetic Resonance Imaging: A Signal Processing Perspective". In collaboration with IEEE Xplore (Online service) and IEEE Engineering in Medicine and Biology Society. IEEE press series on biomedical engineering 4. new York: *IEEE Press*, 2000. 1 page.

[67] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. "A survey on deep learning in medical image analysis". *Medical Image Analysis* 42 (2017), pages 60–88. (Visited on 08/29/2025).

[68] M. Lustig, D. Donoho, and J. M. Pauly. "Sparse MRI: The application of compressed sensing for rapid MR imaging". *Magnetic Resonance in Medicine* 58.6 (2007), pages 1182–1195. (Visited on 08/29/2025).

[69] D. J. C. MacKay. "A Practical Bayesian Framework for Backpropagation Networks". *Neural Computation* 4.3 (1992), pages 448–472. (Visited on 09/01/2025).

[70] D. J. C. MacKay. "The Evidence Framework Applied to Classification Networks". *Neural Computation* 4.5 (1992), pages 720–736. (Visited on 09/01/2025).

[71] S. Martinot, N. Bus, M. Vakalopoulou, C. Robert, E. Deutsch, and N. Paragios. "High-Particle Simulation of Monte-Carlo Dose Distribution with 3D ConvLSTMs". *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*. Springer. 2021, pages 499–508.

[72] M. T. McCann, K. H. Jin, and M. Unser. "Convolutional Neural Networks for Inverse Problems in Imaging: A Review". *IEEE Signal Processing Magazine* 34.6 (2017), pages 85–95. (Visited on 08/29/2025).

[73] F. A. Medeiros, L. M. Zangwill, C. Bowd, K. Mansouri, and R. N. Weinreb. "The Structure and Function Relationship in Glaucoma: Implications for Detection of Progression and Measurement of Rates of Change". *Investigative ophthalmology & visual science* 53.11 (2012), pages 6939–6946.

[74] D. Mirzania, A. C. Thompson, and K. W. Muir. "Applications of deep learning in detection of glaucoma: a systematic review". *European Journal of Ophthalmology* 31.4 (2021), pages 1618–1642.

[75] J. N. Morshuis, S. Gatidis, M. Hein, and C. F. Baumgartner. "Adversarial Robustness of MR Image Reconstruction under Realistic Perturbations". *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer. 2022, pages 24–33.

[76] J. N. Morshuis, M. Hein, and C. F. Baumgartner. "Segmentation-guided MRI reconstruction for meaningfully diverse reconstructions". *MICCAI Workshop on Deep Generative Models*. Springer. 2024, pages 180–190.

[77] J. N. Morshuis, C. Schlarmann, T. Küstner, C. F. Baumgartner, and M. Hein. "Mind the Detail: Uncovering Clinically Relevant Image Details in Accelerated MRI with Semantically Diverse Reconstructions". 2025. arXiv: 2507.00670 [eess.IV]. URL: https://arxiv.org/abs/2507.00670.

[78] A. N. Mumuni, F. Hasford, N. I. Udeme, M. O. Dada, and B. O. Awojoyogbe. "A SWOT analysis of artificial intelligence in diagnostic imaging in the developing world: making a case for a paradigm shift". *Physical Sciences Reviews* 9.1 (2024), pages 443–476. (Visited on 08/29/2025).

[79] D. Narnhofer, A. Effland, E. Kobler, K. Hammernik, F. Knoll, and T. Pock. "Bayesian Uncertainty Estimation of Learned Variational MRI Reconstruction". 2021. arXiv: 2102.06665 [eess.IV].

[80] R. M. Neal. "Bayesian Learning for Neural Networks". Berlin, Heidelberg: *Springer-Verlag*, 1996.

[81] A. Neishabouri, N. Wahl, A. Mairani, U. Köthe, and M. Bangert. "Long short-term memory networks for proton dose calculation in highly heterogeneous tissues". *Medical physics* 48.4 (2021), pages 1893–1908.

[82] Z. Obermeyer and E. J. Emanuel. "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine". *New England Journal of Medicine* 375.13 (2016), pages 1216–1219. (Visited on 08/29/2025).

[83] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett. "Deep Learning Techniques for Inverse Problems in Imaging". arXiv:2005.06001 [cs, eess, stat]. 2020. URL: http://arxiv.org/abs/2005.06001 (visited on 02/05/2023).

[84] T. A. Ozkan, A. T. Eruyar, O. O. Cebeci, O. Memik, L. Ozcan, and I. Kuskonmaz. "Interobserver variability in Gleason histological grading of prostate cancer". *Scandinavian Journal of Urology* 50.6 (2016), pages 420–424. (Visited on 08/29/2025).

[85] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. "Inductive Confidence Machines for Regression". *Machine Learning: ECML 2002*. Edited by T. Elomaa, H. Mannila, and H. Toivonen. Redacted by G. Goos, J. Hartmanis, and J. Van Leeuwen. Volume 2430. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: *Springer Berlin Heidelberg*, 2002, pages 345–356. (Visited on 09/01/2025).

[86] L. Pineda, S. Basu, A. Romero, R. Calandra, and M. Drozdzal. "Active MR k-space Sampling with Reinforcement Learning". *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pages 23–33.

[87] J. Platt. "Probabilistic Outputs for Support vector Machines and Comparisons to Regularized Likelihood Methods". 1999.

[88] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger. "SENSE: sensitivity encoding for fast MRI". *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 42.5 (1999), pages 952–962.

[89] H. A. Rasheed, T. Davis, E. Morales, Z. Fei, L. Grassi, A. De Gainza, K. Nouri-Mahdavi, and J. Caprioli. "RimNet: A Deep Neural Network Pipeline for Automated Identification of the Optic Disc Rim". *Ophthalmology Science* 3.1 (2023), page 100244.

[90] A. Reinke, M. D. Tizabi, C. H. Sudre, M. Eisenmann, T. Rädsch, M. Baumgartner, L. Acion, M. Antonelli, T. Arbel, S. Bakas, P. Bankhead, A. Benis, M. Blaschko, F. Buettner, M. J. Cardoso, J. Chen, V. Cheplygina, E. Christodoulou, B. Cimini, G. S. Collins, S. Engelhardt, K. Farahani, L. Ferrer, A. Galdran, B. v. Ginneken, B. Glocker, P. Godau, R. Haase, F. Hamprecht, D. A. Hashimoto, D. Heckmann-Nötzel, P. Hirsch, M. M. Hoffman, M. Huisman, F. Isensee, P. Jannin, C. E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, A. E. Kavur, H. Kenngott, J. Kleesiek, A. Kleppe, S. Kohler, F. Kofler, A. Kopp-Schneider, T. Kooi, M. Kozubek, A. Kreshuk, T. Kurc, B. A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A. L. Martel, P. Mattson, E. Meijering, B. Menze, D. Moher, K. G. M. Moons, H. Müller, B. Nichyporuk, F. Nickel, M. A. Noyan, J. Petersen, G. Polat, S. M. Rafelski, N. Rajpoot, M. Reyes, N. Rieke, M. Riegler, H. Rivaz, J. Saez-Rodriguez, C. I. Sánchez, J. Schroeter, A. Saha, M. A. Selver, L. Sharan, S. Shetty, M. v. Smeden, B. Stieltjes, R. M. Summers, A. A. Taha, A. Tiulpin, S. A. Tsaftaris, B. V. Calster, G. Varoquaux, M. Wiesenfarth, Z. R. Yaniv, P. Jäger, and L. Maier-Hein. "Common Limitations of Image Processing Metrics: A Picture Story". 2023. arXiv: 2104.05642[eess]. URL: http://arxiv.org/abs/2104.05642 (visited on 08/29/2025).

[91] D. J. Rezende, S. Mohamed, and D. Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". *Proceedings of the 31st International Conference on Machine Learning*. Edited by E. P. Xing and T. Jebara. Volume 32. Proceedings of Machine Learning Research 2. Bejing, China: *PMLR*, 2014, pages 1278–1286.

[92] C. P. Robert and G. Casella. "Monte Carlo Statistical Methods". Springer Texts in Statistics. New York, NY: *Springer New York*, 2004. (Visited on 09/01/2025).

[93] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Edited by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi. Volume 9351. Series Title: Lecture Notes in Computer Science. Cham: *Springer International Publishing*, 2015, pages 234–241. (Visited on 08/29/2025).

[94] J. Schlemper, D. C. Castro, W. Bai, C. Qin, O. Oktay, J. Duan, A. N. Price, J. Hajnal, and D. Rueckert. "Bayesian Deep Learning for Accelerated MR Image Reconstruction". *Machine Learning for Medical Image Reconstruction: First International Workshop, MLMIR 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 1*. Springer. 2018, pages 64–71.

[95] J. Schlemper, O. Oktay, W. Bai, D. C. Castro, J. Duan, C. Qin, J. V. Hajnal, and D. Rueckert. "Cardiac MR Segmentation from Undersampled k-space Using Deep Latent Representation Learning". *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer. 2018, pages 259–267.

[96] A. Sevastopolsky. "Optic Disc and Cup Segmentation Methods for Glaucoma Detection with Modification of U-Net Convolutional Neural Network". *Pattern Recognition and Image Analysis* 27.3 (2017), pages 618–624.

[97] A. Sevastopolsky, S. Drapak, K. Kiselev, B. M. Snyder, J. D. Keenan, and A. Georgievskaya. "Stack-U-net: refinement network for improved optic disc and cup image segmentation". *Medical Imaging 2019: Image Processing*. Volume 10949. SPIE. 2019, pages 576–584.

[98] G. Shafer and V. Vovk. "A tutorial on conformal prediction". 2007. arXiv: 0706.3188[cs]. URL: http://arxiv.org/abs/0706.3188 (visited on 09/01/2025).

[99] L. K. Singh, Pooja, H. Garg, and M. Khanna. "Deep learning system applicability for rapid glaucoma prediction from fundus images across various data sets". *Evolving Systems* 13.6 (2022), pages 807–836.

[100] F. B. Smith, J. Kossen, E. Trollope, M. van der Wilk, A. Foster, and T. Rainforth. "Rethinking Aleatoric and Epistemic Uncertainty". 2025. arXiv: 2412.20892 [cs.LG]. URL: https://arxiv.org/abs/2412.20892.

[101] K. Sohn, H. Lee, and X. Yan. "Learning Structured Output Representation using Deep Conditional Generative Models". *Advances in Neural Information Processing Systems*. Edited by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Volume 28. *Curran Associates, Inc.*, 2015.

[102] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther. "Ladder Variational Autoencoders". *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: *Curran Associates Inc.*, 2016, 3745–3753.

[103] G. L. Spaeth, J. Henderer, C. Liu, M. Kesen, U. Altangerel, A. Bayer, L. J. Katz, J. Myers, D. Rhee, and W. Steinmann. "The disc damage likelihood scale: reproducibility of a new method of estimating the amount of optic nerve damage caused by glaucoma." *Transactions of the American Ophthalmological Society* 100 (2002), page 181.

[104] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". *Journal of Machine Learning Research* 15.56 (2014), pages 1929–1958.

[105] K. C. Tezcan, N. Karani, C. F. Baumgartner, and E. Konukoglu. "Sampling Possible Reconstructions of Undersampled Acquisitions in MR Imaging With a Deep Learned Prior". *IEEE Transactions on Medical Imaging* 41.7 (2022), pages 1885–1896.

[106] A. A. Tolpadi, U. Bharadwaj, K. T. Gao, R. Bhattacharjee, F. G. Gassert, J. Luitjens, P. Giesler, J. N. Morshuis, P. Fischer, M. Hein, et al. "K2S Challenge: From Undersampled K-Space to Automatic Segmentation". *Bioengineering* 10.2 (2023), page 267.

[107] A. Tulsani, P. Kumar, and S. Pathan. "Automated segmentation of optic disc and optic cup for glaucoma assessment using improved UNET++ architecture". *Biocybernetics and Biomedical Engineering* 41.2 (2021), pages 819–832.

[108] U.S. Food and Drug Administration. "Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices". Publicly available database. Data downloaded as a CSV file for analysis. Accessed: 2025-08-01. 2023. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices.

[109] A. Vahdat and J. Kautz. "NVAE: A Deep Hierarchical Variational Autoencoder". 2021. arXiv: 2007.03898[stat]. URL: http://arxiv.org/abs/2007.03898 (visited on 09/01/2025).

[110] V. Vovk, A. Gammerman, and G. Shafer. "Algorithmic Learning in a Random World". Cham: *Springer International Publishing*, 2022. (Visited on 09/01/2025).

[111] D. E. J. Waddington, N. Hindley, N. Koonjoo, C. Chiu, T. Reynolds, P. Z. Y. Liu, B. Zhu, D. Bhutto, C. Paganelli, P. J. Keall, and M. S. Rosen. "On Real-time Image Reconstruction with Neural Networks for MRI-guided Radiotherapy". arXiv:2202.05267 [physics]. 2022. URL: http://arxiv.org/abs/2202.05267 (visited on 02/05/2023).

[112] E. Wan and R. Van Der Merwe. "The Unscented Kalman Filter for Nonlinear Estimation". *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373)*. Symposium on Adaptive Systems for Signal Processing Communications and Control. Lake Louise, Alta., Canada: *IEEE*, 2000, pages 153–158. (Visited on 09/01/2025).

[113] Z. Wang, B. Li, H. Yu, Z. Zhang, M. Ran, W. Xia, Z. Yang, J. Lu, H. Chen, J. Zhou, et al. "Promoting fast MR imaging pipeline by full-stack AI". *Iscience* 27.1 (2024).

[114] L. Wasserman. "All of Statistics: A Concise Course in Statistical Inference". Springer Texts in Statistics. New York, NY: *Springer New York*, 2004. (Visited on 09/01/2025).

[115] Z. Wu, T. Yin, Y. Sun, R. Frost, A. van der Kouwe, A. V. Dalca, and K. L. Bouman. "Learning Task-Specific Strategies for Accelerated MRI". *IEEE Transactions on Computational Imaging* (2024).

[116] A. M. Wundram, P. Fischer, S. Wunderlich, H. Faber, L. M. Koch, P. Berens, and C. F. Baumgartner. "Leveraging Probabilistic Segmentation Models for Improved Glaucoma Diagnosis: A Clinical Pipeline Approach". *Medical Imaging with Deep Learning*. 2024.

[117] L. Yang, I. C. Ene, R. Arabi Belaghi, D. Koff, N. Stein, and P. Santaguida. "Stakeholders' perspectives on the future of artificial intelligence in radiology: a scoping review". *European Radiology* 32.3 (2022), pages 1477–1495. (Visited on 08/29/2025).

[118] B. Zadrozny and C. Elkan. "Transforming Classifier Scores into Accurate Multiclass Probability Estimates". *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD02: The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton Alberta Canada: *ACM*, 2002, pages 694–699. (Visited on 09/01/2025).

[119] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. "UNet++: A Nested U-Net Architecture for Medical Image Segmentation". *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Edited by D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madabhushi. Volume 11045. Series Title: Lecture Notes in Computer Science. Cham: *Springer International Publishing*, 2018, pages 3–11. (Visited on 08/29/2025).

Appendix

Peer-Reviewed Publications

A.

Uncertainty Estimation and Propagation in Accelerated MRI Reconstruction

Paul Fischer[1(✉)], K. Thomas[2], and Christian F. Baumgartner[1]

[1] Cluster of Excellence – Machine Learning for Science, University of Tübingen, Tübingen, Germany
paul.fischer@uni-tuebingen.de
[2] Medical Image and Data Analysis Lab, University Hospital of Tübingen, Tübingen, Germany

Abstract. MRI reconstruction techniques based on deep learning have led to unprecedented reconstruction quality especially in highly accelerated settings. However, deep learning techniques are also known to fail unexpectedly and hallucinate structures. This is particularly problematic if reconstructions are directly used for downstream tasks such as real-time treatment guidance or automated extraction of clinical parameters (e.g. via segmentation). Well-calibrated uncertainty quantification will be a key ingredient for safe use of this technology in clinical practice. In this paper we propose a novel probabilistic reconstruction technique (PHiRec) building on the idea of conditional hierarchical variational autoencoders. We demonstrate that our proposed method produces high-quality reconstructions as well as uncertainty quantification that is substantially better calibrated than several strong baselines. We furthermore demonstrate how uncertainties arising in the MR reconstruction can be propagated to a downstream segmentation task, and show that PHiRec also allows well-calibrated estimation of segmentation uncertainties that originated in the MR reconstruction process.

1 Introduction

Fast magnetic resonance imaging (MRI) techniques play a vital role in clinical practice, allowing to scan an increased number of patients while alleviating patient discomfort caused by prolonged acquisition times. Highly accelerated MR acquisitions also hold the key unlocking novel applications such as real-time MR-guided radiation therapy [31], or shortened scans for directly estimating clinical parameters via segmentations potentially without human oversight [4,26,30].

In recent years, MRI reconstruction techniques relying on deep learning (DL) have gained substantial interest due to their excellent performance at very high acceleration rates [11,12,20] and ability to provide real-time reconstructions [8,35]. Although DL-based reconstructions often appear realistic and of high quality, they have also been shown to fail unexpectedly [7], and hallucinate structural details [18]. Crucially, they lack the ability to indicate regions

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-44336-7_9.

C. H. Sudre et al. (Eds.): UNSURE 2023, LNCS 14291, pp. 84–94, 2023.
https://doi.org/10.1007/978-3-031-44336-7_9

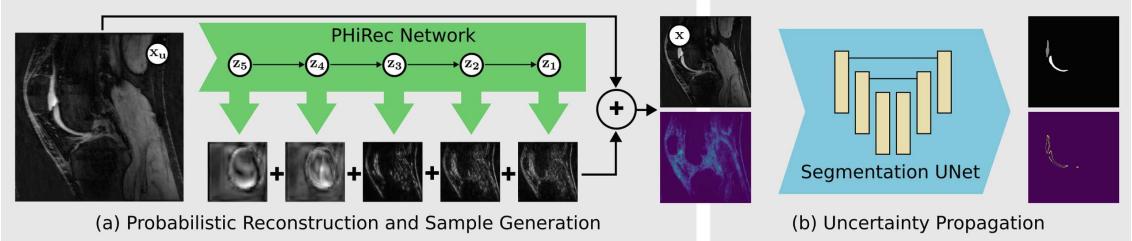

Fig. 1. (a) In our proposed Probabilistic Hierarchical Reconstruction (PHiRec) model, five latent variables $\mathbf{z}_l$ generate residual changes at different resolution scales. These residual changes are added to the undersampled input image $\mathbf{x}_u$ to generate the final output $\mathbf{x}$. The model can be used to sample likely reconstructions and to obtain reliable uncertainty estimates. (b) These samples can then be propagated to a subsequent segmentation network allowing to estimate the resulting segmentation uncertainty.

in the reconstructed images that are uncertain. This problem is exacerbated in scenarios where reconstructed images are used directly for downstream tasks, such as real-time treatment guidance [31], or extraction of clinical parameters via segmentation [30].

The reconstruction process inherently contains *aleotoric*, or irreducible, uncertainty due to the fact that a single undersampled acquisition can correspond to an infinite number of potential reconstructions with varying likelihoods [33]. Moreover, *epistemic* uncertainty may arise when the reconstruction model is applied outside the domain on which it was trained. Developing DL-based reconstruction techniques that are able to reflect those uncertainties is crucial, especially in applications without human oversight.

Several approaches have been proposed to model uncertainty of DL-based MRI reconstruction techniques. Schlemper et al. [25] proposed to estimate epistemic and aleotoric uncertainty using MC dropout and a heteroscedastic variance term, respectively. Hepp et al. [9] presented promising preliminary results for estimating epistemic uncertainty using an ensemble of networks. However, the approach does not scale well since the number of samples that can be generated is equal to the number of networks in the ensemble. Narnhofer et al. [19] modelled epistemic uncertainty by combining the previously proposed Total Deep Variation approach [14] with the Bayes-by-Backprop technique, which models every network weight as a Gaussian with individual mean and variance [3]. We note that this approach has heavy GPU-memory requirements and limits the complexity of the network architecture that can be used. Tezcan et al. [29] have proposed learning the prior distribution of MR images using a variational autoencoder (VAE) and using Markov-Chain Monte-Carlo (MCMC) to sample possible reconstructions. However, this approach is severely limited by the sampling times and is not suitable for real-time applications. Angelopoulos et al. [1] proposed an uncertainty quantification method based on conformal prediction, which offers a straightforward implementation and comes with mathematical guarantees. However, this method does not allow generating samples which could be used to explore potential reconstructions or propagate uncertainty to subsequent tasks. Recently diffusion models have demonstrated exceptional reconstruction performance [5,11,22,32]. While uncertainty quantification is feasible in those models, it is currently hindered by

extremely long sampling times rendering them unsuitable for real-time applications. In closely related work on a different imaging modality, Zhang et al. [34] proposed a PET reconstruction method based on conditional VAEs (cVAE) [27]. The approach employed a bottleneck architecture which only allows to model the uncertainty at a low spatial resolution and is prone to producing blurry reconstructions. Lastly, a major limitation of the existing literature is that none of the above studies present a quantitative evaluation of uncertainty quantification or a thorough comparison with baseline methods, instead relying solely on qualitative interpretation of the uncertainty maps.

In this paper, we propose a novel approach for estimating aleotoric uncertainty based on a hierarchical conditional VAE [27]. Hierarchical cVAEs have been shown to perform exceptionally well for estimating aleotoric uncertainty in segmentation tasks [2,15]. Specifically, they address two issues encountered in non-hierarchical cVAEs: blurry samples and limited expressivity to model high-dimensional spatial probability distributions [16,34]. However, despite their promise, hierarchical cVAEs remain unexplored in the context of MRI reconstruction. Here, we build on the Probabilistic Hierarchical Segmentation (PHiSeg) model by Baumgartner et al. [2], which was originally proposed for segmentation to create a novel *P*robabilistic *Hi*erarchical *R*econstruction technique which we coin PHiRec[1]. Our contributions are as follows:

- We propose PHiRec and show that it outperforms several strong baselines in terms of calibration of its uncertainty quantification.
- We demonstrate that the uncertainties originating in the MR reconstruction can be *propagated* to a downstream segmentation task in order to estimate the resulting segmentation uncertainties. This is, to the best of our knowledge, the first work to explore the propagation of uncertainties arising in DL-based MRI reconstruction to a downstream task.
- We present the first comprehensive *quantitative* evaluation of uncertainty quantification for MRI reconstruction contrasting several baselines.

