Uncertainty-based Deep Learning Methods for Robust and Reliable Cardiac MRI Segmentation and Analysis

Tewodros Weldebirhan Arega

UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

UNIVERSITÉ DE BOURGOGNE

THÈSE DE DOCTORAT

DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

PRÉPARÉE À L'UNIVERSITÉ DE BOURGOGNE

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat d'Instrumentation, Informatique de l'image

par

TEWODROS WELDEBIRHAN AREGA

Uncertainty-based Deep Learning Methods for Robust and Reliable Cardiac MRI Segmentation and Analysis

Thèse présentée et soutenue à Dijon, le 30 novembre 2023

Composition du Jury :

GARREAU MIREILLE	Professeure à l'Université de Rennes 1	Examinatrice
PETITJEAN CAROLINE	Professeure à l'Université de Rouen Normandie	Rapporteuse
CLARYSSE PATRICK	Directeur de Recherche CNRS, CREATIS	Président et Rapporteur
MERIAUDEAU FABRICE	Professeur à l'Université de Bourgogne	Directeur de thèse
BRICQ STÉPHANIE	Maitresse de Conférences HDR à l'Université de Bourgogne	Codirectrice de thèse

N° 0 0 1

ABSTRACT

Deep learning-based segmentation methods have shown promise in automating the segmentation of cardiac MRI images, but they still face challenges in robustly segmenting small, and ambiguous regions with irregular shapes like myocardial scars. Additionally, these models struggle with domain shifts and out-of-distribution (OOD) samples, which makes them unreliable and limits their usage in clinical practice. The main objective of this thesis is to enhance the robustness and reliability of deep learning models for cardiac MRI segmentation and analysis by leveraging uncertainty estimates.

To improve the segmentation of myocardial scars, a segmentation model is proposed that integrates uncertainty information into the learning process. Uncertainty estimation is achieved by utilizing Monte-Carlo dropout-based Bayesian neural networks during training, which are then incorporated into the loss function. This approach yields improved segmentation accuracy and probability calibration, achieving state-of-the-art performance on publicly available datasets focused on scar segmentation from Late Gadolinium Enhancement (LGE) MRI. The method demonstrates superior performance, particularly for visually challenging images with higher epistemic uncertainty.

To enhance the reliability of segmentation models, an uncertainty-based quality control (QC) framework is introduced to identify failed segmentations before further analysis. The QC framework utilizes a Bayesian Swin transformer-based U-Net for the segmentation of T1 mapping images and employs image-level uncertainty features to detect poorly segmented images. Experimental results on private and public datasets demonstrate that the proposed QC method significantly outperforms other state-of-the-art uncertainty-based QC methods, particularly in challenging scenarios. After rejecting inaccurate segmentations, T1 mapping, and Extracellular volume (ECV) values are automatically computed, enabling reliable characterization of myocardial tissues in healthy and pathological cases.

Furthermore, a post-hoc OOD detection method is proposed to identify and reject outlier images. This method utilizes the encoder features of the segmentation model and similarity metrics to enhance the trustworthiness of segmentation models. Experimental results demonstrate that the proposed method outperforms state-of-the-art feature space-based and uncertainty-based OOD detection methods across the various OOD datasets. This further safeguards performance by rejecting unsuitable outliers.

Keywords: Cardiac MRI Segmentation, Myocardial scar, T1 mapping MRI, LGE MRI, ECV, Uncertainty Estimation, Quality Control, Out-of-distribution (OOD) detection

RÉSUMÉ

Les méthodes de segmentation basées sur l'apprentissage profond se sont révélées prometteuses pour automatiser la segmentation des images IRM cardiaques, mais elles sont toujours confrontées à des défis pour segmenter de manière robuste des régions petites et ambiguës aux formes irrégulières comme les cicatrices myocardiques. De plus, ces modèles sont confrontés aux changements de domaine et aux échantillons hors distribution (OOD), ce qui les rend peu fiables et limite leur utilisation dans la pratique clinique. L'objectif principal de cette thèse est d'améliorer la robustesse et la fiabilité des modèles d'apprentissage profond pour la segmentation et l'analyse d'IRM cardiaque en exploitant les estimations d'incertitude.