2 Methods

We denote a fully-sampled MR image as $\mathbf{x} \in \mathbb{C}^N$ where N is the number of pixels. In a multi-coil MR acquisition, the acquired k-space data can be modelled as $\mathbf{y} = \mathcal{M}\mathcal{F}\mathcal{S}\mathbf{x} + \eta$, where $\mathcal{M}$ is an undersampling operator, $\mathcal{F}$ is the Fourier operator, $\mathcal{S}$ is an operator encoding the spatial sensitivity of each coil, and η is used to model thermal scanner noise. The goal of MRI reconstruction is to estimate the maximimum a-posteriori of the distribution $p(\mathbf{x}|\mathbf{y})$. For uncertainty estimation we are additionally interested in the spread of this distribution.

We pose the reconstruction as a de-aliasing problem by modelling the distribution $p(\mathbf{x}|\mathbf{x}_u)$, where the undersampled image $\mathbf{x}_u$ is obtained by applying the inverse Fourier operator to the zero-filled measurement data $\mathbf{y}$. In the following, we show how a hierarchical cVAE approach can be employed to model the distribution $p(\mathbf{x}|\mathbf{x}_u)$. As shown in Fig. 1, we model the distribution using a cVAE

[1] The code for PHiRec is available at https://github.com/paulkogni/MR-Recon-UQ.

that has $L = 5$ separate latent variables $\mathbf{z}_l$ each operating on a different resolution scale. For instance, $\mathbf{z}_1$ operates at the original image resolution, while $\mathbf{z}_5$ operates at a resolution that was four times downpooled by a factor of 2. Each resolution level is responsible for probabilistically generating residual changes that are added to the input image $\mathbf{x}_u$ in order to remove undersampling artifacts and obtain the reconstructed image $\mathbf{x}$. Our modelling assumption includes that the distribution of each $\mathbf{z}_l$ depends on the input image $\mathbf{x}_u$ as well as the latent variable of the resolution level below $\mathbf{z}_{l+1}$. This allows higher resolution levels to have a notion of what changes were already performed in the resolution level below. As was previously shown, this hierarchical approach is a very expressive model for capturing high-dimensional probability distributions [2,15]. Note that in contrast to the hierarchical cVAE methods developed in the context of segmentation [2,15], our proposed PHiRec model contains a skip connection from the input to the output which we found to facilitate the de-aliasing problem. This is reflected by the dependence of the likelihood $p(\mathbf{x}|\mathbf{z}_{1:L}, \mathbf{x}_u)$ on $\mathbf{x}_u$ in the equations below.

Using the above modelling assumptions $p(\mathbf{x}|\mathbf{x}_u)$ can be written as

$$p(\mathbf{x}|\mathbf{x}_u) = \int p(\mathbf{x}|\mathbf{z}_{1:L}, \mathbf{x}_u) p(\mathbf{z}_1|\mathbf{z}_2, \mathbf{x}_u) \dots p(\mathbf{z}_{L-1}|\mathbf{z}_L, \mathbf{x}_u) p(\mathbf{z}_L|\mathbf{x}_u) d\mathbf{z}_{1:L}.$$

Following the standard variational approach we maximise the evidence lower bound, $\text{ELBO}(\mathbf{x}|\mathbf{x}_u) := \log p(\mathbf{x}|\mathbf{x}_u) - KL(q(\mathbf{z}_{1:L}|\mathbf{x}, \mathbf{x}_u)||(p(\mathbf{z}_{1:L}|\mathbf{x}, \mathbf{x}_u))$, which is a lower bound on the true log likelihood. Using our model assumptions, and following the derivation in Baumgartner et al. [2], we can write the ELBO as

$$\begin{aligned}\text{ELBO}(\mathbf{x}|\mathbf{x}_u) =& \mathbb{E}_{q(\mathbf{z}_{1:L}|\mathbf{x}_u,\mathbf{x})} \left[\log p(\mathbf{x}|\mathbf{z}_{1:L}, \mathbf{x}_u)\right] - \alpha_L KL\left[q(\mathbf{z}_L|\mathbf{x}, \mathbf{x}_u)||p(\mathbf{z}_L|\mathbf{x}_u)\right] \\ & - \sum_{l=1}^{L} \alpha_l \mathbb{E}_{q(\mathbf{z}_{l+1}|\mathbf{x}_u,\mathbf{x})} \left[KL\left[q(\mathbf{z}_l|\mathbf{z}_{l+1}, \mathbf{x}, \mathbf{x}_u)||p(\mathbf{z}_l|\mathbf{z}_{l+1}, \mathbf{x}_u)\right]\right],\end{aligned}$$

where $\alpha_l := 4^{(l-1)}$ are heuristic weight terms to equalize the magnitude of the KL-terms of the different resolution levels. The prior and posterior distributions are modelled using axis-aligned Normal distributions

$$\begin{aligned}p(\mathbf{z}_l|\mathbf{z}_{l+1}, \mathbf{x}_u) =& \mathcal{N}\left(\mathbf{z}_l|\Phi_l^{(\mu)}(\mathbf{z}_{l+1}, \mathbf{x}_u), \Phi_l^{(\sigma)}(\mathbf{z}_{l+1}, \mathbf{x}_u)\right) \\ q(\mathbf{z}_l|\mathbf{z}_{l+1}, \mathbf{x}, \mathbf{x}_u) =& \mathcal{N}\left(\mathbf{z}_l|\Theta_l^{(\mu)}(\mathbf{z}_{l+1}, \mathbf{x}, \mathbf{x}_u), \Theta_l^{(\sigma)}(\mathbf{z}_{l+1}, \mathbf{x}, \mathbf{x}_u)\right),\end{aligned}$$

where $\Phi_l^{(\mu)}$, $\Phi_l^{(\sigma)}$, $\Theta_l^{(\mu)}$, $\Theta_l^{(\sigma)}$ are neural network functions that estimate each distribution's mean and variance. The likelihood of the final de-aliased reconstruction $p(\mathbf{x}|\mathbf{z}_{1:L}, \mathbf{x}_u)$ is also modelled as a Normal distribution with a fixed variance and a mean that is estimated using another neural network. While in principle a mathematically valid model could be implemented using any neural network architecture the method lends itself to implementation as a U-Net-like architecture with the prior and posteriors implemented as U-Net encoders, and the likelihood as a decoder. For simplicity here we use the architecture proposed

by Baumgartner et al. [2] with the aforementioned addition of a skip connection from input to output (see Fig. 1). A schematic of this architecture is shown in the supplementary materials.

We train the entire architecture end-to-end with pairs of undersampled images $\mathbf{x}_u$ and ground truth reconstructions $\mathbf{x}$ using the ELBO defined above as objective. After training the posterior network is no longer required. The prior network can be used to predict the means and standard deviations of the $\mathbf{z}_l$ variables. Given these values an arbitrary number of latent variable samples can be generated, and decoded using the likelihood network, to obtain final reconstruction samples. The mean prediction as well as the spread of the distribution $p(\mathbf{x}|\mathbf{x}_u)$ can then be calculated from these samples $\{\mathbf{x}_i\}$.

Uncertainty Propagation. Given a separately trained deterministic segmentation network $f : \mathbf{x} \mapsto \mathbf{s}$, we can furthermore estimate the distribution of the segmentations $\mathbf{s}$ given the undersampled image $\mathbf{x}_u$, $p(\mathbf{s}|\mathbf{x}_u)$, using the Monte Carlo method. Specifically, we can segment each of our reconstruction samples $\{\mathbf{x}_i\}$ using f and analyse the resulting distribution of segmentations empirically.

3 Experiments and Results

Baselines. We compared our proposed PHiRec technique to several baseline strategies for estimating aleotoric and epistemic uncertainty. Firstly, we compared with Schlemper et al.'s [25] approach for which we separately evaluated the epistemic uncertainty quantification based on MC Dropout, the aleatoric uncertainty estimation rooted in a heteroscedastic variance term, as well as the combination of the two approaches as originally described. Furthermore, we compared to the ensemble based approach for estimating epistemic uncertainty initially demonstrated by Hepp et al. [9]. Specifically, we created an ensemble of 20 separately trained reconstruction networks. Lastly, we extended the probabilistic U-Net [16] to MRI reconstruction using the same strategy as for PHiRec to allow a comparison to another cVAE-based method. In order to focus the evaluation on the uncertainty quantification mechanism rather than on architectural details, all baseline methods were implemented using U-Nets, or U-Net-like architectures in the case of the probabilistic U-Net and our PHiRec. Furthermore, to ensure a fair comparison to our proposed approach, we implemented a skip connection from input to output for all baselines. We used the mean and standard deviation calculated using 20 samples for all methods and in all experiments to obtain the final prediction and spread of the distribution, respectively.

Data. All experiments were performed on the Stanford Knee MRI Multi-Task Evaluation (SKM-TEA) dataset [6], which comprises raw multi-coil k-space data of knee scans, along with segmentations for six anatomical structures. We employed the supplied undersampling masks, which were designed with a Poisson-Disc pattern. The provided coil sensitivities were used in a SENSE reconstruction [23] to obtain the fully-sampled ground truth reconstruction $\mathbf{x}$ as well as the undersampled network inputs $\mathbf{x}_u$. We divided the dataset into a training, validation, and test set using the official splits.

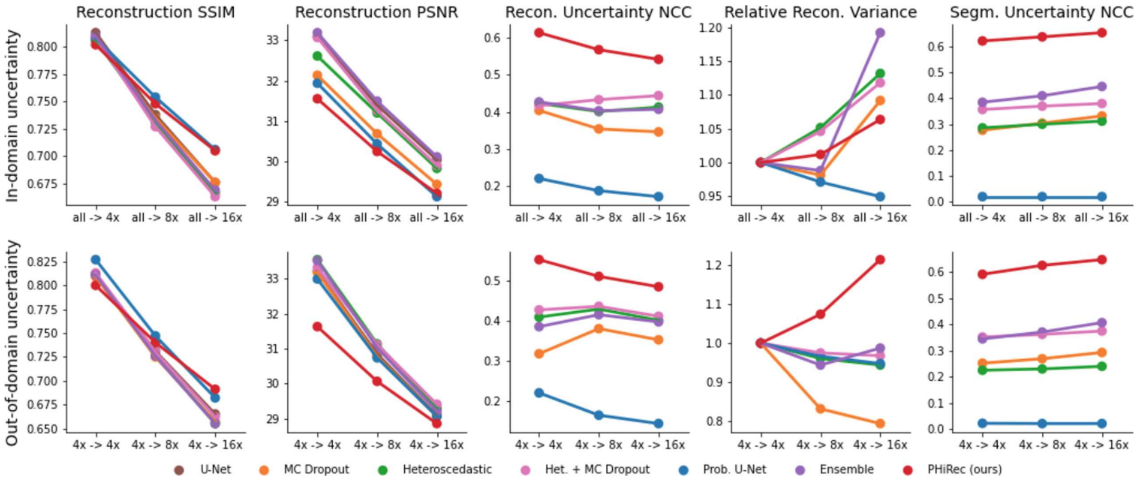

Fig. 2. Quantitative results for the ID (top row), and OOD (bottom row) settings.

Experiment Settings. All experiments were performed in two distinct experimental settings: *in-domain (ID)* and *out-of-domain (OOD)*. In the ID setting, we simultaneously trained and also evaluated on images with acceleration rates 4x, 8x, and 16x. Since all acceleration factors have been seen during training, this setting is dominated by aleotoric uncertainty, that stems from the fact that there are multiple plausible solutions for each undersampled image. In the OOD setting, we trained only on images that have been accelerated 4x, but again tested on images with 4x, 8x, and 16x acceleration. In this setting, for 8x and 16x, there is an additional component of epistemic uncertainty in addition to the aleotoric uncertainty as the testing data moves away from the data that the model has seen during training.

Training Details. All models were implemented in PyTorch [21] and were trained with the Adam optimizer [13] with a learning rate of 10^{-4} using a batch size of 6. The models were trained on NVIDIA RTX 2080 GPUs except the heteroschedastic models and PHiRec which were trained on an NVIDIA Tesla V100 GPU due to increased GPU memory demands. We trained all models for 10 days and model selection was performed based on structural similarity index (SSIM) of the reconstructions on a held-out validation set.

Evaluation of Reconstruction Quality. For both the ID and OOD setting, we evaluated the reconstruction quality in terms of SSIM and peak signal to noise ratio (PSNR). Here, we additionally compared against a standard reconstruction U-Net without uncertainty quantification [10] to ensure the uncertainty quantification does not lead to a general performance degradation. The results are shown in the first two columns of Fig. 2. We observed that, as expected, the performance of all methods degraded with increasing acceleration rates for both settings. This can also be visually confirmed by the squared error maps in Fig. 3 for the OOD setting. Similar effects were observed for the ID setting (see results in supplementary materials). We further observed that all methods performed similarly in terms of pure reconstruction quality. However, PHiRec slightly underperformed in terms of PSNR, but slightly outperformed the other

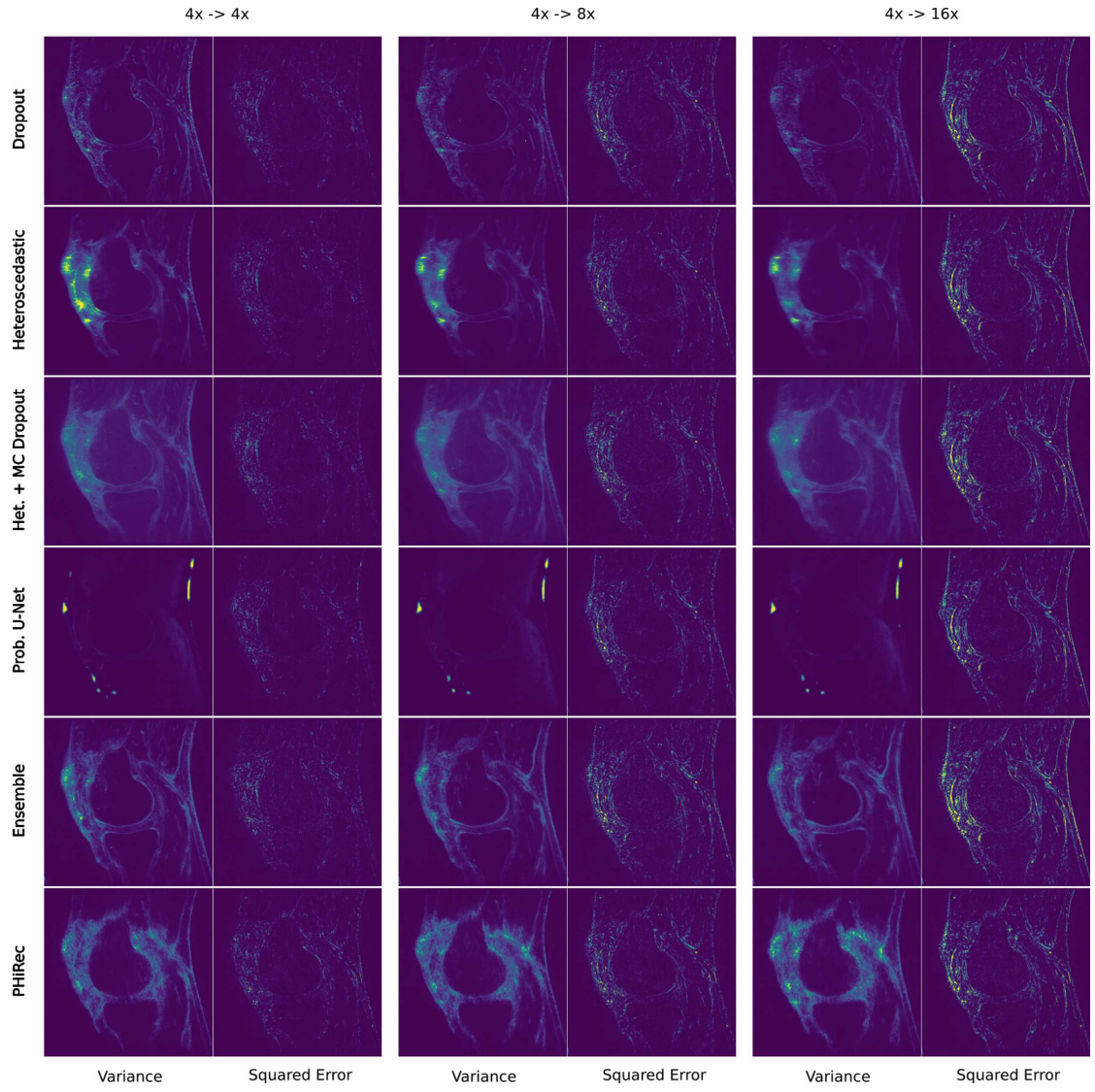

Fig. 3. Variance maps and reconstruction squared error maps in the OOD setting where the column labels have the format "train acceleration" → "test acceleration".

methods in terms of SSIM. This is consistent with the qualitative observation that, while PHiRec reproduced the structural properties exceptionally well, it had a slightly blurry quality. Example reconstructions for all methods are shown in the supplementary materials.

Evaluation of Reconstruction Uncertainty. A crucial quality for a robust uncertainty quantification method is that the model is *calibrated*, i.e. that the uncertainty correlates with the model error [17]. To assess calibration, we computed the average normalised cross correlation (NCC) between the reconstruction uncertainty and the reconstruction squared error for all test images. The results in the third column of Fig. 2 show that PHIRec is substantially better calibrated than the baselines in all settings. It is followed by the MC Dropout + Heteroscedastic Variance approach by Schlemper et al. [25]. The probabilistic U-Net [16], which was originally proposed for segmentation in a multi-annotator

regime, performed the worst in this category due to its poor sample diversity. These results can also be visually confirmed by comparing the uncertainty maps and corresponding error maps in Fig. 3. It is surprising that PHiRec, which is designed to model aleotoric uncertainties, also performed best in the OOD setting. We believe this might be due to the fact that differing acceleration factors do not constitute a large enough domain shift to add significant epistemic uncertainty.

In addition to calibration, we also measured the intuition that the uncertainty should monotonically *increase* with increasing acceleration rates. To this end, we calculated the relative change of the cumulative variance of all image pixels of the 8x and 16x settings with respect to the 4x setting. The results are shown in column four ("Relative Recon. Variance") of Fig. 2. Surprisingly, PHiRec is the only model for which the uncertainty consistently increases for higher acceleration rates. Instead, we found the uncertainty unexpectedly *decreased* for most models between 4x and 8x acceleration. This can be visually confirmed for the OOD setting in the example image shown in Fig. 3.

Evaluation Uncertainty Propagation. Lastly, we investigated propagating the uncertainty to a downstream segmentation task. To this end, we trained a standard deterministic segmentation U-Net with pairs of ground truth images $\mathbf{x}$ and corresponding segmentation masks $\mathbf{s}$ from the SKM-TEA training set. As before we generated a set of 20 reconstruction samples $\{\mathbf{x}_i\}$ for our accelerated test images using all investigated techniques and obtained the corresponding segmentation $\{\mathbf{s}_i\}$ using the segmentation net. We calculated the spread of the segmentation distribution using γ-maps which were defined by Baumgartner et al. [2] as $\gamma(\{\mathbf{s}_i\}) = \mathbb{E}[\mathrm{CE}(\bar{\mathbf{s}}, \mathbf{s}_i)]$, where CE denotes the cross-entropy and $\bar{\mathbf{s}}$ is the mean segmentation. Figure 4 shows pairs of γ-maps and segmentation error maps for an example image in the OOD setting with 16x acceleration. The scaling of the color maps is shared for all images. Qualitatively samples generated by PHIRec exhibited an excellent correlation between the error and the variance outperforming the baselines. We also computed the NCC between the error maps and the segmentation variance (i.e. γ-maps) to measure the calibration of the segmentation uncertainty. The results are shown in column five of Fig. 2. Again, PHiRec clearly outperformed the baseline methods.

4 Discussion

Well-calibrated uncertainty estimation is a crucial component for safely applying DL-based techniques to MRI reconstruction. In this paper, we described PHiRec, a novel reconstruction approach based on hierarchical conditional VAEs, which produces uncertainty estimates substantially better calibrated than several strong baselines. We further demonstrated, how uncertainties originating in the reconstruction process can be propagated to the downstream task of segmentation. In addition to our methodological contributions, we also present the, to our knowledge, first thorough quantitative comparison of different methods for uncertainty quantification in MRI reconstruction.

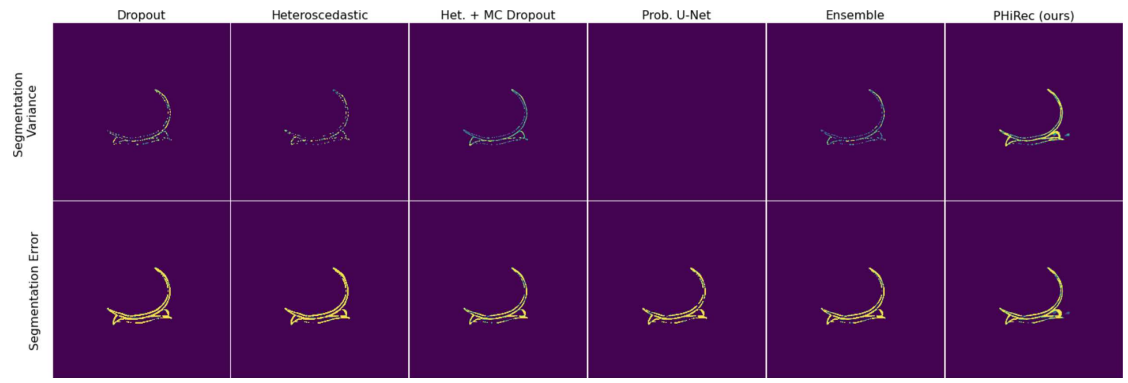

Fig. 4. Segmentation variance maps (measured by γ-maps) and segmentation error maps in the OOD setting with 16x acceleration.

Propagation of uncertainty to downstream tasks may allow to build fail-safe mechanisms to identify when uncertainties are too large to safely make a clinical decision or to guide a treatment. While our study used the simple U-Net as a base architecture and did not enforce data consistency with the measured k-space data, in future work we aim to combine our findings with state-of-the-art methods that are using multiple prediction and data consistency stages (e.g. [14,24,28]). Future work will also focus on investigating the interplay of aleotoric and epistemic uncertainty for larger domain shifts such as changes in anatomy.