Pour améliorer la segmentation des cicatrices myocardiques, un modèle de segmentation est proposé qui intègre les informations d'incertitude dans le processus d'apprentissage. L'estimation de l'incertitude est obtenue en utilisant des réseaux neuronaux bayésiens basés sur une méthode Monté Carlo Drop out pendant la formation, qui sont ensuite incorporés dans la fonction de perte. Cette approche permet d'améliorer la précision de la segmentation et l'étalonnage des probabilités, obtenant ainsi des performances de l'état de l'art sur des ensembles de données accessibles au public axés sur la segmentation des cicatrices à partir de l'IRM avec rehaussement tardif au gadolinium (LGE). La méthode démontre des performances supérieures, en particulier pour les images visuellement difficiles avec une incertitude épistémique plus élevée.

Pour améliorer la fiabilité des modèles de segmentation, un cadre de contrôle qualité (CQ) basé sur l'incertitude est introduit pour identifier les segmentations ayant échoué avant une analyse plus approfondie. Le cadre CQ utilise un U-Net basé sur un Transformer bayésien Swin pour la segmentation des images cartographiques T1 et utilise des caractéristiques d'incertitude au niveau de l'image pour détecter les images mal segmentées. Les résultats expérimentaux sur des ensembles de données privés et publics démontrent que la méthode de CQ proposée surpasse considérablement les autres méthodes de CQ de l'état de l'art basées sur l'incertitude, en particulier dans des scénarios difficiles. Après avoir rejeté les segmentations inexactes, la cartographie T1 et les valeurs du volume extracellulaire (ECV) sont automatiquement calculées, permettant une caractérisation fiable des tissus myocardiques dans les cas sains et pathologiques.

De plus, une méthode de détection OOD post-hoc est proposée pour identifier et rejeter les images aberrantes. Cette méthode utilise les fonctionnalités d'encodeur du modèle de

segmentation et les métriques de similarité pour améliorer la fiabilité des modèles de segmentation. Les résultats expérimentaux démontrent que la méthode proposée surpasse les méthodes de détection OOD de l'état de l'art basées sur l'espace des caractéristiques et l'incertitude dans les différents ensembles de données OOD. Cela garantit davantage les performances en rejetant les valeurs aberrantes inappropriées.

Mots clés: Segmentation IRM cardiaque, cicatrice myocardique, IRM de cartographie T1, IRM LGE, volume extracellulaire, estimation de l'incertitude, contrôle qualité (CQ), détection hors distribution (OOD)

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my esteemed thesis supervisors, Prof. Fabrice Meriaudeau and Dr. Stéphanie Bricq for their unwavering motivation and invaluable support throughout the entire duration of this project. Their guidance and encouragement have been instrumental in shaping the outcome of my research. Their mentorship has not only shaped my academic growth but has also inspired me to become a better researcher and individual. Merci beaucoup!

I am also thankful to Dr. Alain Lalande and François Legrand for their expertise, assistance, and contributions during my Ph.D. research. Furthermore, I would like to express my gratitude to the French National Research Agency (ANR) for providing the financial support that enabled me to undertake this Ph.D. research.

I extend my sincere thanks to the members of the thesis committee, Prof. Mireille Garreau, Prof. Caroline Petitjean, and Prof. Patrick Clarysse, for generously dedicating their time to evaluate and review my thesis, and offering insightful feedback.

I am grateful to my colleagues and friends at the ImViA laboratory, including Kibrom, Youssef, Solomon, Adrian, Arslan, Raabid, Alpha, Diya, Ali, and Mahya for their collaboration, insightful discussions, and willingness to share their expertise. I am thankful for the positive influence you have had on my academic and personal development. Additionally, I would like to express my gratitude to the Tegaru community in Dijon and Europe as a whole, particularly Tewele, Desnet, Yeman, Hiliwi, Sara, Aman, Meron, and Abush, for making my time enjoyable and providing a sense of belonging. I thank you all for being there with me both during the bad and good times throughout this journey.

Last but not least, I want to convey my heartfelt gratitude to my beloved family for their unwavering love and support over the years. Your belief in me has been a constant source of motivation, and I am eternally grateful for your presence in my life. Lastly, I would like to thank God for granting me the resilience and patience needed to overcome obstacles and reach this milestone in my life.

Contents

List of Figures

List of Tables

Acronyms

- **AUC:** Area under the ROC Curve
- **BS:** Brier Score
- **bSSFP:** balanced Steady-State Free Precession
- **CMR:** Cardiovascular Magnetic Resonance
- **DCM:** Dilated cardiomyopathy
- **ECV:** Extracellular volume
- **FCN:** Fully Convolutional Network
- **HCM:** Hypertrophic cardiomyopathy
- **HD:** Hausdorff Distance
- **ID:** In-distribution
- **LGE:** Late Gadolinium Enhancement
- **MI:** Myocardial Infarction
- **MOLLI:** Modified Look-Locker Inversion Recovery
- **MRI:** Magnetic Resonance Imaging
- **MSP:** Maximum Softmax Probability
- **OOD:** out-of-distribution
- **PCC:** Pearson Correlation Coefficient
- **QC:** Quality Control
- **SVM:** Support-Vector Machine

I

Context and Background

1

Introduction

The introduction chapter of this thesis begins with a presentation of the main motivations behind the research and an overview of the proposed solutions including the contributions of the thesis. The chapter concludes with a thesis outline that provides a brief summary of the subsequent chapters and the key works that will be presented, which includes a list of publications resulting from the research conducted.

1.1/ Motivation and Contributions

Cardiovascular diseases (CVDs) are a group of diseases that affect the heart and blood vessels, and they are the leading cause of death globally, accounting for 31% of all deaths [WHO, 2017]. Heart attacks and strokes account for the majority of deaths from cardiovascular diseases. To diagnose and evaluate these diseases, medical imaging techniques, such as Cardiovascular Magnetic Resonance (CMR) imaging, cardiac computed tomography (CT), and echocardiography, are frequently used. These imaging techniques help clinicians to non-invasively assess the qualitative and quantitative characteristics of cardiac anatomical structures and functions [Chen et al., 2019b].

Compared to other cardiac imaging modalities, cardiovascular magnetic resonance (CMR) imaging is widely recognized as the gold-standard non-invasive imaging tool for many CVDs. It is particularly effective in visualizing and measuring cardiovascular anatomy, volumes, and function, as well as in characterizing myocardial tissues [Schulz-Menger et al., 2020]. This is due to the fact that CMR provides superior image quality with excellent soft tissue contrast and allows for multi-planar imaging, which helps it to provide a more comprehensive view of the heart and blood vessels from different angles. Moreover, it does not use ionizing radiation, unlike cardiac CT, which can expose patients to potentially harmful levels of radiation [Ripley et al., 2016, Bai et al., 2017].

To extract relevant information from the CMR images, medical experts have been manually segmenting the cardiac structures and pathologies in their clinical workflow. The

obtained information is essential for the diagnosis, management, treatment planning, and prognosis evaluation of CVDs. For example, Native T1 mapping and extracellular volume (ECV) values which are extracted from cardiac structures of native and post-contrast cardiac T1 mapping CMR images, can be used to quantify diffuse myocardial fibrosis and to characterize myocardial tissues [Arega et al., 2023a]. However, a trained expert may spend up to 20 minutes analyzing images of a single subject at two points in the cardiac cycle [Bai et al., 2017], which can be tiresome and tedious, and may suffer from intra- and inter-expert variability. The inter-observer variability can even increase more when the images have low quality and are more challenging. This shows the need for automatic, precise, and reliable segmentation of cardiac images to assist clinicians in their workflow.

Recently, deep learning-based segmentation methods have shown remarkable progress in the automatic segmentation of cardiac structures like the left-ventricular blood pool (LV), myocardium (MYO), and right-ventricular blood pool (RV). This is because these structures have a relatively consistent shape and structure that makes them easier to segment by deep learning models. However, accurately segmenting myocardial scars from CMR images is more challenging, as these scars have irregular shapes, small sizes, and lack contrast with the surrounding region [Lalande et al., 2021]. Therefore, deep learning-based segmentation methods for CMR image segmentation, particularly for pathologies such as myocardial scars, need further improvements to segment the images robustly.

Although deep learning-based segmentation methods have demonstrated significant potential in CMR images, they are not widely used in clinical practice for several reasons. One of the primary reasons is that the models are not robust and reliable enough to handle domain shifts or out-of-distribution (OOD) samples. Various factors can influence the performance of segmentation models. One of these can be image quality deterioration during acquisition resulting in image artifacts like ghosting, blurring, and smearing. These artifacts can arise from factors related to the scanner, patient physiology, or acquisition settings [Galati et al., 2022]. Other factors that can impact the performance of segmentation models include changes in demographics, modalities, acquisition protocols, scanner vendors, anatomical variability, or even adversarial attacks that alter the input images' statistical properties [Galati et al., 2022]. These complexities can cause segmentation models to produce inaccurate results, potentially leading to incorrect clinical decisions.