Acknowledgments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC number 2064/1 - Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Paul Fischer.

References

1. Angelopoulos, A.N., et al.: Image-to-image regression with distribution-free uncertainty quantification and applications in imaging, February 2022. arXiv arXiv:2202.05265 [cs, eess, q-bio, stat]
2. Baumgartner, C.F., et al.: Phiseg: Capturing uncertainty in medical image segmentation (2019). https://doi.org/10.48550/ARXIV.1906.04045, arXiv:1906.04045
3. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: International Conference on Machine Learning, pp. 1613–1622. PMLR (2015)
4. Calivá, F., et al.: Breaking speed limits with simultaneous ultra-fast MRI reconstruction and tissue segmentation. In: Medical Imaging with Deep Learning, pp. 94–110. PMLR (2020)
5. Chung, H., Ye, J.C.: Score-based diffusion models for accelerated MRI. Med. Image Anal. **80**, 102479 (2022)
6. Desai, A.D., et al.: SKM-TEA: a dataset for accelerated MRI reconstruction with dense image labels for quantitative clinical evaluation (2022)
7. Gottschling, N.M., Antun, V., Adcock, B., Hansen, A.C.: The troublesome kernel: why deep learning for inverse problems is typically unstable. arXiv preprint arXiv:2001.01258 (2020)

8. Hauptmann, A., Arridge, S., Lucka, F., Muthurangu, V., Steeden, J.A.: Real-time cardiovascular MR with spatio-temporal artifact suppression using deep learning-proof of concept in congenital heart disease. Magn. Reson. Med. **81**(2), 1143–1156 (2019)
9. Hepp, T., Gatidis, S., Hammernik, K., Küstner, T.: Uncertainty estimation via ensembling for deep learning-based MR image reconstruction. In: ISMRM, vol. 685 (2022)
10. Hyun, C.M., Kim, H.P., Lee, S.M., Lee, S., Seo, J.K.: Deep learning for undersampled MRI reconstruction. Phys. Med. Biol. **63**(13), 135007 (2018)
11. Jalal, A., Arvinte, M., Daras, G., Price, E., Dimakis, A.G., Tamir, J.I.: Robust compressed sensing MRI with deep generative priors, December 2021. arXiv arXiv:2108.01368 [cs, math, stat]
12. Jin, K.H., McCann, M.T., Froustey, E., Unser, M.: Deep convolutional neural network for inverse problems in imaging. IEEE Trans. Image Process. **26**(9), 4509–4522 (2017). https://doi.org/10.1109/TIP.2017.2713099, http://ieeexplore.ieee.org/document/7949028/
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2017)
14. Kobler, E., Effland, A., Kunisch, K., Pock, T.: Total deep variation for linear inverse problems. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7549–7558 (2020)
15. Kohl, S.A.A., et al.: A hierarchical probabilistic u-net for modeling multi-scale ambiguities (2019)
16. Kohl, S.A.A., et al.: A probabilistic u-net for segmentation of ambiguous images (2018). https://doi.org/10.48550/ARXIV.1806.05034, https://arxiv.org/abs/1806.05034
17. Laves, M.H., Ihler, S., Fast, J.F., Kahrs, L.A., Ortmaier, T.: Recalibration of aleatoric and epistemic regression uncertainty in medical imaging. arXiv preprint arXiv:2104.12376 (2021)
18. Morshuis, J.N., Gatidis, S., Hein, M., Baumgartner, C.F.: Adversarial robustness of MR image reconstruction under realistic perturbations. In: Haq, N., Johnson, P., Maier, A., Qin, C., Würfl, T., Yoo, J. (eds.) Machine Learning for Medical Image Reconstruction, vol. 13587, pp. 24–33. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-17247-2_3
19. Narnhofer, D., Effland, A., Kobler, E., Hammernik, K., Knoll, F., Pock, T.: Bayesian uncertainty estimation of learned variational MRI reconstruction. IEEE Trans. Med. Imaging **41**(2), 279–291 (2022). https://doi.org/10.1109/TMI.2021.3112040
20. Ongie, G., Jalal, A., Metzler, C.A., Baraniuk, R.G., Dimakis, A.G., Willett, R.: Deep learning techniques for inverse problems in imaging, May 2020. arXiv arXiv:2005.06001 [cs, eess, stat]
21. Paszke, A., et al.: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32, pp. 8024–8035. Curran Associates, Inc. (2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
22. Peng, C., Guo, P., Zhou, S.K., Patel, V., Chellappa, R.: Towards performant and reliable undersampled MR reconstruction via diffusion model sampling (2022). https://doi.org/10.48550/ARXIV.2203.04292, arXiv:2203.04292
23. Pruessmann, K.P., Weiger, M., Scheidegger, M.B., Boesiger, P.: SENSE: sensitivity encoding for fast MRI. Magn. Reson. Med. Off. J. Int. Soci. Magn. Reson. Med. **42**(5), 952–962 (1999)

24. Schlemper, J., Caballero, J., Hajnal, J.V., Price, A.N., Rueckert, D.: A deep cascade of convolutional neural networks for dynamic MR image reconstruction. IEEE Trans. Med. Imaging **37**(2), 491–503 (2017)
25. Schlemper, J., et al.: Bayesian deep learning for accelerated MR image reconstruction. In: Knoll, F., Maier, A., Rueckert, D. (eds.) MLMIR 2018. LNCS, vol. 11074, pp. 64–71. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00129-2_8
26. Schlemper, J., et al.: Cardiac MR segmentation from undersampled *k*-space using deep latent representation learning. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 259–267. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_30
27. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
28. Sriram, A., et al.: End-to-end variational networks for accelerated MRI reconstruction. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12262, pp. 64–73. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_7
29. Tezcan, K.C., Karani, N., Baumgartner, C.F., Konukoglu, E.: Sampling possible reconstructions of undersampled acquisitions in MR imaging with a deep learned prior. IEEE Trans. Med. Imaging **41**(7), 1885–1896 (2022)
30. Tolpadi, A.A., et al.: K2S challenge: from undersampled k-space to automatic segmentation. Bioengineering **10**(2), 267 (2023)
31. Waddington, D.E.J., et al.: On real-time image reconstruction with neural networks for MRI-guided radiotherapy, May 2022. arXiv:2202.05267 [physics]
32. Xie, Y., Li, Q.: Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention, vol. 13436, pp. 655–664. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_62
33. Zeng, G., et al.: A review on deep learning MRI reconstruction without fully sampled k-space. BMC Med. Imaging **21**(1), 195 (2021). https://doi.org/10.1186/s12880-021-00727-9
34. Zhang, C., Barbano, R., Jin, B.: Conditional variational autoencoder for learned image reconstruction. Comput. **9**(11), 114 (2021)
35. Zhou, Z., et al.: Parallel imaging and convolutional neural network combined fast MR image reconstruction: applications in low-latency accelerated real-time imaging. Med. Phys. **46**(8), 3399–3413 (2019)

Proceedings of Machine Learning Research 250:1725–1740, 2024 MIDL 2024 – Full paper track

Leveraging Probabilistic Segmentation Models for Improved Glaucoma Diagnosis: A Clinical Pipeline Approach

Anna M. Wundram[*1] ANNA.WUNDRAM@STUDENT.UNI-TUEBINGEN.DE
Paul Fischer[*1] PAUL.FISCHER@UNI-TUEBINGEN.DE
Stephan Wunderlich[1,2] STEPHAN.WUNDERLICH@TUM.DE
Hanna Faber[3,4,5] H.FABER@UKE.DE
Lisa M. Koch[6,7] LISA.KOCH@UNIBE.CH
Philipp Berens[1,6] PHILIPP.BERENS@UNI-TUEBINGEN.DE
Christian F. Baumgartner[1,8] CHRISTIAN.BAUMGARTNER@UNILU.CH

[1] *Cluster of Excellence – Machine Learning for Science, University of Tübingen, Germany*
[2] *Ludwig-Maximilians-University of Munich, Germany*
[3] *University Clinic Hamburg, Germany*
[4] *University Eye Clinic Tübingen, Germany*
[5] *Moorfields Eye Hospital, London, UK*
[6] *Hertie Institute for AI in Brain Health, University of Tübingen, Germany*
[7] *Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism UDEM, Inselspital, Bern University Hospital, University of Bern, Switzerland*
[8] *Faculty of Health Sciences and Medicine, University of Lucerne, Switzerland*

Abstract

The accurate segmentation of the optic cup and disc in fundus images is essential for diagnostic processes such as glaucoma detection. The inherent ambiguity in locating these structures often poses a significant challenge, leading to potential misdiagnosis. To model such ambiguities, numerous probabilistic segmentation models have been proposed. In this paper, we investigate the integration of these probabilistic segmentation models into a multistage pipeline closely resembling clinical practice. Our findings indicate that leveraging the uncertainties provided by these models substantially enhances the quality of glaucoma diagnosis compared to relying on a single segmentation only.

1. Introduction

Glaucoma is a chronic eye disease in which nerve fibers gradually degenerate, leading to damage in the optic nerve head. It is the second leading cause of blindness worldwide affecting one in ten people over the age of eighty (Sevastopolsky, 2017). Diagnosis involves the exact localization of the optic disc and cup in the fundus image. Their shape, size, and relationship to each other are crucial disease markers. Therefore, automated segmentation of the cup and disc is an important step for computer-aided diagnosis frameworks. However, delineation of those structures, in particular the optic cup, is highly challenging and subject to large uncertainties with large disagreements even among experts.

[*] Contributed equally

In this paper, we demonstrate that probabilistic segmentation techniques are effective at capturing uncertainties in the segmentation of the optic disc and cup. Going further, we propose a method to propagate this uncertainty through a multi-stage diagnostic pipeline. Specifically, we propose a method for incorporating the uncertainty arising in an upstream step in the final downstream task, and show that it substantially improves classification performance with respect to a deterministic baseline[1].

Several models have been proposed for optic cup and disc segmentation over the past years (Sevastopolsky et al., 2019; Tulsani et al., 2021; Rasheed et al., 2023). However, these methods do not take into account the inherent uncertainties of this problem. A small number of works address expert disagreements. Edupuganti et al. (2018) incorporate multi-annotator information into learning the segmentation by weighing the loss for pixels depending on the agreement or disagreement of the annotators. Cheng et al. (2020) approximate the joint distribution of the input fundus image and the ground truth segmentation to regularize a U-Net segmentation network. However, estimation of the segmentation uncertainty in optic cup and disc segmentation remains unexplored.

Most recent automated glaucoma diagnosis frameworks approach the problem with black box solutions (Mirzania et al., 2021; Singh et al., 2022; Fan et al., 2023; De Vente et al., 2023). While reaching high performance, end-to-end black box models may often not be clinically desirable. In practice, systems that model the individual steps of a diagnostic pipeline are more interpretable, easier to debug, and more likely to find clinical acceptance. Moreover, data shifts can be addressed by retraining individual components rather than the whole method. Indeed, clinical pipeline approaches have shown great promise in similar areas (De Fauw et al., 2018). A number of works propose to first segment the cup and disc and then extract the vertical or area *cup-to-disc ratio* (CDR) as a diagnostic marker (Al-Bander et al., 2018; Sevastopolsky et al., 2019; Bi et al., 2019; Jiang et al., 2019; Bian et al., 2020; Neto et al., 2022; Zhang et al., 2023). The vertical CDR measures the ratio between the diameter of the cup and the disc along a vertical line through the center. The area CDR measures the ratio of pixels belonging to the cup and disc respectively. However, clinical literature shows that the *rim thickness curve* (RTC) (i.e. the distance between the disc and the cup for every point) may be a more informative measure (Spaeth et al., 2002; Kumar et al., 2019). In this work, we follow these insights and propose a clinically inspired multi-stage pipeline based on initial probabilistic segmentation of the cup and disc, automated RTC extraction, followed by glaucoma classification (see Fig. 1).

Uncertainty estimation is typically studied for individual deep learning models, and has been shown to yield improved performance and robustness on individual tasks (Abdar et al., 2023; Schmidt et al., 2023). However, when clinical workflows consist of tasks that are arranged in a cascading sequence – as is the case here – it becomes crucial to understand how uncertainty in certain stages impacts subsequent tasks. Few studies have focused on uncertainty propagation in clinical pipelines. Eaton-Rosen et al. (2018) propagate tumor segmentation uncertainty to tumor volume measurement using the variance sum law. Mehta et al. (2019, 2021) propose to append variance maps obtained using MC Dropout (Kendall et al., 2016) in a segmentation stage as an additional input channel to a subsequent tumor detection network. To our knowledge, these are the only studies showing that incorporat-

1. The code is available at https://github.com/annawundram/glaucoma-diagnosis-pipeline

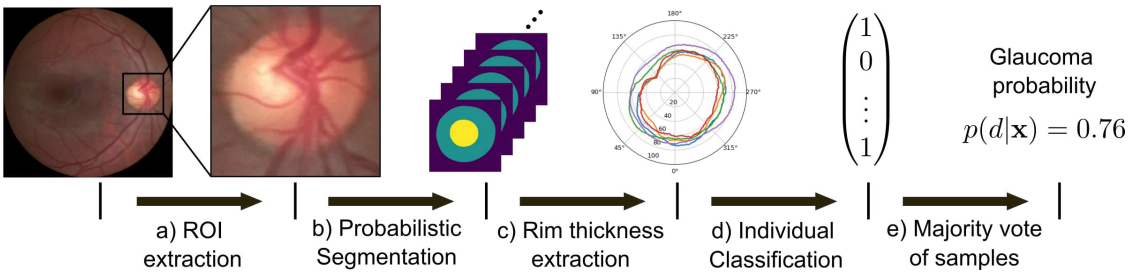

Figure 1: **Proposed pipeline.** a) Automatic ROI extraction from the full fundus images. b) Sampling of possible segmentations using probabilistic segmentation. c) Extraction of rim thickness curves (RTCs) for each segmentation sample. d) Classification of each RTC. e) Marginalization over possible segmentations produces the final glaucoma probability.

ing uncertainty can enhance downstream task performance. However, relying on variance maps instead of the full distribution limits the downstream task to neural networks, as the variance must be added as a channel. Recently, Feiner et al. (2023) and Fischer et al. (2023) proposed two related approaches for propagating uncertainty arising in MR reconstruction to subsequent classification, regression, or segmentation tasks. Both these approaches directly propagate samples from the conditional distribution through the pipeline rather than relying on the variance. Similar, to those works, we propose a sampling-based uncertainty propagation approach. However, we go one step further by using uncertainty to *enhance* downstream performance. This is achieved by marginalizing over possible outcomes in the highly uncertain segmentation stage.

In summary, our contributions are as follows:

1. An interpretable pipeline for glaucoma diagnosis which incorporates clinical knowledge and closely mirrors clinicians' reasoning.
2. The first application of segmentation uncertainty quantification techniques to cup-and-disc segmentation in fundus images with a comparative analysis of four widely used techniques.
3. A sampling-based approach for propagating uncertainty through a multi-stage pipeline as well as a strategy for leveraging this approach to enhance downstream performance.
4. Introduction of RTC as a clinically motivated alternative to CDR in computer-aided glaucoma diagnosis pipelines and empirical verification of its superior performance.

2. Methods

The proposed pipeline consists of five steps as illustrated in Fig. 1. We deterministically extract the region of interest (ROI) containing the cup and disc to obtain the close-up images $\mathbf{x}$. For the highly uncertain step of optic disc and cup segmentation, we use probabilistic segmentation to approximate the distribution $p(\mathbf{s}|\mathbf{x})$ of the segmentations $\mathbf{s}$, and we produce

segmentation samples $\mathbf{s}_i$. The samples (which capture the uncertainty) are propagated through a deterministic rim thickness extraction function $g : \mathbf{s}_i \mapsto \mathbf{r}_i$ producing RTCs $\mathbf{r}_i$. Following this, a deterministic classification function $f : \mathbf{r}_i \mapsto d_i$ takes the RTCs as input and produces a predicted diagnosis $d_i \in \{0, 1\}$ for each $\mathbf{r}_i$. Lastly, we marginalize over the samples to obtain a final probability $p(d|\mathbf{x})$ of the image being "glaucoma suspect". All of these steps will be introduced in detail below.

2.1. Automated Region of Interest Detection

The region of interest (ROI) for glaucoma diagnosis only includes a small area of the fundus containing the optic nerve head, in particular the optic disc and cup (see Fig. 1a).

Prior work has shown that a two-step segmentation approach consisting of ROI extraction, followed by segmentation can improve segmentation results (Kim et al., 2019; Liu et al., 2021). Following these works we use a U-Net for cup and disc segmentation on full-view fundus images. Next, a padded quadratic bounding box is placed around the segmentations, and the resulting ROI is cropped and resized to 320×320 pixels. All experiments in this paper were performed with ROI images $\mathbf{x}$ obtained in this manner.

We use the improved U-Net first employed by Kohl et al. (2018) which operates on 7 rather than 5 resolution levels and uses bilinear upsampling instead of transposed convolutions for all experiments, to ensure consistency with the Prob. U-Net (Kohl et al., 2018) and PHiSeg (Baumgartner et al., 2019) baselines described in the next section.

2.2. Probabilistic Optic Cup and Disc Segmentation

The cup and disc segmentation step is characterized by large uncertainties and variability even among human experts. Therefore, segmentations resulting from a deterministic approach, such as a U-Net, may be insufficient. It may for example produce segmentations with a very thin rim in a subject where a thick and a thin rim are equally likely. The full distribution of possible segmentations $p(\mathbf{s}|\mathbf{x})$ matching a given image $\mathbf{x}$ contains valuable information for the downstream classification task. In recent years, several techniques have been proposed which allow to approximate this conditional probability distribution. In this work, we compare four such techniques: The probabilistic U-Net (Kohl et al., 2018), PHiSeg (Baumgartner et al., 2019), MC Dropout (Kendall et al., 2016), and ensembles (Lakshminarayanan et al., 2017).

The **probabilistic U-Net** (Kohl et al., 2018) is a combination of the conditional VAE (Sohn et al., 2015) approach with a U-Net architecture. **PHiSeg** further extends the idea by a hierarchical latent space and was shown to provide closer approximations of $p(\mathbf{s}|\mathbf{x})$. Both techniques estimate the *aleatoric* uncertainty.

Ensembles are implemented by training ten standard U-Nets with different random seeds (Ensemble_{seeds}). We additionally create an ensemble by training a U-Net for each of the 11 expert annotators in our data ($\text{Ensemble}_{experts}$). The widely used **MC Dropout** technique produces probabilistic segmentation samples by repeatedly predicting segmentations for the same image with dropout enabled. We use a dropout rate of 0.2 on the activation maps for training and testing. Dropout is applied to all layers except the final four segmentation layers. Ensembles and MC Dropout estimate *epistemic* uncertainty. We refer the reader to (Abdar et al., 2021) for definitions of aleatoric and epistemic uncertainty.

We use the improved U-Net architecture first proposed in Kohl et al. (2018) for all approaches with the exception of PHiSeg. PHiSeg employs the same U-Net encoder, but requires a specific decoder. Crucially, all examined probabilistic segmentation methods allow the generation of segmentation samples $\mathbf{s}_i$ from the estimated distribution of $p(\mathbf{s}|\mathbf{x})$.

2.3. Rim Thickness Curve (RTC) Extraction

Next, we extract RTCs $\mathbf{r}_i$ from segmentation samples s_i. Rim thickness is defined as the width of the rim between the borders of the optic cup and disc (Spaeth et al., 2002; Hwang and Kim, 2012). To compute the RTC, a beam centered at the optic cup is rotated by 360 degrees. At every half-degree interval, the points where the beam intersects with the borders of the optic disc and optic cup are determined. The rim thickness is calculated as the Euclidean distance between these two intersections. This process results in a data vector of length 720 for each segmentation sample s_i. The resulting RTCs were visualized as polar plots (see Fig. 1 and 3). We denote the rim-thickness extraction procedure as a deterministic function $g : \mathbf{s}_i \mapsto \mathbf{r}_i$. A visual explanation is shown in Appendix C.

2.4. Glaucoma Classification

The RTCs $\mathbf{r}_i$ are classified as "glaucoma suspect" or "not glaucoma suspect". We use a logistic regression classifier since, in preliminary experiments, more powerful classifiers such as SVMs or Random Forests did not lead to improvements. To prevent overfitting, the RTC data is reduced by grouping the values into 72 bins and calculating their mean. We train an individual classifier for the RTCs obtained from each of the examined segmentation methods. The optimal decision threshold for each classifier is obtained by maximizing the Youden index (sensitivity + specificity - 1) on the validation set. This results in a deterministic classifier $f : \mathbf{r}_i \mapsto d_i$ that maps each rim thickness sample to a binary diagnosis.

2.5. Uncertainty Estimation and Robust Classification

The final probability $p(d|\mathbf{x})$ of an input image $\mathbf{x}$ being "glaucoma suspect" is obtained by marginalizing over all possible segmentations, that is $p(d|\mathbf{x}) = \int p(\mathbf{s}|\mathbf{x}) f(g(\mathbf{s}))d\mathbf{s}$. We approximate this integral using the Monte Carlo method with samples from the respective segmentation techniques. We use 100 samples for all methods, except for ensembles which are limited by the number of networks trained.

The probability $p(d|\mathbf{x})$ is a natural measure for uncertainty as it can be interpreted as a respective agreement or disagreement of the predictions resulting from the segmentation samples. In that sense it is comparable to expert disagreement. We obtain the final robust prediction of our pipeline by thresholding the above probability at 0.5.

3. Experiments and Results

3.1. Data

We used two publicly available fundus image datasets for experiments. The **Chákṣu dataset** (Kumar et al., 2023) contains 1345 fundus images with cup and disc annotations by five experts for each image. Additionally, each expert also provided a diagnosis

Table 1: **Quantitative results.** AUROC, sensitivity and specificity refer to the glaucoma classification task. Dice between the mean predicted segmentation and the consensus ground truth segmentation. The correlation coefficient (CC) between the mean pipeline prediction and the mean expert prediction for all test samples.