While techniques like data augmentation and image harmonization can enhance generalization, models will inevitably encounter unfamiliar shifts. Regardless of the model's performance, there will always be cases where it is unsuitable for accurate predictions. This poses a challenge for deploying segmentation networks as it can result in silent failures that go undetected [González, 2023]. Therefore, it is crucial to develop quality control methods that can detect failed segmentation results from In-distribution (ID) im-

ages, as well as OOD detection techniques that can identify and reject images that are significantly different from the ID images. In this thesis, different deep learning-based methods are proposed to address these issues in cardiac Magnetic Resonance Imaging (MRI) segmentation using different approaches. To improve the segmentation of cardiac pathologies such as myocardial scar, we proposed a segmentation model that leverages uncertainty estimates during training. The proposed method generates uncertainty estimates using Monte-Carlo dropout during training and incorporates it into the loss function to improve the segmentation accuracy and probability calibration. The experimental results show that the proposed method outperforms the top-ranked methods of the challenge and improves the segmentation results, particularly in visually challenging and difficult images with higher epistemic uncertainty.

To identify failed segmentation results and improve the reliability of segmentation models, we proposed an uncertainty-based quality control framework for T1 mapping and extracellular volume (ECV) analysis. The proposed framework consists of three parts: segmentation of cardiac structures using a Bayesian Swin transformer-based U-Net, a novel uncertainty-based quality control method to detect inaccurate segmentation results, and automatic computation of T1 mapping and ECV values to characterize myocardial tissues. The proposed quality control method utilizes image-level uncertainty features as input to a random forest-based classifier/regressor to determine the quality of the segmentation outputs. The experimental results show that the proposed QC method outperforms other state-of-the-art uncertainty-based QC methods and achieves an excellent agreement with the manually computed myocardial T1 and ECV values.

Lastly, to detect and reject OOD images that are far from the ID images and enhance the trustworthiness of the models, we proposed a post-hoc out-of-distribution (OOD) detection method that can be used with any pre-trained segmentation model. The proposed method leverages features extracted from the segmentation model's encoder blocks and employs Mahalanobis distance as a metric to measure the similarity between the input image and the validation set of in-distribution images. The experimental results show that the proposed method outperforms existing feature space-based and uncertainty-based OOD detection methods across various OOD datasets. The proposed method successfully detects near, mild, and far OOD images with high detection accuracy, showcasing the advantage of using the multi-scale and semantically rich representations of the encoder.

During the Ph.D., we also participated in three cardiac MR image segmentation MICCAI challenges. These three challenges involve the segmentation of cardiac structures from Multi-disease, Multi-center, and Multi-view Cardiac MR images (M&M2 2021 challenge) and from Cardiac MR images with Motion Artifacts (CMRxMotion 2022 challenge) and left atrial and scar quantification and segmentation from multi-center cardiac MR images (LAScarQS 2022 challenge). We won the LAScarQS 2022 challenge and ranked second

in the M&M2 2021 and CMRxMotion 2022 (in segmentation task) challenges. The details of the proposed methods that are used to solve the challenges can be found in Appendix A.

1.2/ THESIS OUTLINE

The remainder of the thesis is organized as follows:

Chapter 2 and 3 provide the necessary background on the clinical context, technical methods, and problem domain central to this thesis. Chapter 2 overviews relevant cardiac anatomy, physiology, and major cardiovascular diseases to establish the clinical motivation. Common diagnostic approaches and imaging modalities like cardiac MRI and echocardiography used to assess these diseases are discussed. Public datasets in cardiac MRI are also briefly highlighted. Chapter 3 provides the necessary background on deep learning and uncertainty estimation techniques, along with a review of state-of-the-art methods in cardiac MRI analysis. First, core machine and deep learning concepts are introduced, including neural network architectures, training procedures, and evaluation metrics. A literature review on recent developments for deep learning-based cardiac MRI segmentation is also conducted. Finally, different uncertainty estimation techniques are introduced, along with their application to medical image analysis tasks. These two chapters lay the groundwork for the proposed methods detailed in subsequent chapters (Chapters 4 - 6).

Chapters 4 - 6 discuss the main contributions of this thesis. Chapter 4 details the first contribution which proposes a novel segmentation model that leverages uncertainty estimates during the learning process. It employs Monte Carlo dropout to generate uncertainty estimates (sample variance) during training and incorporates this information into the loss function. This integration aims to enhance segmentation accuracy and probability calibration. The proposed method is validated on two publicly available datasets, namely EMIDEC MICCAI 2020 and LAScarQS MICCAI 2022. These datasets focus on segmenting infarcted myocardium and left atrial (LA) scars from LGE MRI.