		AUROC		Sensitivity		Specificity		Dice	CC
	Segm. source	RTC	CDR	RTC	CDR	RTC	CDR		
Det.	U-Net	0.857	0.8881	0.701	0.771	0.838	0.845	0.919	-
Probabilistic	$\text{Ensemble}_{experts}$	0.863	**0.890**	0.719	0.789	0.802	0.853	0.907	0.629
	Ensemble_{seeds}	**0.899**	0.886	0.824	0.736	0.792	**0.888**	**0.926**	0.639
	Prob. U-Net	0.876	0.869	**0.877**	0.824	0.749	0.799	0.871	0.612
	PHiSeg	0.884	0.882	0.807	**0.842**	0.741	0.752	0.900	**0.653**
	MC Dropout	0.885	0.887	0.736	0.736	**0.835**	0.874	0.894	0.592
	Expert Annotations	0.957	0.930	0.921	0.883	0.884	0.866	-	-
	ResNet50 (black box)	0.884		0.754		0.853		-	-

of "glaucoma suspect" or "not glaucoma suspect". The dataset also contains consensus segmentations obtained using the STAPLE algorithm (Warfield et al., 2004). The **RIGA dataset** (Almazroa et al., 2018) consists of 750 fundus images with cup and disc annotations by six experts for each image. In contrast to Chákṣu it contains no diagnosis labels, and was only used for training the segmentation networks in our study.

We split the training portion of the Chákṣu data into a training and validation set according to an 80/20 split. We used the official test split of the Chákṣu data for all our evaluations. Additionally, we split the RIGA dataset into a training, and a validation portion according to a 80/20 split.

3.2. Training

We trained the segmentation-based stages of our pipelines with combined RIGA and Chákṣu datasets (see Sections 2.1 & 2.2), and the classifiers using only the Chákṣu dataset which contains Glaucoma suspect labels (see Section 2.4). More details about the training and model selection can be found in Appendix A.

3.3. Findings

Uncertainty quantification improves downstream predictions

In order to show that accounting for the uncertainty in the segmentation step leads to improved performance in downstream tasks, we included a deterministic U-Net for the cup and disc segmentation as an additional baseline. Again, we used the improved architecture proposed in Kohl et al. (2018). Moreover, we included a classifier trained directly on the expert annotations, as well as a black box ResNet50 network trained on the ROIs of the Chákṣu dataset to get a sense of the maximum achievable performance.

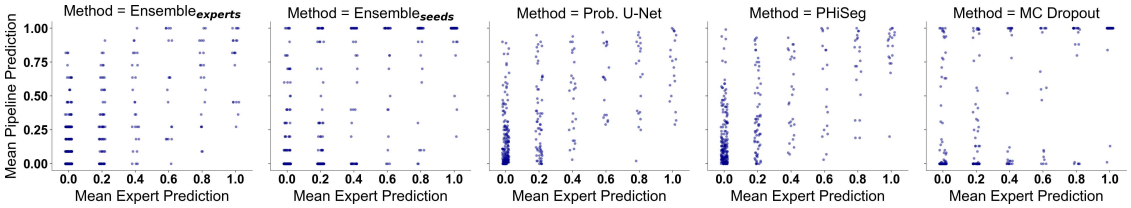

Figure 2: **Mean expert disagreement versus mean pipeline prediction for all methods.** Values close to 1 or 0 indicate large agreement for "glaucoma suspect" or "not glaucoma suspect", respectively. Values close to 0.5 denote high disagreement.

We observed that using probabilistic segmentation techniques consistently led to improvements in the downstream classification performance compared to the deterministic U-Net (Tab. 1), with Ensemble$_{seeds}$ achieving the highest overall AUC score.

Notably, the black box ResNet50 model performed slightly worse than the best probabilistic pipeline approaches. This suggests that our interpretable approach has the potential to outperform black-box models in certain settings. We note that black box approaches may perform better in a data-richer setting (De Vente et al., 2023).

Most methods showed similar performance in the segmentation task, as indicated by the Dice scores in Table 1. However, despite its high Dice score, the deterministic U-Net fell short on AUROC scores, indicating that accurate segmentation alone is insufficient, and that considering the entire probability distribution of the upstream task leads to improvements.

RTCs outperformed CDR for glaucoma diagnosis

In order to confirm the hypothesis that RTCs are better suited for glaucoma diagnosis than the widely used CDR, we additionally extracted the area CDR from all segmentations and trained an additional set of logistic regression classifiers on those values. We observed that the best AUROC scores were achieved with RTC, with particularly large improvements over CDR for the highly performing Ensemble$_{seeds}$.

Propagated uncertainty correlates with expert disagreement

We additionally calculated the Pearson's correlation coefficient (CC) between the mean expert prediction (i.e. all expert predictions averaged) and the mean pipeline prediction $p(d|\mathbf{x})$ in the last column of Tab. 1. PHiSeg achieved the highest CC with Ensemble$_{seeds}$ also performing very well. Visual inspection of the mean expert and pipeline predictions confirmed these findings (see Fig. 2). This indicates that the propagated uncertainty correlates with the expert disagreement, and thus is an informative measure for prediction uncertainty. However, further improvements may be achieved in future work by specifically optimizing downstream calibration.

Qualitative analysis shows good segmentation and RTC agreement with experts

The entropy maps and RTCs in Fig. 3 confirm that the distributions of the annotator disagreement approximately matched the estimated segmentation and RTC uncertainties.

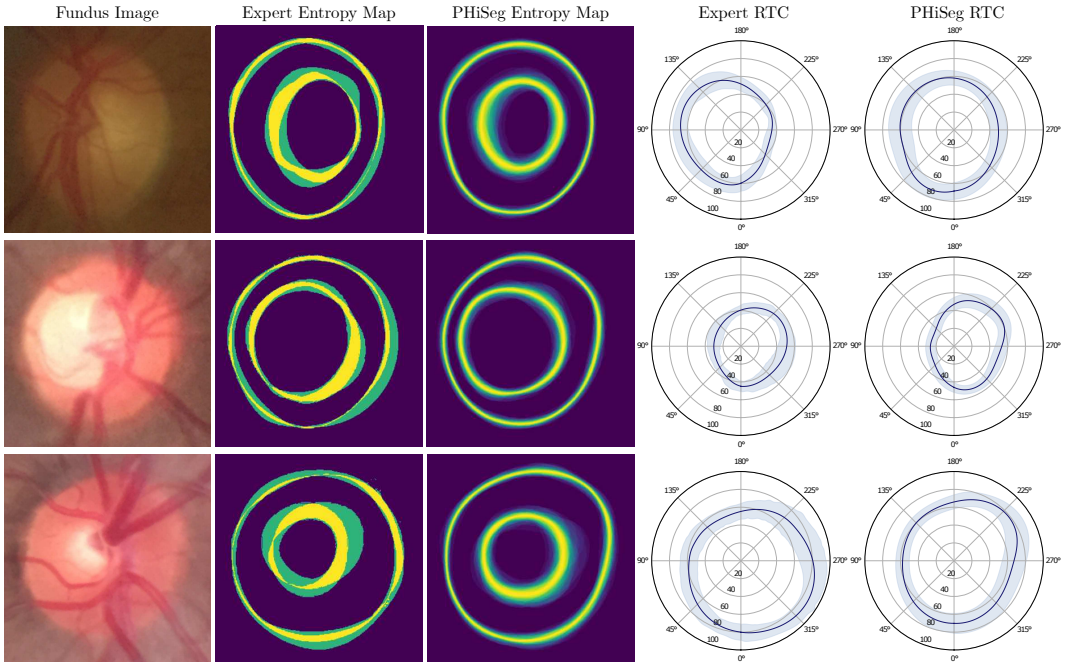

Figure 3: **Entropy maps and RTCs for expert annotations as well as PHiSeg.** The mean rim-thickness (blue line) as well as the standard deviation (light blue shading) are displayed. Three different scenarios are shown: an uncertain case (top), a certain glaucoma suspect (middle) and a certain healthy eye (bottom).

We observed that the method correctly predicted higher uncertainties in areas where the rim was obscured by blood vessels. Additional samples and methods are shown in Appendix B.

4. Discussion and Conclusion

Here, we proposed a pipeline for human-interpretable glaucoma prediction. We showed that probabilistic segmentation techniques are suitable for capturing uncertainties in the location of the cup and disc, and demonstrated an approach for propagating these uncertainties through the pipeline steps to the final prediction. Knowledge about the uncertainty adds an additional level of interpretability to the individual pipeline steps. We furthermore proposed a simple strategy for obtaining robust predictions by marginalizing over the distribution of possible segmentations, and showed that accounting for the uncertainty in this manner led to improved downstream predictions. Our analysis of different probabilistic segmentation techniques revealed that a simple random seed ensemble provided the best balance. However, PHiSeg provided the best qualitative results and correlation with expert disagreements.

A limitation of our work is the sole focus on rim thickness as diagnostic feature. Future work will focus on incorporating diagnostic markers such as the color and intensity of the optic cup, and whether fundus corresponds to the right or left eye into our pipeline. These features are known to be diagnostically important and may further improve performance.

Acknowledgments

This work was supported by the German Science Foundation (BE5601/8-1 and the Excellence Cluster 2064 "Machine Learning — New Perspectives for Science", project number 390727645) and the Hertie Foundation. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Paul Fischer.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

Moloud Abdar, Soorena Salari, Sina Qahremani, Hak-Keung Lam, Fakhri Karray, Sadiq Hussain, Abbas Khosravi, U Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. Uncertaintyfusenet: robust uncertainty-aware hierarchical feature fusion model with ensemble monte carlo dropout for covid-19 detection. *Information Fusion*, 90:364–381, 2023.

Baidaa Al-Bander, Bryan M Williams, Waleed Al-Nuaimy, Majid A Al-Taee, Harry Pratt, and Yalin Zheng. Dense fully convolutional segmentation of the optic disc and cup in colour fundus for glaucoma diagnosis. *Symmetry*, 10(4):87, 2018.

Ahmed Almazroa, Sami Alodhayb, Essameldin Osman, Eslam Ramadan, Mohammed Hummadi, Mohammed Dlaim, Muhannad Alkatee, Kaamran Raahemifar, and Vasudevan Lakshminarayanan. Retinal fundus images for glaucoma analysis: the RIGA dataset. In Jianguo Zhang and Po-Hao Chen, editors, *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, volume 10579, page 105790B. International Society for Optics and Photonics, SPIE, 2018. doi: 10.1117/12.2293584. URL https://doi.org/10.1117/12.2293584.

Christian F. Baumgartner, Kerem C. Tezcan, Krishna Chaitanya, Andreas M. Hötker, Urs J. Muehlematter, Khoschy Schawkat, Anton S. Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 119–127, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32245-8.

Lei Bi, Yuyu Guo, Qian Wang, Dagan Feng, Michael Fulham, and Jinman Kim. Automated segmentation of the optic disk and cup using dual-stage fully convolutional networks, 2019.

Xuesheng Bian, Xiongbiao Luo, Cheng Wang, Weiquan Liu, and Xiuhong Lin. Optic disc and optic cup segmentation based on anatomy guided cascade network. *Computer Methods and Programs in Biomedicine*, 197:105717, 2020.

Pujin Cheng, Junyan Lyu, Yijin Huang, and Xiaoying Tang. Probability distribution guided optic disc and cup segmentation from fundus images. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1976–1979. IEEE, 2020.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

Coen De Vente, Koenraad A Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, et al. Airogs: artificial intelligence for robust glaucoma screening challenge. *IEEE transactions on medical imaging*, 2023.

Zach Eaton-Rosen, Felix Bragman, Sotirios Bisdas, Sébastien Ourselin, and M Jorge Cardoso. Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 691–699. Springer, 2018.

Venkata Gopal Edupuganti, Akshay Chawla, and Amit Kale. Automatic optic disk and cup segmentation of fundus images using deep learning. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 2227–2231. IEEE, 2018.

Rui Fan, Kamran Alipour, Christopher Bowd, Mark Christopher, Nicole Brye, James A. Proudfoot, Michael H. Goldbaum, Akram Belghith, Christopher A. Girkin, Massimo A. Fazio, Jeffrey M. Liebmann, Robert N. Weinreb, Michael Pazzani, David Kriegman, and Linda M. Zangwill. Detecting glaucoma from fundus photographs using deep learning without convolutions: Transformer for improved generalization. *Ophthalmology Science*, 3(1):100233, 2023. ISSN 2666-9145. doi: https://doi.org/10.1016/j.xops.2022.100233. URL https://www.sciencedirect.com/science/article/pii/S2666914522001221.

Leonhard F Feiner, Martin J Menten, Kerstin Hammernik, Paul Hager, Wenqi Huang, Daniel Rueckert, Rickmer F Braren, and Georgios Kaissis. Propagation and attribution of uncertainty in medical imaging pipelines. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 1–11. Springer, 2023.

Paul Fischer, K Thomas, and Christian F Baumgartner. Uncertainty estimation and propagation in accelerated mri reconstruction. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 84–94. Springer, 2023.

Young Hoon Hwang and Yong Yeon Kim. Glaucoma Diagnostic Ability of Quadrant and Clock-Hour Neuroretinal Rim Assessment Using Cirrus HD Optical Coherence Tomography. *Investigative ophthalmology & visual science*, 53(4):2226–2234, 2012.

Yuming Jiang, Lixin Duan, Jun Cheng, Zaiwang Gu, Hu Xia, Huazhu Fu, Changsheng Li, and Jiang Liu. Jointrcnn: a region-based convolutional neural network for optic disc and cup segmentation. *IEEE Transactions on Biomedical Engineering*, 67(2):335–343, 2019.

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding, 2016.

Mijung Kim, Jong Chul Han, Seung Hyup Hyun, Olivier Janssens, Sofie Van Hoecke, Changwon Kee, and Wesley De Neve. Medinoid: Computer-aided diagnosis and localization of glaucoma using deep learning †. *Applied Sciences*, 9(15), 2019. ISSN 2076-3417. doi: 10.3390/app9153064. URL https://www.mdpi.com/2076-3417/9/15/3064.

Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/473447ac58e1cd7e96172575f48dca3b-Paper.pdf.

JR Harish Kumar, Chandra Sekhar Seelamantula, Yogish Subraya Kamath, and Rajani Jampala. Rim-to-disc ratio outperforms cup-to-disc ratio for glaucoma prescreening. *Scientific reports*, 9(1):7099, 2019.

JR Harish Kumar, Chandra Sekhar Seelamantula, JH Gagan, Yogish S Kamath, Neetha IR Kuzhuppilly, U Vivekanand, Preeti Gupta, and Shilpa Patil. Chákṣu: A glaucoma specific fundus image database. *Scientific data*, 10(1):70, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017.

Bingyan Liu, Daru Pan, and Hui Song. Joint optic disc and cup segmentation based on densely connected depthwise separable convolution deep network. *BMC Medical Imaging*, 21(1):14, Jan 2021. ISSN 1471-2342. doi: 10.1186/s12880-020-00528-6. URL https://doi.org/10.1186/s12880-020-00528-6.

Raghav Mehta, Thomas Christinck, Tanya Nair, Paul Lemaitre, Douglas Arnold, and Tal Arbel. Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures: First International Workshop, UNSURE 2019, and 8th International Workshop, CLIP 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 8*, pages 23–32. Springer, 2019.

Raghav Mehta, Thomas Christinck, Tanya Nair, Aurélie Bussy, Swapna Premasiri, Manuela Costantino, M Mallar Chakravarthy, Douglas L Arnold, Yarin Gal, and Tal Arbel. Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference. *IEEE Transactions on Medical Imaging*, 41(2):360–373, 2021.

Delaram Mirzania, Atalie C Thompson, and Kelly W Muir. Applications of deep learning in detection of glaucoma: a systematic review. *European Journal of Ophthalmology*, 31 (4):1618–1642, 2021.

Alexandre Neto, José Camara, and António Cunha. Evaluations of deep learning approaches for glaucoma screening using retinal images from mobile device. *Sensors*, 22(4):1449, 2022.

Haroon Adam Rasheed, Tyler Davis, Esteban Morales, Zhe Fei, Lourdes Grassi, Agustina De Gainza, Kouros Nouri-Mahdavi, and Joseph Caprioli. Rimnet: A deep neural network pipeline for automated identification of the optic disc rim. *Ophthalmology Science*, 3(1): 100244, 2023.

Arne Schmidt, Pablo Morales-Alvarez, and Rafael Molina. Probabilistic attention based on gaussian processes for deep multiple instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

A. Sevastopolsky. Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network. *Pattern Recognition and Image Analysis*, 27(3):618–624, Jul 2017. ISSN 1555-6212. doi: 10.1134/S1054661817030269. URL https://doi.org/10.1134/S1054661817030269.

Artem Sevastopolsky, Stepan Drapak, Konstantin Kiselev, Blake M Snyder, Jeremy D Keenan, and Anastasia Georgievskaya. Stack-u-net: refinement network for improved optic disc and cup image segmentation. In *Medical Imaging 2019: Image Processing*, volume 10949, pages 576–584. SPIE, 2019.

Law Kumar Singh, Pooja, Hitendra Garg, and Munish Khanna. Deep learning system applicability for rapid glaucoma prediction from fundus images across various data sets. *Evolving Systems*, 13(6):807–836, Dec 2022. ISSN 1868-6486. doi: 10.1007/s12530-022-09426-4. URL https://doi.org/10.1007/s12530-022-09426-4.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

George L Spaeth, Jeffrey Henderer, Connie Liu, Muge Kesen, Undraa Altangerel, Atilla Bayer, L Jay Katz, Jonathan Myers, Douglas Rhee, and William Steinmann. The disc damage likelihood scale: reproducibility of a new method of estimating the amount of optic nerve damage caused by glaucoma. *Transactions of the American Ophthalmological Society*, 100:181, 2002.

Akshat Tulsani, Preetham Kumar, and Sumaiya Pathan. Automated segmentation of optic disc and optic cup for glaucoma assessment using improved unet++ architecture. *Biocybernetics and Biomedical Engineering*, 41(2):819–832, 2021.

Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.

Baoliang Zhang, Xiaoxin Guo, Guangyu Li, Zhengran Shen, Xiaoying Hu, and Songtian Che. Multiple graph reasoning network for joint optic disc and cup segmentation. *Applied Intelligence*, pages 1–15, 2023.

Appendix A. Training Details

The ROI extraction U-Net network (see Sec. 2.1) was trained with the combined RIGA and Chákṣu training data sampling different expert ground truths for each batch. Model selection was performed using the Dice score on the validation set.

The probabilistic segmentation networks described in Sec. 2.2 were trained in the same fashion based on the regions extracted in the first step. Model selection for the probabilistic U-Net, PHiSeg, and MC Dropout was performed using the generalized energy distance (GED) metric (Kohl et al., 2018) between the expert annotations and the samples on the validation sets. The ensembles were analogously trained by sampling different experts for each batch, and model selection was performed based on the Dice score of the individual networks.

We trained classifiers with RTCs resulting from each of the segmentation methods as described in Sec. 2.4. The classifiers were trained using the Chákṣu training data, and the optimal threshold was determined using the Chákṣu validation data. We used 100 RTC curve samples per training image for all methods except the $\text{Ensemble}_{experts}$ and Ensemble_{seeds}, which were limited by design to 11 and 10 samples, respectively.

The ResNet50 baseline was initialized using ImageNet weights, and then fine-tuned on predicting the glaucoma label from the automatically extracted ROI crops of the Chákṣu dataset. During training we used random horizontal flips, and small random rotations in $[-10°, 10°]$ to augment the small dataset. Note that we did not use any augmentation for the segmentation networks, as they were additionally trained with RIGA data and segmentation tasks typically require fewer training data points due to the dense segmentation annotations. We used the AUC computed on the validation set for model selection. In order to compute the sensitivity and specificity in Table 1, we obtained the decision threshold which maximizes the Youden index.

Appendix B. Additional qualitative results

In Fig. 4 we show additional qualitative results of the entropy for experts and models as well as the corresponding RTCs.

Appendix C. Visual explanation of RTC extraction

Fig. 5 provides some intuition for the calculation of the RTC.

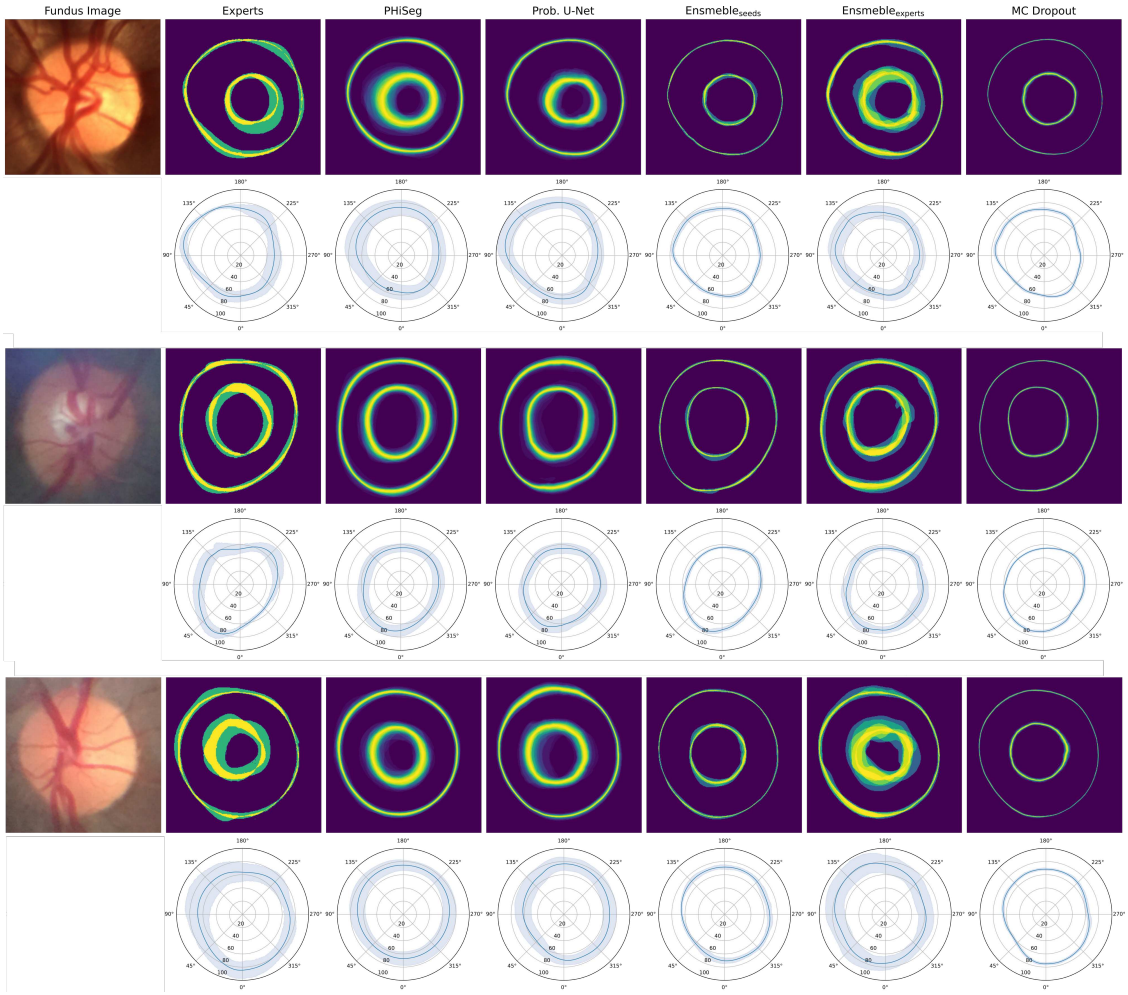

Figure 4: Entropy maps and RTC plots for every model for three representative example subjects.

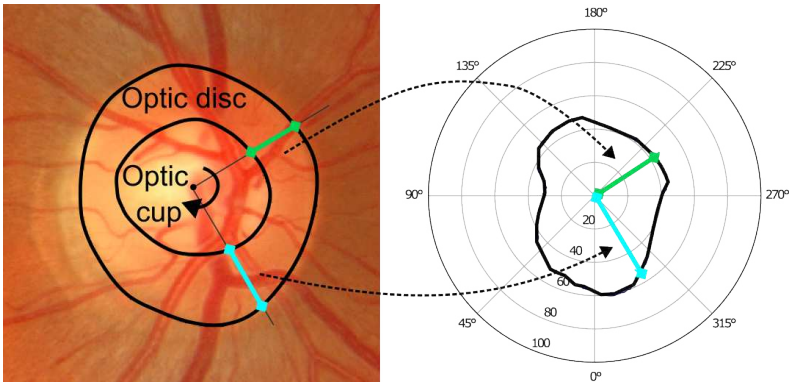

Figure 5: **Rim-thickness calculation.** A beam centered at the cup is rotated 360 degrees. At every 0.5 degrees the Euclidean distance between the cup and the disc is recorded in the rim thickness plot.

Subgroup-Specific Risk-Controlled Dose Estimation in Radiotherapy

Paul Fischer[1,2(✉)], Hannah Willms[1], Moritz Schneider[3], Daniela Thorwarth[1,3], Michael Muehlebach[4], and Christian F. Baumgartner[1,2]

[1] Cluster of Excellence – ML for Science, University of Tübingen, Tübingen, Germany
[2] Faculty of Health Sciences and Medicine, University of Lucerne, Lucerne, Switzerland
paul.fischer@uni-tuebingen.de
[3] Section for Biomedical Physics, Department of Radiation Oncology, University of Tübingen, Tübingen, Germany
[4] Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract. Cancer remains a leading cause of death, highlighting the importance of effective radiotherapy (RT). Magnetic resonance-guided linear accelerators (MR-Linacs) enable imaging during RT, allowing for inter-fraction, and perhaps even intra-fraction, adjustments of treatment plans. However, achieving this requires fast and accurate dose calculations. While Monte Carlo simulations offer accuracy, they are computationally intensive. Deep learning frameworks show promise, yet lack uncertainty quantification crucial for high-risk applications like RT. Risk-controlling prediction sets (RCPS) offer model-agnostic uncertainty quantification with mathematical guarantees. However, we show that naive application of RCPS may lead to only certain subgroups such as the image background being risk-controlled. In this work, we extend RCPS to provide prediction intervals with coverage guarantees for multiple subgroups with unknown subgroup membership at test time. We evaluate our algorithm on real clinical planing volumes from five different anatomical regions and show that our novel subgroup RCPS (SG-RCPS) algorithm leads to prediction intervals that jointly control the risk for multiple subgroups. In particular, our method controls the risk of the crucial voxels along the radiation beam significantly better than conventional RCPS.

1 Introduction

Cancer remains one of the leading causes of death for people under the age of 70. Radiotherapy (RT) has been proven to be a critical treatment modality for var-

P. Fischer and H. Willms—Contributed equally.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-72117-5_65.

M. G. Linguraru et al. (Eds.): MICCAI 2024, LNCS 15010, pp. 696–706, 2024.
https://doi.org/10.1007/978-3-031-72117-5_65

ious tumour entities. Recent advancements in medical imaging have led to the development of the magnetic resonance-guided linear accelerator (MR-Linac), which integrates MR imaging with RT [8]. Aligning the computed tomography (CT) planing volume to an MR scan acquired at the beginning of each treatment fraction, and re-evaluating the treatment plan based on the transformed CT volume allows to better adjust the plan to the patient's inter-fraction changes. This holds the potential for more precise tumor targeting, enhancing treatment outcomes. However, in order to successfully realize these benefits, fast and accurate dose deposition calculations are needed. Current algorithms are based on Monte Carlo simulations, which provide highly accurate results [4]. However, these calculations are computationally intensive, taking minutes to hours. Therefore recent research has focused on optimizing treatment planning efficiency [17]. Deep learning (DL) frameworks have shown encouraging results in dose estimation for RT, with fast calculation time and accurate predictions [13,15,16,18]. However, despite RT's high-risk nature, prior approaches have not adequately addressed risk assessment within DL-based dose estimation.

A natural approach for assessing the risk associated with a prediction is quantifying the prediction uncertainty. Recent years have seen the development of various methods for uncertainty quantification in medical imaging. Examples include deep ensembles [14], Monte Carlo dropout [11], or approaches based on variational autoencoders [3,6,12]. A major limitation of those techniques is that they do not provide any guarantees about the usefulness or correctness of the uncertainty estimates. Recently, risk-controlling prediction sets [2] (RCPS) has gained popularity as a simple, model-agnostic strategy to adapt heuristic notions of uncertainty into uncertainty measures with guarantees. RCPS allows to construct a set of predictions with a guarantee that the correct solution is inside this set with a user-defined probability. Such prediction sets can indicate poor model performance through excessively large intervals, revealing that the models may not be acceptable for certain high-risk applications.

While RCPS has already seen successful adoption in medical image analysis (e.g. [1]), a remaining limitation is that it can only provide guarantees on a global level. There are many situations where we are interested in obtaining guarantees for different subgroups of our population, or different image regions. For instance, in RT, we would like the method to be calibrated along the beam as well as the background. If there is an imbalance between different subgroups (e.g. more background voxels) naive application of RCPS will focus mostly on the majority group (e.g. background) and fail to meet the guarantees for the minority subgroup (e.g. the beam). Calculating RCPS for the different subgroups separately is only a solution if the subgroup is known at test time. However, if this is not the case, as in our RT example, it is not possible to determine the correct RCPS model to use for prediction.

In this paper, we address this problem by proposing a novel calibration algorithm for RCPS that takes into account subgroups and can provide subgroup as well as global guarantees. Our contributions are:

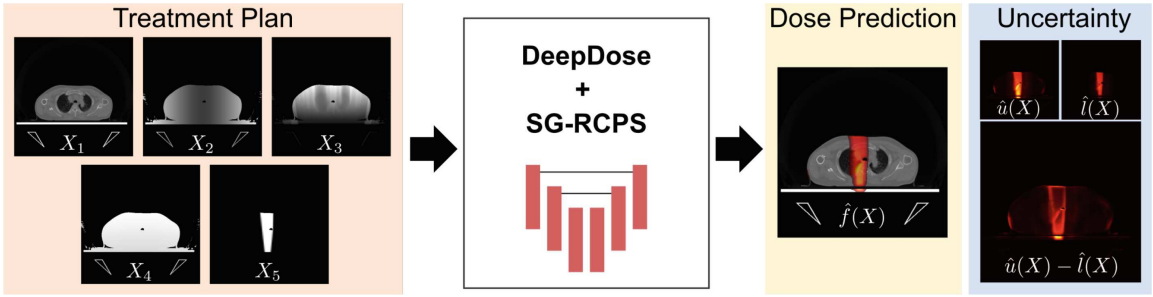

Fig. 1. Overview. We use the DeepDose network [13] to convert a personalized RT plan defined by input CT scan, beam center distance, radiological depth, source distance map, and beam shape (X_1 to X_5) to a voxel-wise dose prediction $\hat{f}(x)$. Extending DeepDose by our novel subgroup risk-controlled prediction sets algorithm (SG-RCPS) allows to obtain a calibrated upper and lower bound for the dose ($\hat{u}(X)$ & $\hat{l}(X)$), as well as the voxel-wise size of the interval ($\hat{u}(X) - \hat{l}(X)$) which serves as final uncertainty measure.

1. The first application of uncertainty quantification in neural network-based dose estimation for RT.
2. A novel algorithm that yields mathematical guarantees for uncertainty intervals for subgroups in the dataset.
3. A quantitative and qualitative evaluation of the algorithm on RT dose prediction on a real-world multi-organ dataset.

2 Methods

The RCPS framework [1,2] ensures that a set-valued predictor $\mathcal{T}$ maintains a *risk* below a user-specified level α with a user-defined probability of $1 - \delta$. In regression problems such as ours, the prediction set is often a prediction interval characterised by a lower and an upper bound value. The risk $\mathcal{R}(\mathcal{T}) = \mathbb{E}[L(Y, \mathcal{T}(X))]$ is defined through a loss function L tailored to the application, which encodes a notion of consequence if the desired property is not fulfilled. In this work, we demonstrate the application of RCPS to DL-based dose prediction and extend the method to ensure risk guarantees for multiple subgroups.

In the following, we will first describe our dose estimation framework (Sect. 2.1). We will then show how heuristic prediction intervals can be obtained using quantile regression (Sect. 2.2). Next, we will discuss how to define the concept of risk, and how the heuristic prediction intervals can be adjusted to control the risk with the desired levels (Sect. 2.3). Lastly, we will describe our novel subgroup RCPS algorithm which allows controlling the risk for multiple subgroups without knowledge of the subgroup membership at test time (Sect. 2.4).

2.1 Dose Estimation Using DeepDose

In order to construct a voxel-wise dose predictor $\hat{f}$ we build on the previously proposed DeepDose network [13,18], which is in turn derived from a 3D UNet

architecture [5]. DeepDose takes the beam shape, center beam line distance, source distance, CT image and radiological depth as input $X \in \mathbb{R}^{5\times W\times H\times D}$ and outputs a predicted dose $\hat{f}(X)_i \in \mathbb{R}$ for each voxel i (see Fig. 1). The network is trained with a highly accurate Monte Carlo simulation as ground truth Y_i for each voxel. For notational clarity, we will omit the index i in the following.

2.2 Heuristic Dose Prediction Intervals Using Quantile Regression

To extend the DeepDose network by the ability to provide prediction intervals, we adopt a voxel-wise quantile regression approach [1]. Specifically, we add two additional output channels $\tilde{l}(X)$ and $\tilde{u}(X)$ to estimate the voxel-wise upper and lower bound, respectively. Similar to [1], we train the two additional network heads using pinball losses, which allow to estimate a specific quantile. For a general feature x, label y, and quantile β, the pinball loss is given by

$$\begin{aligned}\mathcal{L}_\beta(\hat{q}_\beta(x), y) = (y - \hat{q}_\beta(x))\beta \mathbb{1}_{\{y>\hat{q}_\beta(x)\}} + \\ (\hat{q}_\beta(x) - y)(1-\beta)\mathbb{1}_{\{y\leq\hat{q}_\beta(x)\}},\end{aligned} \tag{1}$$

where $\hat{q}_\beta(x)$ is the corresponding quantile estimator and $\mathbb{1}$ denotes the indicator function. We use $\hat{f}(X)-\tilde{l}(X)$ and $\hat{f}(X)+\tilde{u}(X)$ for the respective lower and upper quantile estimators. Our overall training objective $\mathcal{L}$ is comprised of losses for the upper and lower quantiles as well as the standard MSE loss for the point prediction

$$\mathcal{L} = \mathcal{L}_{\alpha/2}(\tilde{l}(X), Y) + \mathcal{L}_{1-\alpha/2}(\tilde{u}(X), Y + \text{MSE}(Y, \hat{f}(X)), \tag{2}$$

where each loss is only applied to the corresponding head. This gives our framework the ability to not only output a per-voxel dose prediction $\hat{f}(X)$, but also a heuristic prediction interval

$$\mathcal{T}(X) = [\hat{f}(X) - \tilde{l}(X), \hat{f}(X) + \tilde{u}(X)]. \tag{3}$$

2.3 RCPS for Radiotherapy Dose Estimation

We now show how the RCPS framework [1,2] can be used to obtain dose prediction intervals that are guaranteed to keep a *risk* below a user-specified level. First, we define the risk of $\mathcal{T}$ as the predicted interval *not* containing the ground truth dose Y

$$\mathcal{R}(\mathcal{T}) = \mathbb{E}\left[\mathbb{1}_{\{Y\notin\mathcal{T}(X)\}}\right] = \Pr(Y \notin \mathcal{T}(X)). \tag{4}$$

We then define new lower and upper bounds by scaling them with a non-negative factor $\hat{\lambda}$

$$\hat{l}(X) = \hat{\lambda}\tilde{l}(X) \quad \text{and} \quad \hat{u}(X) = \hat{\lambda}\tilde{u}(X). \tag{5}$$

RCPS provides a strategy to choose $\hat{\lambda}$ based on a calibration dataset such that $\mathcal{R}(\mathcal{T}) \leq \alpha$ with a probability of at least $1-\delta$ on future test data under the assumption of exchangeability of the test and calibration sets. We use $\alpha =$

$\delta = 0.1$ for all experiments, meaning that with a probability of at least 90%, a minimum of 90% of the ground truth dose depositions should be contained in the predicted dose interval.

Since the calibration data is only a random sample of the data distribution, $\hat{\lambda}$ cannot be chosen by simply minimising the risk on the calibration set. Rather a point-wise *upper confidence bound (UCB)* $\mathcal{R}^+ : \Pr(\mathcal{R}(\lambda)) \leq \mathcal{R}^+$ must be obtained. This UCB accounts for the calibration sample size and the desired probability of the guarantee holding on future data. Following [1] we use the Hoeffding bound [9] to define $\mathcal{R}^+$ as

$$R^+(\lambda) = \frac{1}{nWHD} \sum_{k=1}^{n} \#\{\mathcal{T}_\lambda(X_k) \notin Y_k\} + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}} \,. \tag{6}$$

In the original RCPS approach, Bates et al. [2] proposed a greedy optimization algorithm for obtaining the smallest possible prediction interval still fulfilling the desired guarantees. It requires initializing $\hat{\lambda}$ with a very large value and reducing it until $\mathcal{R}^+$ falls under the desired risk level, that is,

$$\hat{\lambda} = \min \left\{ \lambda : \hat{R}^+(\lambda) \leq \alpha \right\}. \tag{7}$$

This procedure produces a dose interval predictor

$$\mathcal{T}_{\hat{\lambda}}(X) = [\hat{f}(X) - \hat{\lambda}\tilde{l}(X), \hat{f}(X) + \hat{\lambda}\tilde{u}(X)], \tag{8}$$

for the predicted dose that satisfies the desired risk-properties *on average* for all voxels of the test data distribution. However, it does not offer conditional guarantees for individual data subgroups. For instance, as we will show in Sect. 3 a naive application of RCPS results in the voxels along the radiation beam violating the desired guarantees.

2.4 Risk-Controlled Prediction Sets with Multiple Subgroups

If our dataset comprises M imbalanced subgroups, using Eq. 7 to obtain $\hat{\lambda}$ provides guarantees for the overall dataset but not for each individual subgroup. This can lead to a systematic miscalibration of uncertainty intervals for underrepresented subgroups. If the subgroups are known at test time, this problem can be addressed by calibrating for each subgroup separately, by using a subgroup-specific parameter $\hat{\lambda}_z$ during test time. In our RT application, we are interested in calibrated uncertainty quantification in the area of the beam as well as the background, which also receives small dose intensities. While we know the absorbed dose during training, we do not have direct access to this information during testing. Hence, naive application of RCPS yields good calibration overall, but not in the critical area of the beam.

Therefore, we propose an extension of the risk-controlling framework for scenarios where the subgroup membership Z is unknown at test time. Our proposed extension provides the same guarantees as RCPS for each individual subgroup.

That is for every subgroup Z in our dataset, it holds that, at a risk level of α, the ground truth is included with a probability of at least $1 - \delta$.

In order to achieve the desired guarantees we reformulate the risk conditioned on the subgroups Z as follows

$$\mathcal{R}(\mathcal{T}) = \Pr(Y \notin \mathcal{T}(X)) = \mathbb{E}_Z[\mathbb{E}_{XY|Z}\left[\mathbb{1}_{\{Y \notin \mathcal{T}(X)\}}\right]] \,. \tag{9}$$

This leads to a novel subgroup risk-controlled predictions set (SG-RCPS) procedure which is summarised in Algorithm 1. Similar to the original RCPS algorithm, we start off with a large $\hat{\lambda}$ that satisfies the risk for each subgroup. We then iteratively reduce the interval size until the first confidence bound of a subgroup no longer satisfies the criterion[1]. A proof that Algorithm 1 leads to the desired subgroup guarantees is presented in Appendix A.

Algorithm 1: Pseudocode for SG-RSPC

Input : Calibration sets $(X_k, Y_k)_z$, $k = 1, \ldots, n_z$ where $z = 1, \ldots, M$ indicates the subgroup; in our case we have $M = 3$ for foreground, background and the combined image; risk level α; error rate δ; predictor $\hat{f}$; heuristic lower and upper interval predictions $\tilde{l}$ and $\tilde{u}$; initial max value $\lambda_{\max}$; step size $d\lambda > 0$

Output: Optimal interval scaling $\hat{\lambda}$

$\lambda \leftarrow \lambda_{\max}$
for $z \leftarrow 1$ **to** M **do**
  $UCB_z \leftarrow 1$
end
while $UCB_1 \leq \alpha \,\&\, \ldots \,\&\, UCB_M \leq \alpha$ **do**
  $\lambda \leftarrow \lambda - d\lambda$
  for $z \leftarrow 1$ **to** M **do**
    for $k \leftarrow 0$ **to** n_z **do**
      $L_{k,z} \leftarrow \#\{\mathcal{T}_\lambda(X_{k,z}) \notin Y_{k,z}\}/WHD$
    end
    $UCB_z \leftarrow \frac{1}{n_z}\sum_{k=1}^{n_z} L_{k,z} + \sqrt{\frac{1}{2n_z}\log\frac{1}{\delta}}$
  end
end
$\hat{\lambda} \leftarrow \lambda + d\lambda$

3 Experiments and Results

3.1 Dataset

To assess the performance of our model, we trained and tested it on a dataset containing CT data and RT treatment plans of 125 patients obtained from patients at the Department of Radiation Oncology at the University of Tübingen. The

[1] The code is available at https://github.com/paulkogni/SG-RCPS.

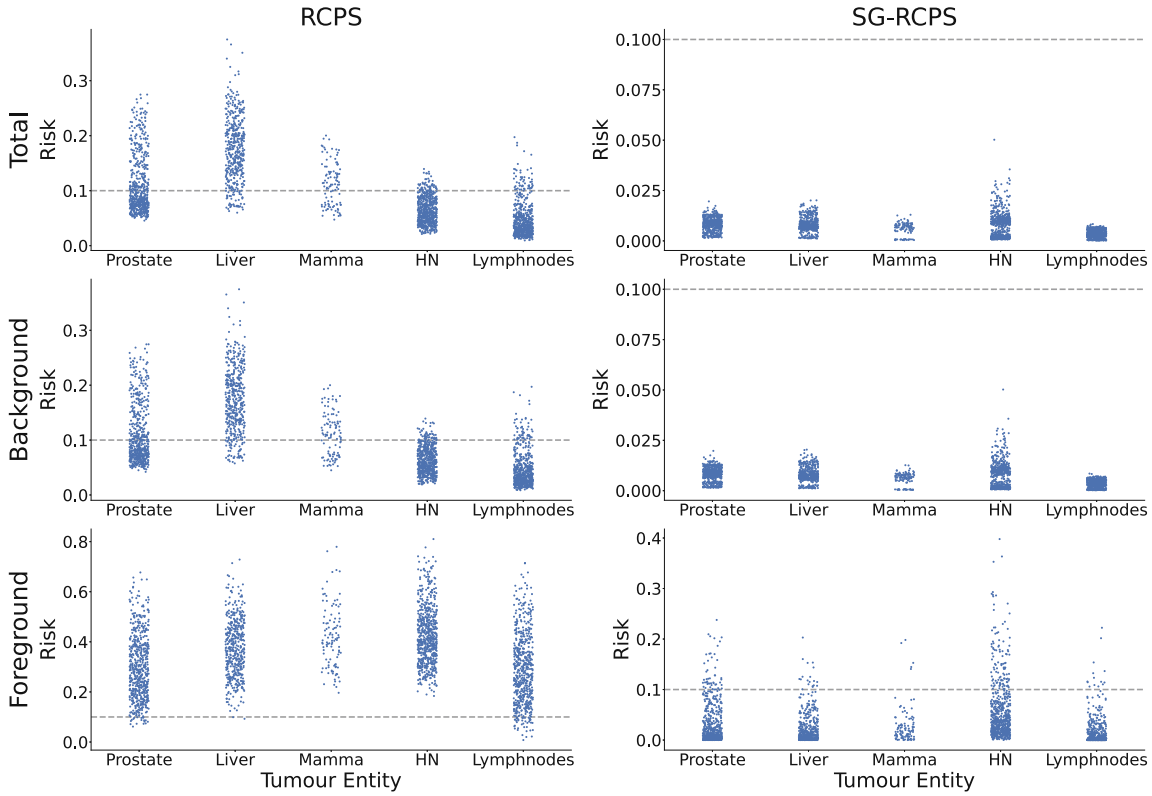

Fig. 2. Tumor-specific risks for the original calibration method (left) and our method (right) for the total image (top row), the background radiation (middle row) and foreground radiation (bottom row).

study was approved by the institutional review board and all patients gave written informed consent (NCT04172753).

The training dataset comprises four anatomical entities: prostate, liver, breast (mamma), and head and neck (HN). The test and calibration datasets contained data from the same four tumour entities as the training dataset, along with lymph nodes, as an additional out-of-domain (OOD) entity. Testing the neural network on an OOD entity allowed us to assess whether the calibration is able to generalize, which is highly desirable in real-world scenarios. For each patient, we extracted multi-leaf collimator (MLC) segments, resulting in a total of 6638 segments. For validation, we randomly selected 20 diverse prostate segments from the dataset. For calibration, we randomly selected three random segments from one patient for each entity. More detailed information about the data splits is provided in Appendix B. To train the DeepDose network, each segment was divided into patches of size $5 \times 32 \times 32 \times 32$. The ground truth dose estimations were generated using Monte Carlo simulations with the EGSnrc open-source software package [7,10]. For inference, we performed predictions at the patch level and combined them in a sliding window fashion.

Table 1. Quantitative results. Empirical risks for RCPS and SG-RCPS averaged over all segments. Controlled risks with $\alpha \leq 0.1$ for more than $1 - \delta$ of the cases are highlighted in bold.

Method	Prostate		Liver		Mamma		HN		Lymph.	
	RCPS	SG-RCPS	RCPS	SG-RCPS	RCPS	SG-RCPS	RCPS	SG-RCPS	RCPS	SG-RCPS
Total image	0.415	**0.0**	0.898	**0.0**	0.620	**0.0**	**0.089**	**0.0**	**0.094**	**0.0**
Backgr. rad.	0.403	**0.0**	0.896	**0.0**	0.610	**0.0**	**0.083**	**0.0**	**0.094**	**0.0**
Foregr. rad.	0.970	**0.084**	0.996	**0.041**	1.0	**0.048**	1.0	0.176	0.934	**0.024**

3.2 Findings

We trained a DeepDose network using training data from all four entities. To evaluate the model's dose estimation performance we calculated the 3 mm/3% gamma pass rate (γ-PR) criterion and observed a γ-PR of 98.9%.

We then compared the two calibration strategies discussed above: RCPS, and our proposed subgroup RCPS (SG-RCPS). We considered three subgroups: the beam foreground and background determined by thresholding the ground truth dose (see Fig. 1), and the whole image. We used a target risk level of $\alpha = 0.1$, and an error rate of $\delta = 0.1$.

Figure 2 and Table 1 show the empirical risk for all segments in the test set grouped by entity. The empirical risk for each segment is defined as proportion of ground truth doses not contained in the predicted interval. Based on our risk settings we expect at most 10% of the segments to fall above the specified risk of 10%.

From Fig. 2 it can be seen that the calibration is dominated by the background class. We found that the normal RCPS algorithm only controlled for the risk in the head & neck, and lymph node entities when considering the total image, and the background only. However, the risk was not controlled to the desired levels in the foreground subgroup (i.e. the beam) for any of the entities. Interestingly, the risks for liver, prostate as well as mamma are not controlled even in the total image. This is likely caused by a mismatch in the proportion of background voxels in the calibration and the test set.

Our proposed SG-RCPS algorithm was able control the risk substantially better for all anatomical areas. There were no dose predictions outside the predicted interval when considering the total image and the background only. We note that because our algorithm estimates a single $\hat{\lambda}$ that controls the risk for all subgroups jointly, the estimates for the total image and background group were more conservative. When considering the foreground (i.e. the beam) only, we found that all entities except head & neck were risk-controlled with the desired levels. As can be seen in Table 1 the empirical risk for the head & neck entity fell slightly short of the desired levels. Notably, the risk for out-of-distribution entity, lymph nodes, was also well controlled. Predicting risk-controlled intervals is particularly important for the foreground, as the beam is where the most accurate uncertainty estimation is required.

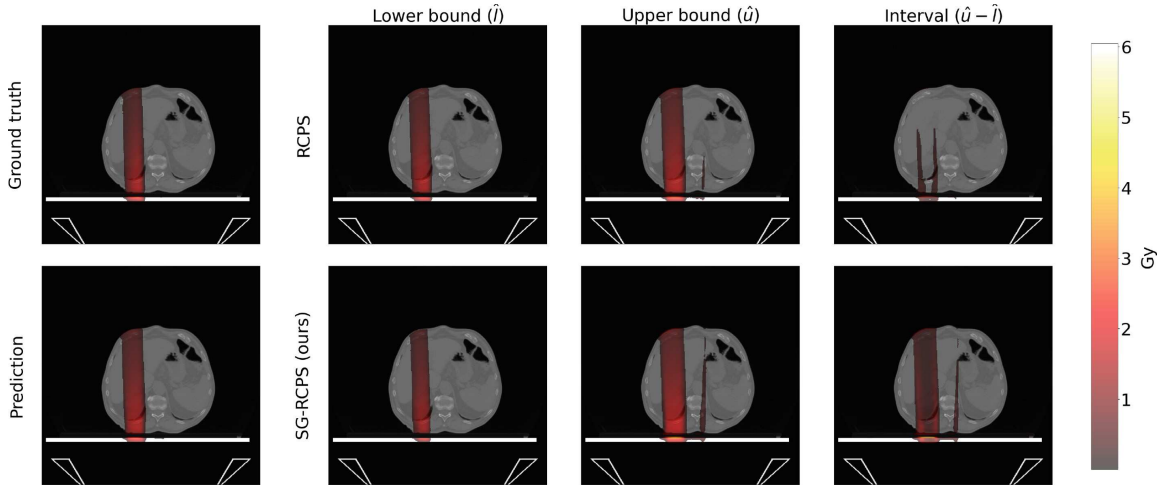

Fig. 3. A representative example for a liver tumor visualizing the qualitative differences between the uncertainty intervals for the non-subgroup-specific calibration and our method. The uncertainty intervals provided by our method are significantly wider ($p < 0.001$) than the ones generated by classical RCPS. All values are given in Gray (Gy). (Color figure online)

A qualitative example of dose predictions and prediction intervals for RCPS and SG-RCPS is shown in Fig. 3.

4 Conclusion

We have proposed subgroup RCPS, an extension of the RCPS algorithm allowing to control risk for multiple subgroups with unknown subgroup membership at test time. We validated our method on a clinical RT dataset comprising five anatomical entities. Our results demonstrate that in case of imbalances between subgroups our method substantially improves calibration for individual subsets. Specifically, in contrast to regular RCPS, our SG-RCPS approach allows to control the risk for the beam *and* the background thereby increasing safety and trustworthiness in this high-risk application. A potential drawback of our method is that it requires a separate calibration set for each subgroup. Additionally, this method usually yields more conservative prediction intervals. In the future, we will apply this algorithm to datasets that include other under-represented subgroups, such as ethnicity or gender.

Acknowledgments. This work was supported by the Excellence Cluster 2064 "Machine Learning—New Perspectives for Science", project number 390727645 and received funding from the German Research Council under DFG Grant No. ZI 736/2-1. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Paul Fischer. The authors acknowledge support by the state of Baden-Württemberg through bwHPC (INST39/963-1 FUGG bwForCluster NEMO) and through the Research and Training Network "AI4MedBW".

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Angelopoulos, A.N., et al.: Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In: International Conference on Machine Learning, pp. 717–730. PMLR (2022)
2. Bates, S., Angelopoulos, A., Lei, L., Malik, J., Jordan, M.: Distribution-free, risk-controlling prediction sets. J. ACM (JACM) **68**(6), 1–34 (2021)
3. Baumgartner, C.F., et al.: PHiSeg: capturing uncertainty in medical image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 119–127. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_14
4. Bol, G., Hissoiny, S., Lagendijk, J., Raaymakers, B.: Fast online monte carlo-based imrt planning for the mri linear accelerator. Phys. Med. Biol. **57**(5), 1375 (2012)
5. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
6. Fischer, P., Thomas, K., Baumgartner, C.F.: Uncertainty estimation and propagation in accelerated mri reconstruction. In: Sudre, C.H., Baumgartner, C.F., Dalca, A., Mehta, R., Qin, C., Wells, W.M. (eds.) UNSURE 2023. LNCS, pp. 84–94. Springer, Heidelberg (2023). https://doi.org/10.1007/978-3-031-44336-7_9
7. Friedel, M., Nachbar, M., Mönnich, D., Dohm, O., Thorwarth, D.: Development and validation of a 1.5 t mr-linac full accelerator head and cryostat model for monte carlo dose simulations. Med. Phys. **46**(11), 5304–5313 (2019)
8. Hall, W., et al.: Mr linac atlantic consortium and the viewray c2t2 research consortium. the transformation of radiation oncology using real-time magnetic resonance guidance: a review. Eur. J. Cancer **122**, 42–52 (2019)
9. Hoeffding, W.: Probability inequalities for sums of bounded random variables. In: The Collected Works of Wassily Hoeffding, pp. 409–426 (1994)
10. Kawrakow, I., Rogers, D., Mainegra-Hing, E., Tessier, F., Townson, R., Walters, B.: Egsnrc toolkit for monte carlo simulation of ionizing radiation transport (2000). https://doi.org/10.4224/40001303
11. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint arXiv:1511.02680 (2015)
12. Kohl, S., et al.: A probabilistic u-net for segmentation of ambiguous images. Adv. Neural Inf. Process. Syst. **31** (2018)
13. Kontaxis, C., Bol, G., Lagendijk, J., Raaymakers, B.: Deepdose: towards a fast dose calculation engine for radiation therapy using deep learning. Phys. Med. Biol. **65**(7), 075013 (2020)
14. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Adv. Neural Inf. Process. Syst. **30** (2017)
15. Martinot, S., Bus, N., Vakalopoulou, M., Robert, C., Deutsch, E., Paragios, N.: High-particle simulation of Monte-Carlo dose distribution with 3D ConvLSTMs. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 499–508. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_48

16. Neishabouri, A., Wahl, N., Mairani, A., Köthe, U., Bangert, M.: Long short-term memory networks for proton dose calculation in highly heterogeneous tissues. Med. Phys. **48**(4), 1893–1908 (2021)
17. Randall, J.W., Rammohan, N., Das, I.J., Yadav, P.: Towards accurate and precise image-guided radiotherapy: clinical applications of the mr-linac. J. Clin. Med. **11**(14), 4044 (2022)
18. Tsekas, G., Bol, G., Raaymakers, B., Kontaxis, C.: Deepdose: a robust deep learning-based dose engine for abdominal tumours in a 1.5 t mri radiotherapy system. Phys. Med. Biol. **66**(6), 065017 (2021)

Manuscripts Under Review

B.

CUTE-MRI: Conformalized Uncertainty-based framework for Time-adaptivE MRI

Paul Fischer[a,b], Jan Nikolas Morshuis[b], Thomas Küstner[c,b], Christian Baumgartner[a,b]

[a] *University of Lucerne, Faculty of Health Sciences and Medicine, Lucerne, Switzerland*
[b] *University of Tübingen, Cluster of Excellence – Machine Learning for Science, Tübingen, Germany*
[c] *University Hospital of Tübingen, Medical Image and Data Analysis Lab, Tübingen, Germany*

Abstract

Magnetic Resonance Imaging (MRI) offers unparalleled soft-tissue contrast but is fundamentally limited by long acquisition times. While deep learning-based accelerated MRI can dramatically shorten scan times, the reconstruction from undersampled data introduces ambiguity resulting from an ill-posed problem with infinitely many possible solutions that propagates to downstream clinical tasks. This uncertainty is usually ignored during the acquisition process as acceleration factors are often fixed *a priori*, resulting in scans that are either unnecessarily long or of insufficient quality for a given clinical endpoint. This work introduces a dynamic, uncertainty-aware acquisition framework that adjusts scan time on a per-subject basis. Our method leverages a probabilistic reconstruction model to estimate image uncertainty, which is then propagated through a full analysis pipeline to a quantitative metric of interest (e.g., patellar cartilage volume or cardiac ejection fraction). We use conformal prediction to transform this uncertainty into a rigorous, calibrated confidence interval for the metric. During acquisition, the system iteratively samples k-space, updates the reconstruction, and evaluates the confidence interval. The scan terminates automatically once the uncertainty meets a user-predefined precision target. We validate our framework on both knee and cardiac MRI datasets. Our results demonstrate that this adaptive approach reduces scan times compared to fixed protocols while providing formal statistical guarantees on the precision of the final image. This framework moves beyond fixed acceleration factors, enabling patient-specific acquisitions that balance scan efficiency with diagnostic confidence, a critical

step towards personalized and resource-efficient MRI.

Keywords:

---

1. Introduction

Magnetic Resonance Imaging (MRI) is a cornerstone of modern medical diagnostics. Its ability to non-invasively generate images with exceptional soft-tissue contrast makes it indispensable for the diagnosis, staging, and monitoring of a wide range of diseases, from neurological disorders to musculoskeletal injuries and cardiovascular conditions [1]. However, the high diagnostic value of MRI is often counterbalanced by its inherently long acquisition times. These lengthy scans can lead to patient discomfort, increase the likelihood of motion artifacts that degrade image quality, and limit patient throughput, thereby increasing operational costs and wait times [2]. Consequently, accelerated MRI techniques, which aim to reconstruct high-quality images from undersampled k-space data, are of paramount importance for making MRI more efficient, cost-effective, and patient-friendly [3, 4].

While accelerated MRI promises to alleviate these challenges, the majority of current methods, both in clinical practice and in research, rely on static acquisition strategies [3, 4, 5]. These approaches employ fixed, pre-determined undersampling rates that are designed offline and are not adapted to the specific patient. This inflexibility represents a central, unaddressed limitation: the acquisition process remains agnostic to the content and complexity of the image being formed. This can lead to a suboptimal use of scanner time, as less data may be sufficient, especially when a specific downstream metric is the primary interest.

The evolution of accelerated MRI has been marked by two major paradigms. The first encompasses classic reconstruction techniques, such as parallel imaging and compressed sensing, while the second is defined by the rise of deep learning (DL). Classic methods, rooted in parallel imaging (e.g., SENSE, GRAPPA) and compressed sensing (CS), leverage explicit priors like signal sparsity to recover images from limited data [3, 6, 7]. While they provide a strong theoretical foundation, their performance tends to degrade at high acceleration factors, where severe aliasing artifacts can become diagnostically prohibitive. In contrast, the second paradigm of deep learning has revolutionized the field. Models trained on large datasets learn complex, implicit priors and have demonstrated high-quality reconstructions, even from highly

undersampled data [8, 9, 10, 11]. These methods often outperform traditional techniques in terms of pure reconstruction quality and speed.

Despite their impressive performance, deep learning models often function as "black boxes," and their predictions come with no inherent guarantees of correctness. This can lead to a critical problem of misplaced trust, where models may produce plausible-looking but factually incorrect reconstructions, a phenomenon often termed "hallucination" [12, 13, 14]. The risk is particularly acute in the ill-posed problem of MR reconstruction, where uncertainty arises not only from the missing k-space measurements but also from physiological and anatomical variability between patients, pathologies, and motion. This unquantified uncertainty does not just affect the reconstructed image; it can silently propagate to and corrupt downstream clinical tasks, such as segmentation, registration, or disease classification, that rely on these images for diagnosis and treatment planning [15, 16].

Recognizing this challenge, a growing body of research has focused on uncertainty quantification (UQ) for deep learning in medical imaging. Various methods, such as Bayesian neural networks, ensembles and variational autoencoder-based methods have been developed to estimate model uncertainty [17, 18, 19, 20, 21, 22, 23, 24, 25]. Several works have successfully demonstrated how this uncertainty can be propagated from the reconstruction to a downstream task to provide a more complete picture of diagnostic confidence [15, 26].

However, while significant research has focused on estimating and propagating uncertainty for post-hoc analysis, its potential to actively guide and optimize the MRI acquisition process itself in real-time remains largely unexplored. Daudé et al. [27] proposed an adaptive method where scan quality, specifically the Signal-to-Noise Ratio SNR, is estimated periodically during acquisition. The scan is terminated once the SNR surpasses a pre-defined quality threshold, enabling personalized scan durations. However, this approach relies on a classical, signal-based metric and does not account for the reconstruction uncertainty or potential for artifacts, such as hallucinations, common in modern learning-based methods. Pineda et al. [28] for example analyzed how to find the optimal sampling trajectory for accelerated MR acquisition using reinforcement learning, however they did not consider the effect of downstream applications. Wang et al. [29] analyzed jointly the influence of k-space acquisition and segmentation quality by iteratively sampling k-space up to a fixed undersampling rate such that segmentation quality based on reconstructions within this pipeline is as high as possible. However,

this work does not incorporate the intrinsic uncertainty within this pipeline as well as evaluating when there is "enough" k-space data. It becomes apparent that prior work on optimizing scan duration has typically focused on pre-calculating sampling trajectories or defining stopping criteria based on image-level metrics, without considering model confidence along the diagnostic pipeline [30, 31]. This reveals a critical gap: current static acquisition protocols are inherently inefficient. They may waste valuable scanner time on anatomically "easy" cases that could have been reconstructed with sufficient quality from fewer measurements, or conversely, they may terminate prematurely for "hard" or unusual cases, yielding diagnostically inadequate images. This one-size-fits-all approach fails to account for the simple fact that some diagnostic tasks or anatomies do not require perfectly reconstructed images to yield clinically reliable results.

In this work, we hypothesize that by monitoring the uncertainty of a reconstruction model and its downstream clinical application, one can create a patient-specific, adaptive stopping rule for k-space acquisition. The core idea is to halt the scan precisely when the system reaches a pre-defined level of diagnostic confidence, rather than adhering to a fixed sampling budget. Such a dynamic stopping criterion would optimize the scan duration for each individual, allowing for fast scan times while keeping the diagnostic quality high. This would not only improve patient comfort and scanner throughput but would do so without sacrificing the diagnostic integrity required for clinical decision-making.

To address this gap, we introduce CUTE-MRI: a Conformalized Uncertainty-based framework for Time-adaptivE MRI. This novel framework leverages uncertainty estimation to determine an optimal, patient-specific stopping point for the scan, ensuring that the resulting images are fit for a specified clinical purpose. Our main contributions are threefold:

1. We propose a complete framework for dynamically terminating an MR acquisition based on the propagation of uncertainty through a diagnostic pipeline, from reconstruction to a downstream clinical measurement.
2. We demonstrate that naïve uncertainty estimates from deep learning models without adjustment are poorly calibrated and thus unsuitable for reliable decision-making. We show how to transform these estimates into rigorous confidence intervals with formal statistical guarantees using the principled technique of conformal prediction.
3. We validate our framework on two distinct and clinically relevant ap-

plications: the estimation of patellar cartilage volume from knee MRI and the computation of left ventricular ejection fraction from cardiac CINE MRI, demonstrating its effectiveness and generalizability.

2. Methods

We propose a dynamic acquisition pipeline that iterates over a set of undersampling rates, assesses the uncertainty of derived clinical metrics and stops the scan once a predefined confidence threshold is reached. The pipeline operates as follows: after each k-space acquisition step, we first generate a set of M plausible reconstructions $\{\mathbf{x}^{(m)}\}_{m=1}^{M}$ from the currently undersampled k-space data $\mathbf{y}_t$ using a probabilistic reconstruction model, PHiRec [15], which we describe in Section 2.1. In Section 2.2 we showcase how to propagate uncertainty where each candidate reconstruction $\mathbf{x}^{(m)}$ is segmented by a deterministic segmentation network, $S(\cdot)$, yielding a set of segmentations $\{\mathbf{s}^{(m)}\}_{m=1}^{M}$, where $\mathbf{s}^{(m)} = S(\mathbf{x}^{(m)})$. From these segmentations, a clinical metric of interest, $\mathbf{w}$, is computed via a function $f(\cdot)$, resulting in a set of metric samples $\{\mathbf{w}^{(m)}\}_{m=1}^{M}$, where $\mathbf{w}^{(m)} = f(\mathbf{s}^{(m)})$. In our experiments, these metrics are the left ventricular ejection fraction and patellar cartilage volume. We quantify the uncertainty of the metric $\mathbf{w}$ by its empirical standard deviation, which is then calibrated using a scaling factor derived from conformal prediction (Section 2.3). This entire process—reconstruction, segmentation, metric estimation, and uncertainty calibration—is repeated after each acquisition step. The acquisition is terminated when the calibrated uncertainty bound falls below a user-defined threshold, ε. A schematic of this iterative process is provided in Figure 1.

2.1. Probabilistic Hierarchical Reconstruction (PHiRec)

The goal of MR reconstruction is to recover a high-fidelity image $\mathbf{x} \in \mathbb{C}^D$ from undersampled k-space measurements $\mathbf{y} \in \mathbb{C}^M$, where $M \ll D$. The relationship is described by the forward model:

$$\mathbf{y} = \mathcal{A}(\mathbf{x}) + \mathbf{n} = \mathcal{M}\mathcal{F}\mathcal{S}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathcal{S}$ denotes the coil sensitivity mapping, $\mathcal{F}$ is the Fourier transform, $\mathcal{M}$ is the binary sampling mask, and $\mathbf{n}$ represents measurement noise. The combined operator $\mathcal{A}$ is the forward encoding model.

Instead of seeking a single point estimate, we aim to model the full posterior distribution $p(\mathbf{x} \mid \mathbf{y})$. This inverse problem can be framed as a de-aliasing

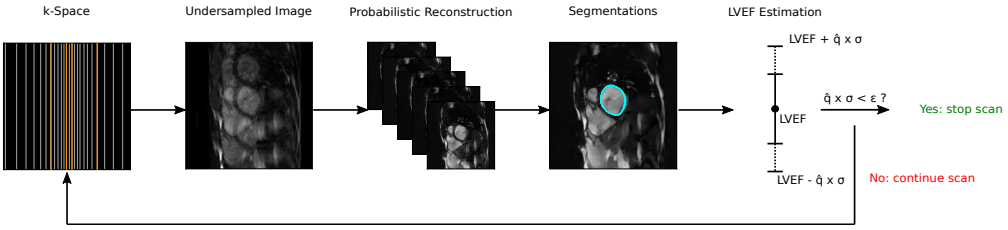

Figure 1: Overview of the proposed dynamic and iterative MR acquisition framework. At each time step t, k-space data $\mathbf{y}_t$ is acquired. A probabilistic model generates M candidate reconstructions $\{\mathbf{x}^{(m)}\}$, which are then passed to a segmentation network. The resulting segmentations are used to compute a distribution of a clinical metric (e.g., LVEF). The uncertainty of this metric is estimated and calibrated. Based on a user-defined stopping criterion (i.e., if the uncertainty is below a threshold ε), the scan is either terminated or continued with the acquisition of the next k-space segment.

task by conditioning on the zero-filled reconstruction $\mathbf{x}_u = \mathcal{A}^*(\mathbf{y})$, where $\mathcal{A}^*$ is the adjoint of the forward operator. We thus seek to model the distribution $p(\mathbf{x} \mid \mathbf{x}_u)$.

To this end, we employ our previously proposed Probabilistic Hierarchical Reconstruction (PHiRec) model [15], a state-of-the-art method for uncertainty quantification in MR reconstruction. Its high sampling speed, compared to alternatives like diffusion models, makes it particularly suitable for the real-time requirements of our dynamic acquisition setting. PHiRec is a hierarchical conditional variational autoencoder (CVAE) that models the distribution of reconstruction artifacts across multiple scales. It uses a hierarchy of latent variables $\mathbf{z}_{1:L} = \{\mathbf{z}_1, \ldots, \mathbf{z}_L\}$, where each level l corresponds to a different image resolution. The generative process is defined as:

$$p(\mathbf{x}|\mathbf{x}_u) = \int p(\mathbf{x}|\mathbf{z}_1, \mathbf{x}_u) \left(\prod_{l=1}^{L-1} p(\mathbf{z}_l|\mathbf{z}_{l+1}, \mathbf{x}_u) \right) p(\mathbf{z}_L|\mathbf{x}_u)\, d\mathbf{z}_{1:L}. \tag{2}$$

The model is trained by maximizing the evidence lower bound (ELBO) on the log-likelihood of the data, which, for a given ground truth image $\mathbf{x}$, is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{ELBO}}(\mathbf{x}, \mathbf{x}_u) =& \mathbb{E}_{q(\mathbf{z}_{1:L}|\mathbf{x},\mathbf{x}_u)}[\log p(\mathbf{x}|\mathbf{z}_{1:L}, \mathbf{x}_u)] \\ & - \sum_{l=1}^{L} \mathrm{KL}\left(q(\mathbf{z}_l|\mathbf{z}_{>l}, \mathbf{x}, \mathbf{x}_u) \parallel p(\mathbf{z}_l|\mathbf{z}_{>l}, \mathbf{x}_u)\right). \end{aligned} \tag{3}$$

Here, $q(\cdot)$ is the approximate posterior (encoder) and $p(\cdot)$ is the prior (decoder). Assuming Gaussian distributions for the likelihood and the latent priors, maximizing the ELBO is equivalent to minimizing a loss function composed of two main terms: a reconstruction loss (typically mean squared error) corresponding to the first term, and a regularization term that penalizes the divergence between the approximate posterior and the prior distributions for each latent level, given by the sum of KL-divergences.

2.1.1. Segmentation

For the downstream segmentation task, we employed a standard 2D U-Net architecture [32]. The network follows a symmetric encoder-decoder structure with four downsampling stages. The encoder path begins with an initial block of two 3×3 convolutions, mapping the input channels to 64 feature maps. Each subsequent downsampling stage consists of a 2×2 max-pooling operation followed by two more 3×3 convolutions, doubling the number of feature channels at each step ($64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow 1024$).

The decoder path symmetrically mirrors this design. At each stage, it uses a 2×2 transposed convolution to upsample the feature maps, followed by concatenation with the corresponding feature maps from the encoder path via skip connections. These concatenated features are then processed by two 3×3 convolutions. All convolutional layers, except for the final one, are followed by Batch Normalization and a ReLU activation function. A final 1×1 convolution maps the 64 feature channels from the last upsampling block to the number of output classes, producing the segmentation logits. The model was trained with the fully sampled reconstructions as input, using a hybrid loss function, defined as the sum of a soft Dice loss ($\mathcal{L}_{\text{Dice}}$) and a standard Cross-Entropy loss ($\mathcal{L}_{\text{CE}}$):

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}} \tag{4}$$

2.2. Uncertainty Propagation through the Processing Pipeline

To quantify how uncertainty from the reconstruction stage affects downstream clinical metrics, we propagate samples through the entire analysis pipeline. This Monte Carlo approach allows us to estimate the posterior distribution of a given metric, conditioned on the undersampled k-space data.

Let $\mathbf{y}$ denote the undersampled k-space measurements for a given scan. Our probabilistic reconstruction network is trained to sample from the posterior distribution of the fully-sampled image, $p(\mathbf{x}|\mathbf{y})$. For each $\mathbf{y}$, we draw

a set of M plausible image reconstructions:

$$\{\hat{\mathbf{x}}^{(m)}\}_{m=1}^{M} \sim p(\mathbf{x}|\mathbf{y}), \tag{5}$$

where each $\hat{\mathbf{x}}^{(m)}$ is a sample. Let $T(\cdot)$ be a deterministic function representing a downstream task (e.g., segmentation followed by volume calculation) that computes a scalar metric of interest, $\mathbf{w}$. By applying this function to each reconstruction sample, we generate a set of metric samples:

$$\{\mathbf{w}^{(m)} = T(\hat{\mathbf{x}}^{(m)})\}_{m=1}^{M}. \tag{6}$$

These samples, $\{\mathbf{w}^{(m)}\}$, form an empirical estimate of the metric's posterior distribution, $p(\mathbf{w}|\mathbf{y})$. From this set, we can compute the final prediction as the sample mean, $\hat{\mathbf{w}}$, and an estimate of its uncertainty as the sample standard deviation, $\sigma_{\mathbf{w}}$:

$$\hat{\mathbf{w}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{w}^{(m)}, \qquad \sigma_{\mathbf{w}} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\mathbf{w}^{(m)} - \hat{\mathbf{w}})^2} \tag{7}$$

This allows us to define a one-standard-deviation interval, $\mathcal{I}_{\text{std}} = [\hat{\mathbf{w}} - \sigma_{\mathbf{w}}, \hat{\mathbf{w}} + \sigma_{\mathbf{w}}]$. A smaller interval suggests a more certain prediction.

For both datasets, the function $T(\cdot)$ involves applying a trained segmentation network, $S(\cdot)$, to the reconstruction samples. For each subject, we generate $M = 20$ reconstructions, yielding a set of M segmentation masks $\{\hat{\mathbf{s}}^{(m)} = S(\hat{\mathbf{x}}^{(m)})\}_{m=1}^{M}$. These masks are then used to compute the final clinical metrics.

2.3. Uncertainty Calibration via Conformal Prediction

While the standard deviation $\sigma_{\mathbf{w}}$ provides a useful heuristic for uncertainty, the resulting intervals lack formal statistical guarantees. To construct prediction intervals with rigorous theoretical properties, we employ the split conformal prediction framework [33, 34]. This method transforms heuristic uncertainty estimates into valid prediction intervals that are guaranteed to contain the true, unknown value with a user-specified probability.

Formally, for a new test sample with undersampled data y, we aim to construct a prediction interval $\mathcal{C}(\mathbf{y})$ for the true metric $\mathbf{w}$ that satisfies the marginal coverage guarantee:

$$\mathbb{P}(\mathbf{w} \in \mathcal{C}(\mathbf{y})) \geq 1 - \alpha \tag{8}$$

where $\alpha \in (0, 1)$ is a user-defined tolerable error rate. This procedure requires a dedicated calibration set $D_{\text{calib}} = \{(\mathbf{y}_i, \mathbf{w}_i)\}_{i=1}^{n_{\text{calib}}}$, where samples are assumed to be exchangeable with the test data.

The core idea is to define a nonconformity score that quantifies how "unusual" a prediction is, given our heuristic uncertainty. For our symmetric intervals based on the standard deviation, we define the score for each calibration sample i as the normalized absolute error:

$$sc_i = \frac{|\mathbf{w}_i - \hat{\mathbf{w}}_i|}{\sigma_{\mathbf{w},i}} \tag{9}$$

where $\hat{\mathbf{w}}_i$ and $\sigma_{\mathbf{w},i}$ are the mean prediction and standard deviation derived from the Monte Carlo samples for calibration sample i and $\mathbf{w}_i$ is the ground truth value. These scores $\{sc_i\}_{i=1}^{n_{\text{calib}}}$ measure the error in units of predicted standard deviations.

We then compute a correction factor, $\hat{q}$, by taking the $\lceil (1-\alpha)(n_{\text{calib}}+1) \rceil$-th value of the sorted nonconformity scores. This $\hat{q}$ represents the empirical quantile of the normalized errors on the calibration set. The final conformal prediction interval for a new test prediction $(\hat{\mathbf{w}}, \sigma_{\mathbf{w}})$ is then constructed by scaling the standard deviation by this factor:

$$\mathcal{C}(\mathbf{y}) = [\hat{\mathbf{w}} - \hat{q}\sigma_{\mathbf{w}}, \hat{\mathbf{w}} + \hat{q}\sigma_{\mathbf{w}}] \tag{10}$$

By construction, this interval is guaranteed to achieve the coverage defined in Eq. (8). The width of this interval provides a rigorous, data-driven measure of uncertainty. A wider interval indicates that a larger deviation from the prediction is needed to be considered "conformal," implying higher uncertainty and a greater probability of a large error. This property makes these intervals highly suitable for defining an uncertainty-based stopping criterion for accelerated MRI.

3. Experiments

We demonstrate the dynamic uncertainty-guided MR acquisition strategy described in Section 2 on two datasets that provide raw multi-coil k-space data: The public Stanford Knee MRI Multi-Task Evaluation (SKM-TEA) [35], and an in-house cardiac CINE MR dataset. We simulate the acquisition process by retrospectively undersampling the k-space data. The two datasets contain anatomical segmentations which allow training a segmentation network and quantify anatomical volumes as introduced earlier, as well

as evaluating the method. In the following we describe the experimental setup and the experimental details.

3.1. Experimental Setup

To simulate dynamic MR acquisition, we retrospectively undersample the fully-sampled raw k-space data using predefined sampling masks corresponding to various acceleration factors $R \in 4, 8, \ldots, 32$, as described in the Data section. Starting from a highly undersampled input, we incrementally reveal additional k-space data by successively applying sampling masks of increasing density. The acquisition simulation proceeds by moving to the next predefined k-space subset at each acquisition step, mimicking a real-time, progressive acquisition process. At each step, we generate a reconstruction sample and compute the downstream metric of interest (i.e., patellar cartilage volume or LVEF) along with a calibrated uncertainty interval as described below. The scan is automatically terminated when the uncertainty interval for the downstream metric becomes sufficiently tight—i.e., once the width of the interval falls below a user-defined threshold ε. For the patellar cartilage volume, we defined $\varepsilon_v = 0.5cm^3$ and for the LVEF as $\varepsilon_{LVEF} = 15\%$.

3.2. Data and Preprocessing

3.2.1. SKM-TEA

The SKM-TEA dataset provides raw multi-coil k-space measurements of knee MRIs, accompanied by manual segmentations of six anatomical structures. While the original dataset includes undersampling masks for up to 16x acceleration based on a Poisson-Disc sampling pattern, we generated a new set of masks to explore higher acceleration factors. We followed the same sampling methodology to create masks for a set of acceleration factors $R \in \{4, 8, 12, 16, 20, 24, 28, 32\}$. The input images for our models were obtained by applying the adjoint operator ($\mathcal{A}^*$) to the zero-filled, retrospectively undersampled multi-coil k-space data. As in the original dataset, consistent spatial dimension across subjects was ensured by zero-padding the undersampled k-space.

For our experiments, a dedicated calibration set was required. We created this set by reallocating five subjects from the original training set and five from the original validation set. The test set remained unchanged, as defined by the original benchmark. This partitioning resulted in final splits of 81, 28, 10, and 36 subjects for training, validation, calibration, and testing, respectively.

3.2.2. CINE

Our in-house CINE dataset comprises raw multi-coil k-space measurements from cardiac MRI scans, with corresponding manual segmentations for the left ventricle (LV), myocardium (Myo), and right ventricle (RV). Multi-slice 2D Cartesian data was acquired with a balanced steady-state free precession (bSSFP) CINE (2x GRAPPA accelerated) in 8 breath-holds of 12s duration (2 slices per breathhold) each with 20 seconds pause in between. Further imaging parameters include 1.9×1.9mm in-plane (acquired and reconstructed) resolution, slice thickness 8 mm, temporal resolution 40 ms, 25 cardiac phases (reconstructed), TE=1.06ms, TR=2.12ms, flip angle 52°, bandwidth=915Hz/px.

Due to the dynamic nature of the CINE acquisition, we employed a Variable-density Incoherent Spatio-Temporal Acquisition (VISTA) sampling pattern [36] to generate the retrospective undersampling masks. Masks were generated for the same set of acceleration factors R as used for the SKM-TEA dataset. Similarly, input images were reconstructed by applying the adjoint operator ($\mathcal{A}^*$)) to the zero-filled multi-coil k-space data. Consistent spatial dimension across subjects was ensured by zero-padding the undersampled k-space. Like for the SKM-TEA dataset, consistent spatial dimension across subjects was ensured by zero-padding the undersampled k-space.

The full CINE cohort includes 134 subjects suitable for the reconstruction task. A subset of 40 subjects has corresponding ground truth segmentations (manually annotated by experienced radiologists with ¿10 years of experience in cardiovascular imaging), enabling the segmentation task. This disparity required us to define two distinct data splits. To ensure a fair comparison and prevent data leakage, the test and calibration sets were kept consistent across both splits.

- **Reconstruction Task:** The 134 subjects were partitioned into 95 for training, 24 for validation, and 10 for testing.
- **Segmentation Task:** The 40 subjects with annotations were split into 20 for training, 5 for validation, 5 for calibration, and the same 10 for testing.

3.3. Training Procedures

This section outlines the training protocols for the reconstruction and segmentation models. For reproducibility, we maintained consistent hyperparameters where appropriate and detail any dataset-specific adaptations.

3.3.1. Reconstruction

A separate PHiRec model was trained for each acceleration factor $R \in \{4, 8, \ldots, 32\}$. The models operate on 2D complex-valued image slices, which are processed as two-channel real-valued tensors ($\mathbb{R}^{H \times W \times 2}$ of real and imaginary parts), where H and W represent the image height and width. For both datasets, the images were normalized per-slice as in the original paper.

The model architecture was adapted to the different spatial dimensions of the datasets: 512×512 for SKM-TEA and 192×192 for CINE. This was achieved by setting the number of resolution levels in the PHiRec network to seven for SKM-TEA and five for CINE. All other model parameters were kept consistent.

We trained each reconstruction model using the Adam optimizer [37] with a learning rate of 1×10^{-4} and a batch size of 12. To improve generalization, we applied spatial data augmentation in the form of random flips and rotations. Training was performed for a fixed duration of 10 days on a single NVIDIA A100 GPU, which was sufficient to ensure convergence. For each acceleration factor, we selected the model checkpoint that achieved the highest Structural Similarity Index (SSIM) [38] on the validation set for final evaluation.

3.3.2. Segmentation

The U-Net was trained on normalized 2D image slices with spatial dimensions of 512×512 for SKM-TEA and 192×192 for CINE. We used the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 12. Also here, we used random flips and rotations to increase generalization and model robustness. Training was performed on NVIDIA RTX 2080Ti GPUs. The final model for each dataset was selected based on the checkpoint that achieved the highest mean Dice Similarity Coefficient (DSC) on the validation set.

3.4. Downstream Metrics and Uncertainty Quantification

3.4.1. Patellar Cartilage Volume for SKM-TEA

For the SKM-TEA dataset, we used the patellar cartilage volume as our downstream metric. We defined a function $V(\cdot)$ that calculates the volume from a segmentation mask in cm^3, using the voxel spacing provided in the image metadata. This yields a set of volume samples $\{v^{(m)} = V(\hat{\mathbf{s}}^{(m)})\}_{m=1}^{M}$. From these samples, we compute the final volume prediction, $\hat{v}$, and its associated uncertainty, σ_v.

3.4.2. Ejection Fraction for CINE

For the CINE dataset, the metric of interest was the Left Ventricular Ejection Fraction (LVEF), a critical biomarker for cardiac function. Calculating LVEF requires segmenting the left ventricle at two specific cardiac phases: end-diastole (ED) and end-systole (ES).

For each subject, we generate 20 reconstruction samples for both the ED scan, $\{\hat{\mathbf{x}}_{\text{ED}}^{(m)}\}$, and the ES scan, $\{\hat{\mathbf{x}}_{\text{ES}}^{(m)}\}$. We then apply the segmentation network to each, obtaining paired sets of segmentation masks: $\{\hat{\mathbf{s}}_{\text{ED}}^{(m)}\}$ and $\{\hat{\mathbf{s}}_{\text{ES}}^{(m)}\}$. The corresponding ED and ES volumes, $v_{\text{ED}}^{(m)}$ and $v_{\text{ES}}^{(m)}$, are calculated for each possible pairing. This yields us $M = 20 \times 20 = 400$ LVEF samples using its clinical definition:

$$\text{LVEF}^{(m)} = \frac{v_{\text{ED}}^{(m)} - v_{\text{ES}}^{(m)}}{v_{\text{ED}}^{(m)}} \times 100\% \tag{11}$$

This process yields an empirical distribution of LVEF values, from which we compute the final prediction, $\hat{\text{LVEF}}$, and its uncertainty, σ_{LVEF}.

While the standard deviation $\sigma_{\mathbf{w}}$ provides an intuitive measure of uncertainty, the resulting interval $\mathcal{I}_{\text{std}}$ offers no formal guarantees on its coverage probability (i.e., how often it contains the true, unknown metric value). To construct prediction intervals with rigorous statistical guarantees, we leverage the conformal prediction framework, as detailed in the following section.

3.4.3. Calibration Details

For all experiments, we set the target error rate to $\alpha = 0.1$, aiming for 90% coverage. The calibration procedure was performed independently for each acceleration factor R. This was done using the dedicated calibration sets described previously, with $n_{\text{calib}} = 10$ for SKM-TEA and $n_{\text{calib}} = 5$ for CINE.

4. Results

After training the models and calibrating the uncertainties as described in the Methods section, we evaluated our proposed framework in three steps. First, we quantified the performance of the underlying reconstruction and segmentation models. Second, we analyzed the behavior of the dynamic stopping mechanism, comparing outcomes with and without uncertainty calibration. Finally, we present qualitative examples to visualize the method's performance.

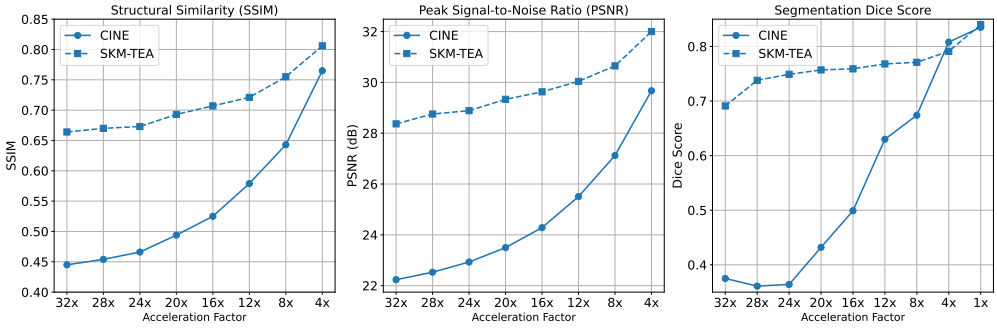

Figure 2: Quantitative evaluation of reconstruction and segmentation performance across different acceleration factors for two datasets (SKM-TEA and CINE). Each subplot shows one metric: Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Segmentation Dice Score. The x-axis denotes the acceleration factor (higher values correspond to stronger undersampling). Performance consistently improves with decreasing acceleration, where the models for SKM-TEA yield better metrics compared to the models for the CINE dataset due to a differences in undersampling.

4.1. Reconstruction and Segmentation Performance

To validate the underlying models, we evaluated reconstruction quality using the Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR), and segmentation accuracy using the Dice Similarity Coefficient (DSC). Figure 2 shows that for both datasets, all metrics improved as the acceleration rate decreased, with the highest scores achieved in the fully-sampled setting. This trend is expected, as more k-space data provides more information for both reconstruction and the downstream segmentation task.

We also observed that performance on the CINE dataset was notably lower than on the SKM-TEA dataset across all acceleration factors. This difference can be attributed to the more challenging VISTA undersampling pattern used for the CINE data, which tends to produce stronger aliasing artifacts in zero-filled images compared to the Poisson-disk sampling used for SKM-TEA.

4.2. Dynamic Stopping Behavior and Coverage

We next analyzed the behavior of the uncertainty-guided stopping mechanism. As shown in Figure 3, our method successfully determines patient-specific scan durations rather than relying on a fixed acquisition time. To assess the impact of calibration, we compared the distribution of stopping points determined by uncalibrated versus calibrated uncertainties. Without

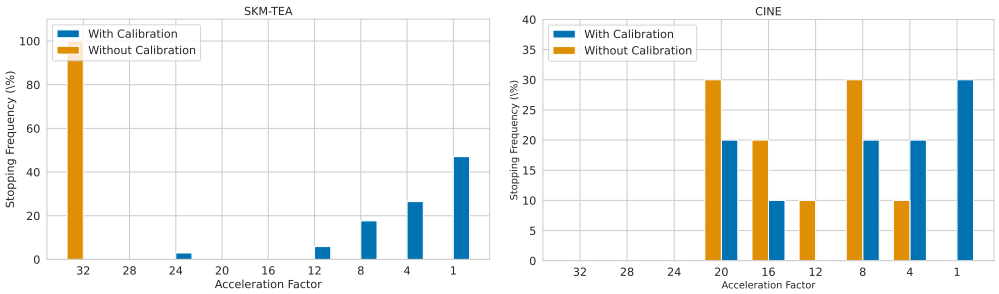

Figure 3: Distribution of stopping frequencies across experimental settings for SKM-TEA (left) and CINE (right) datasets. The x-axis shows different experimental settings (acceleration factors), and the y-axis indicates the percentage of runs that stopped at each setting. Results are shown separately for models with and without calibration. Notably, models without calibration tend to stop earlier compared to models with calibration for both datasets.

calibration, the mechanism consistently terminated scans prematurely. This was particularly pronounced for the SKM-TEA dataset, where every scan was stopped at the highest acceleration factor (32x). In contrast, applying conformal calibration resulted in significantly longer and more varied scan durations.

To evaluate the statistical reliability of the uncertainty intervals at the moment of stopping, we measured the empirical coverage—the percentage of test cases where the ground truth metric fell within the predicted interval. For SKM-TEA, uncalibrated intervals achieved only 17.6% coverage, which increased to 61.1% after calibration. For the CINE dataset, coverage improved from 20.0% to 85.7% with calibration. While calibration substantially improved reliability, the empirical coverage for both datasets remained below the target of 90%.

Finally, our method is computationally efficient and suitable for real-time implementation. The entire pipeline—encompassing reconstruction, segmentation, and calibrated uncertainty estimation—requires approximately 28 ms per slice on an NVIDIA A100 GPU. This translates to an overhead of less than 0.4 seconds for a typical CINE volume and under 4.5 seconds for a full SKM-TEA volume, making the approach practical for inline clinical decision-making.

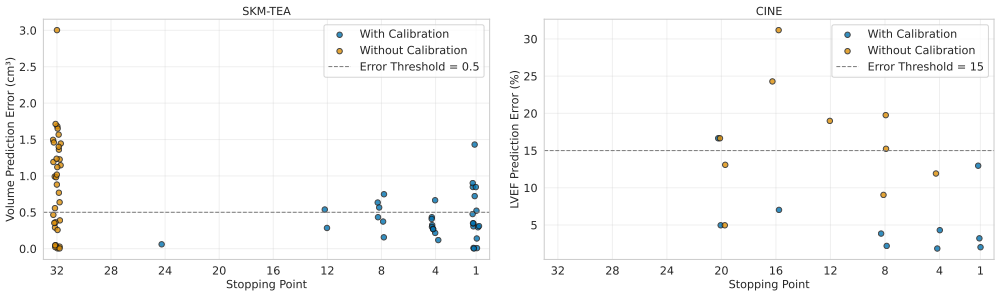

Figure 4: Performance of uncertainty-guided early stopping with and without calibration. The prediction error at point of stopping is plotted against the stopping point determined by our uncertainty criterion. Results are shown for (left) the SKM-TEA dataset, with prediction error measured in cm^3, and (right) the CINE dataset, with LVEF prediction error shown in percent. Each point represents a single reconstruction. The model with calibration (blue) reliably terminates acquisition at lower acceleration rates with errors mostly below the task-specific thresholds (dashed lines). In contrast, the uncalibrated model (orange) often produces reconstructions with unacceptable errors while stopping comparably early.

4.3. Qualitative Results

To provide a qualitative understanding of our dynamic stopping mechanism, Figures 5 and 6 present representative cases of both early and late scan terminations. Each figure visualizes the evolution of the reconstruction, segmentation, and the downstream metric along with its calibrated uncertainty as more k-space data is acquired. As expected, we observe a consistent trend across all examples: as the acquisition progresses, reconstruction quality and segmentation accuracy visibly improve. Additional reconstruction examples are displayed in Figure 7 and 8. Concurrently, the downstream metric estimation converges toward the ground truth value while the corresponding uncertainty bands narrow. Crucially, instances of high uncertainty consistently correspond to visible artifacts, segmentation errors, and larger deviations in the final metric, confirming that our uncertainty estimates effectively track acquisition quality.

5. Discussion

5.1. Principal Findings

Our study demonstrates that downstream uncertainty can effectively guide dynamic MRI scan termination, enabling patient-specific acquisition times.

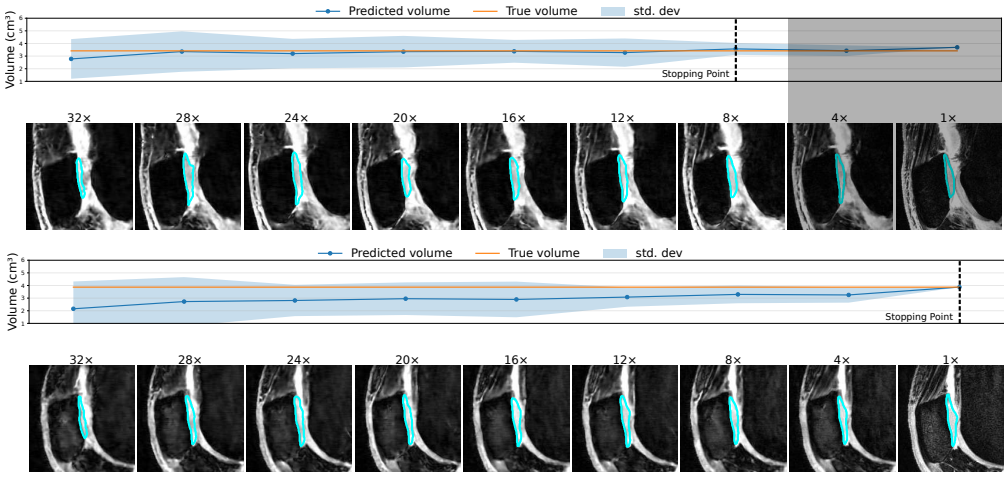

Figure 5: Patellar cartilage volume estimations along with calibrated uncertainty bounds and examples of reconstructions and segmentations for all acceleration factors for the SKM-TEA dataset. The top subject (MTR_196) displays a case of lower uncertainty (and notably lower error) whereas the bottom subject (MTR_120) displays higher uncertainty and therefore a longer scan time. The grayed out area indicates the skipped scans.

We establish that conformal calibration is indispensable for this task, as uncalibrated uncertainty estimates from deep learning models are systematically overconfident and lead to premature scan termination with unacceptably high error rates. By providing statistically meaningful uncertainty intervals, our calibrated approach offers a robust framework for balancing scan time and diagnostic confidence.

5.2. Interpretation of Key Findings

Our results confirm the expected trade-off between acquisition speed and image quality, where both reconstruction and segmentation performance improve with increased k-space sampling. The performance gap between the SKM-TEA and CINE datasets highlights the significant impact of the k-space sampling strategy on task difficulty. To place our results in context, we verified that the performance of our models on fully-sampled data is comparable to benchmarks reported in the original SKM-TEA publication [35] and related CINE segmentation work [39], confirming the validity of our underlying models.

The core contribution of this work lies in the dynamic stopping mechanism. The dramatic difference between uncalibrated and calibrated stopping

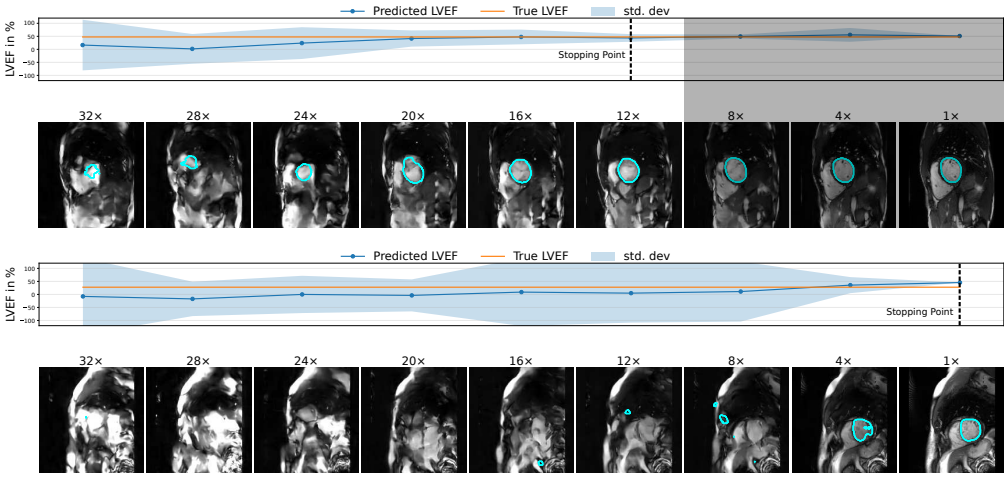

Figure 6: LVEF estimates along with calibrated uncertainty bounds and examples of reconstructions and segmentations for all acceleration factors for the CINE dataset. The top subject displays a case of lower uncertainty whereas the bottom subject displays higher uncertainty. One can clearly see the differences in segmentation quality that lead to the high uncertainty for the lower subject. The grayed out area indicates the skipped scans.

points (Figure 3) reveals a critical insight: raw neural network uncertainties are not reliable proxies for model error. The uncalibrated models were consistently overconfident, terminating scans when the downstream metric error was still high (Figure 4). This misalignment poses a significant clinical risk. Conformal calibration corrects this by widening the uncertainty intervals to better reflect the true potential for error, leading to more appropriate and safer stopping decisions. This finding aligns with a growing body of literature emphasizing the necessity of calibration for deploying machine learning models in high-stakes medical applications [26, 40].

Furthermore, our qualitative results (Figures 5, 6) visually corroborate these quantitative findings. The clear correlation between wider uncertainty bands, visible image artifacts, and inaccurate segmentations provides intuitive evidence that the calibrated uncertainty is a meaningful and trustworthy indicator of quality.

5.3. Limitations and Future Work

Several limitations of this study present avenues for future work. First, our reconstruction model does not enforce data consistency, which could potentially improve image quality and reduce uncertainty, leading to earlier,

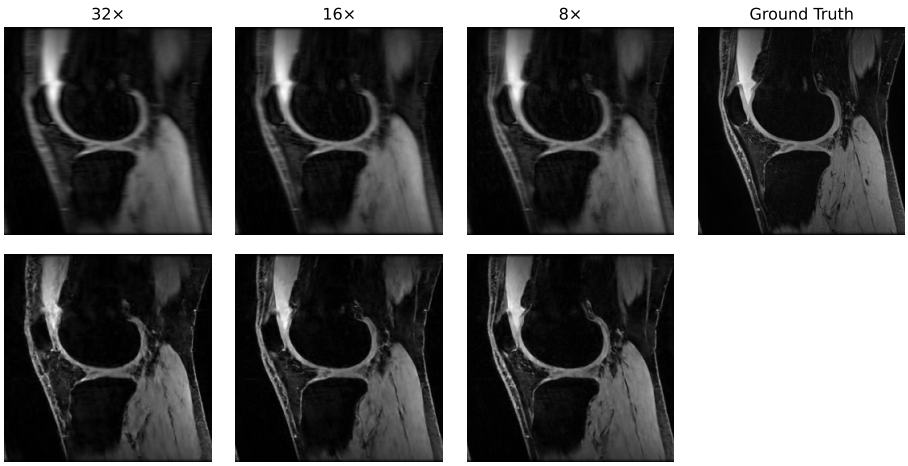

Figure 7: Example reconstructions for the SKM-TEA dataset. The top row shows the undersampled input images along with the ground truth, and the bottom row shows the corresponding model reconstructions at 32x, 16x, and 8x acceleration.

more efficient scan termination. Integrating a data consistency term within the probabilistic framework is a clear next step.

Second, our framework adapts the scan duration but not the acquisition strategy, as it relies on a discrete set of predefined undersampling masks. A more advanced approach would optimize the k-space trajectory in real-time, selecting the most informative measurements to reduce uncertainty as quickly as possible. This could be achieved using techniques like reinforcement learning or Bayesian experimental design.

Finally, while calibration significantly improved the reliability of our uncertainty intervals, the empirical coverage on the test sets did not consistently meet the 90% target. This indicates a potential distribution shift between the calibration and test sets. Moreover, the resulting calibrated intervals, while statistically valid, may still be too wide for certain clinical applications. Future work should investigate more advanced calibration techniques, such as those that account for subgroup shifts, and explore alternative stopping criteria that can achieve a better trade-off between statistical rigor and clinical utility.

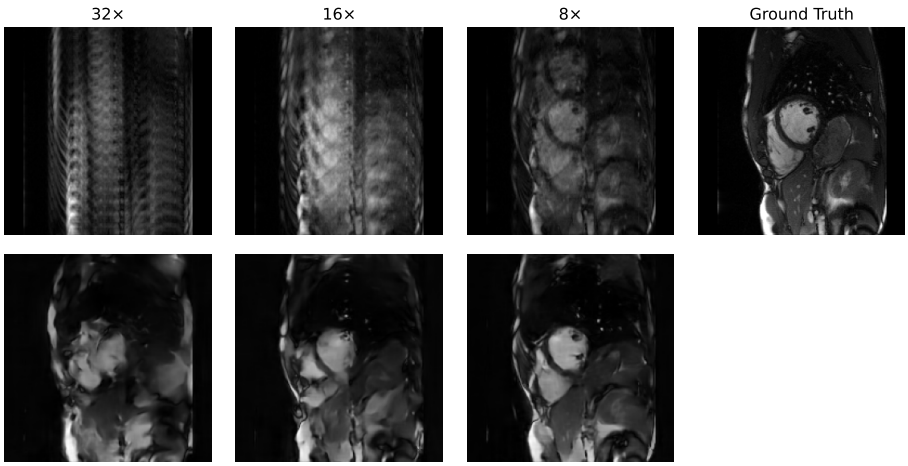

Figure 8: Example reconstructions for the CINE dataset. The top row shows the undersampled input images along with the ground truth, and the bottom row shows the corresponding model reconstructions at 32x, 16x, and 8x acceleration. Note the different artifact patterns in the inputs, stemming from the different undersampling schemes compared to the SKM-TEA dataset (VISTA vs. Poisson-disk).

6. Conclusion

Deep learning has shown tremendous promise for accelerated magnetic resonance imaging, offering the potential to significantly reduce scan times while maintaining diagnostic image quality [11]. However, the intrinsic uncertainty inherent in these reconstruction methods necessitates careful modeling and quantification to ensure clinical reliability. Despite this critical need, limited work has explored how to effectively leverage uncertainty estimates within clinical imaging pipelines, and it remains unclear when sufficient k-space data has been acquired for reliable downstream applications.

Our work addresses this fundamental gap by providing a principled approach for leveraging uncertainty arising during accelerated MR acquisition to determine reliable stopping points based on prediction certainty. We demonstrate that uncertainty estimates can be effectively utilized to enable dynamic scan termination, allowing for patient-specific optimization of scan duration. Our methodology is validated across two distinct datasets, and we further enhance the reliability of stopping decisions through uncertainty calibration with mathematical guarantees.

Additional research directions include the incorporation of temporal con-

sistency constraints for dynamic imaging, the development of more sophisticated undersampling strategies that better reflect clinical acquisition patterns, and the exploration of alternative uncertainty quantification methods that may provide more informative estimates for stopping decisions.

References

[1] C. Westbrook, J. Talbot, MRI in Practice, John Wiley & Sons, 2018.

[2] J. B. Andre, B. W. Bresnahan, M. Mossa-Basha, M. N. Hoff, C. P. Smith, Y. Anzai, W. A. Cohen, Toward quantifying the prevalence, severity, and cost associated with patient motion during clinical mr examinations, Journal of the American College of Radiology 12 (7) (2015) 689–695.

[3] M. Lustig, D. Donoho, J. M. Pauly, Sparse MRI: The application of compressed sensing for rapid MR imaging, Magnetic Resonance in Medicine 58 (6) (2007) 1182–1195. doi:10.1002/mrm.21391.
URL https://onlinelibrary.wiley.com/doi/10.1002/mrm.21391

[4] F. Knoll, K. Hammernik, C. Zhang, S. Moeller, T. Pock, D. K. Sodickson, M. Akcakaya, Deep-learning methods for parallel magnetic resonance imaging reconstruction: A survey of the current approaches, trends, and issues, IEEE signal processing magazine 37 (1) (2020) 128–140.

[5] O. N. Jaspan, R. Fleysher, M. L. Lipton, Compressed sensing mri: a review of the clinical literature, The British journal of radiology 88 (1056) (2015) 20150487.

[6] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, P. Boesiger, Sense: sensitivity encoding for fast mri, Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 42 (5) (1999) 952–962.

[7] M. A. Griswold, P. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, A. Haase, Generalized autocalibrating partially parallel acquisitions (grappa), Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 47 (6) (2002) 1202–1210.

[8] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, F. Knoll, Learning a variational network for reconstruction of accelerated mri data, Magnetic resonance in medicine 79 (6) (2018) 3055–3071.

[9] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, D. Rueckert, A deep cascade of convolutional neural networks for dynamic mr image reconstruction, IEEE transactions on Medical Imaging 37 (2) (2017) 491–503.

[10] K. Hammernik, T. Küstner, B. Yaman, Z. Huang, D. Rueckert, F. Knoll, M. Akçakaya, Physics-driven deep learning for computational magnetic resonance imaging: Combining physics and machine learning for improved medical imaging, IEEE Signal Processing Magazine 40 (1) (2023) 98–114. doi:10.1109/MSP.2022.3215288.

[11] R. Heckel, M. Jacob, A. Chaudhari, O. Perlman, E. Shimron, Deep learning for accelerated and robust mri reconstruction, Magnetic Resonance Materials in Physics, Biology and Medicine 37 (3) (2024) 335–368.

[12] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On faithfulness and factuality in abstractive summarization, arXiv preprint arXiv:2005.00661 (2020).

[13] S. K. Aithal, P. Maini, Z. C. Lipton, J. Z. Kolter, Understanding hallucinations in diffusion models through mode interpolation (2024). arXiv:2406.09358.
URL https://arxiv.org/abs/2406.09358

[14] V. Antun, F. Renna, C. Poon, B. Adcock, A. C. Hansen, On instabilities of deep learning in image reconstruction and the potential costs of ai, Proceedings of the National Academy of Sciences 117 (48) (2020) 30088–30095. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1907377117, doi:10.1073/pnas.1907377117.
URL https://www.pnas.org/doi/abs/10.1073/pnas.1907377117

[15] P. Fischer, K. Thomas, C. F. Baumgartner, Uncertainty estimation and propagation in accelerated mri reconstruction, in: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, Springer, 2023, pp. 84–94.

[16] A. M. Wundram, P. Fischer, S. Wunderlich, H. Faber, L. M. Koch, P. Berens, C. F. Baumgartner, Leveraging probabilistic segmentation models for improved glaucoma diagnosis: A clinical pipeline approach, in: Medical Imaging with Deep Learning, 2024.

[17] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, PMLR, 2016, pp. 1050–1059.

[18] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Advances in neural information processing systems 30 (2017).

[19] C. F. Baumgartner, K. C. Tezcan, K. Chaitanya, A. M. Hötker, U. J. Muehlematter, K. Schawkat, A. S. Becker, O. Donati, E. Konukoglu, Phiseg: Capturing uncertainty in medical image segmentation (2019). doi:10.48550/ARXIV.1906.04045.
URL https://arxiv.org/abs/1906.04045

[20] V. Edupuganti, M. Mardani, S. Vasanawala, J. Pauly, Uncertainty quantification in deep mri reconstruction, IEEE Transactions on Medical Imaging 40 (1) (2020) 239–250.

[21] D. Narnhofer, A. Effland, E. Kobler, K. Hammernik, F. Knoll, T. Pock, Bayesian uncertainty estimation of learned variational mri reconstruction, IEEE transactions on medical imaging 41 (2) (2021) 279–291.

[22] N. R. Huttinga, T. Bruijnen, C. A. van den Berg, A. Sbrizzi, Gaussian processes for real-time 3d motion and uncertainty estimation during mr-guided radiotherapy, Medical Image Analysis 88 (2023) 102843.

[23] J. Schlemper, D. C. Castro, W. Bai, C. Qin, O. Oktay, J. Duan, A. N. Price, J. Hajnal, D. Rueckert, Bayesian deep learning for accelerated mr image reconstruction, in: International workshop on machine learning for medical image reconstruction, Springer, 2018, pp. 64–71.

[24] J. N. Morshuis, M. Hein, C. F. Baumgartner, Segmentation-guided mri reconstruction for meaningfully diverse reconstructions, in: MICCAI Workshop on Deep Generative Models, Springer, 2024, pp. 180–190.

[25] J. N. Morshuis, C. Schlarmann, T. Küstner, C. F. Baumgartner, M. Hein, Mind the detail: Uncovering clinically relevant image details in accelerated mri with semantically diverse reconstructions (2025). arXiv:2507.00670.
URL https://arxiv.org/abs/2507.00670

[26] A. M. Wundram, P. Fischer, M. Mühlebach, L. M. Koch, C. F. Baumgartner, Conformal performance range prediction for segmentation output quality control, in: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, Springer, 2024, pp. 81–91.

[27] P. Daudé, R. Ramasawmy, A. Javed, R. J. Lederman, K. Chow, A. E. Campbell-Washburn, Inline automatic quality control of 2d phase-contrast flow mri for subject-specific scan time adaptation, Magnetic Resonance in Medicine 92 (2) (2024) 751–760.

[28] L. Pineda, S. Basu, A. Romero, R. Calandra, M. Drozdzal, Active mr k-space sampling with reinforcement learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 23–33.

[29] Z. Wang, B. Li, H. Yu, Z. Zhang, M. Ran, W. Xia, Z. Yang, J. Lu, H. Chen, J. Zhou, et al., Promoting fast mr imaging pipeline by full-stack ai, Iscience 27 (1) (2024).

[30] Z. Huang, J. Duan, Y. Xie, Y. Liu, Udnet: Unified deep network based on transformer and multi-stage fusion for brain tumor classification from undersampled mri, Neurocomputing 619 (2025) 129109.

[31] Z. Wu, T. Yin, Y. Sun, R. Frost, A. van der Kouwe, A. V. Dalca, K. L. Bouman, Learning task-specific strategies for accelerated mri, IEEE Transactions on Computational Imaging (2024).

[32] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, Springer, 2015, pp. 234–241.

[33] V. Vovk, A. Gammerman, G. Shafer, Algorithmic learning in a random world, Vol. 29, Springer, 2005.

[34] A. N. Angelopoulos, S. Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification, arXiv preprint arXiv:2107.07511 (2021).

[35] A. D. Desai, A. M. Schmidt, E. B. Rubin, C. M. Sandino, M. S. Black, V. Mazzoli, K. J. Stevens, R. Boutin, C. Ré, G. E. Gold, et al., Skm-tea: A dataset for accelerated mri reconstruction with dense image labels for quantitative clinical evaluation, arXiv preprint arXiv:2203.06823 (2022).

[36] R. Ahmad, H. Xue, S. Giri, Y. Ding, J. Craft, O. P. Simonetti, Variable density incoherent spatiotemporal acquisition (vista) for highly accelerated cardiac mri, Magnetic resonance in medicine 74 (5) (2015) 1266–1278.

[37] D. P. Kingma, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13 (4) (2004) 600–612.

[39] T. Wang, X. Xu, J. Xiong, Q. Jia, H. Yuan, M. Huang, J. Zhuang, Y. Shi, Ica-unet: Ica inspired statistical unet for real-time 3d cardiac cine mri segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23, Springer, 2020, pp. 447–457.

[40] A. N. Angelopoulos, A. P. Kohli, S. Bates, M. Jordan, J. Malik, T. Alshaabi, S. Upadhyayula, Y. Romano, Image-to-image regression with distribution-free uncertainty quantification and applications in imaging, in: International Conference on Machine Learning, PMLR, 2022, pp. 717–730.