Chapter 5 discusses the issue of incorrect segmentation results generated by deep learning-based methods in cardiac MR. It emphasizes the importance of accurate segmentation for downstream tasks, such as myocardial tissue characterization, and highlights the need for quality control methods to detect and reject failed segmentations before further analysis. It proposes a fully automatic framework for quality control in T1 mapping and extracellular volume (ECV) analysis. The framework consists of three parts. Firstly, it focuses on segmenting cardiac structures from native and post-contrast T1 mapping datasets using a Bayesian Swin transformer-based U-Net, achieving accurate initial seg-

mentations. It, then, introduces a novel uncertainty-based Quality Control (QC) method to detect inaccurate segmentation results. The QC method utilizes image-level uncertainty features as input to a random forest-based classifier to determine the quality of the segmentation outputs. After detecting and rejecting inaccurate segmentation results, it presents the automatic computation of T1 mapping and ECV values to characterize myocardial tissues in healthy and cardiac pathological cases.

Chapter 6 addresses the challenge of handling input images that deviate from the training distribution. It presents a novel post-hoc out-of-distribution (OOD) detection method that can be used with any pre-trained segmentation model. The proposed method leverages multi-scale representations extracted from the encoder blocks of the segmentation model. It utilizes the Mahalanobis distance as a metric to measure the similarity between the input image and the in-distribution images used during pre-training. The detection performance of the proposed method is compared with the state-of-the-art feature space-based and uncertainty-based OOD detection methods on 13 different OOD datasets, categorized as near, mild, and far OOD datasets based on their similarity to the in-distribution dataset.

Finally, Chapter 7 concludes the work presented in this thesis and discusses some limitations and potential future work.

1.3/ PUBLICATIONS

The work of this thesis is mainly based on the following three research publications:

1. **Arega, T. W.**, Bricq, S., & Meriaudeau, F. (2023). **Post-hoc Out-of-Distribution Detection for Cardiac MRI Segmentation**, *Computers in Biology and Medicine* [under review]

2. **Arega, T.W.**, Bricq, S., Grand, F.L., Jacquier, A., Lalande, A., & Mériaudeau, F. (2023). **Automatic uncertainty-based quality controlled T1 mapping and ECV analysis from native and post-contrast cardiac T1 mapping images using Bayesian vision transformer**. *Medical image analysis*, 86, 102773 [Arega et al., 2023a]

3. **Arega, T. W.**, Bricq, S., & Meriaudeau, F. (2021). **Leveraging uncertainty estimates to improve segmentation performance in cardiac MR**. *In Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021*, Strasbourg, France, October 1, 2021, Proceedings 3 (pp. 24-33). Springer

International Publishing. [Arega et al., 2021a]

During the Ph.D., the following peer-reviewed journals and conferences were published in which I was either the first author or contributing author. While these works are related to the topics and aims of this thesis, they are not directly included in the thesis itself.

1. **Arega, T. W.**, Bricq, S., Legrand, F., Jacquier, A., Lalande, A., & Meriaudeau, F. (2023). **Uncertainty-based Quality Controlled T1 Mapping and ECV Analysis using Bayesian Vision Transformer**. *In Medical Imaging with Deep Learning, short paper track*. [Arega et al., 2023b]

2. Martín-Isla, C., Campello, V.M., Izquierdo, C., Kushibar, K., Sendra-Balcells, C., Gkontra, P., Sojoudi, A., Fulton, M.J., **Arega, T.W.**, Punithakumar, K., Li, L., Sun, X., Khalil, Y.A., Liu, D., Jabbar, S., Queirós, S., Galati, F., Mazher, M., Gao, Z., Beetz, M., Tautz, L., Galazis, C., Varela, M., Hullebrand, M., Grau, V., Zhuang, X., Puig, D., Zuluaga, M.A., Mohy-ud-Din, H., Metaxas, D.N., Breeuwer, M.M., Geest, R.J., Noga, M.L., Bricq, S., Rentschler, M.E., Guala, A., Petersen, S.E., Escalera, S., Palomares, J.F., & Lekadir, K. (2023). **Deep Learning Segmentation of the Right Ventricle in Cardiac MRI: The M&Ms Challenge**. *IEEE Journal of Biomedical and Health Informatics*, 27, 3302-3313. [Martín-Isla et al., 2023]

3. Li, L., Wu, F., Wang, S., Luo, X., Martín-Isla, C., Zhai, S., Zhang, J., Liu7, Y., Zhang, Z., Ankenbrand, M.J., Jiang, H., Zhang, X., Wang, L., **Arega, T.W.**, Altunok, E., Zhao, Z., Li, F., Ma, J., Yang, X., Puybareau, É., Oksuz, I., Bricq, S., Li, W., Punithakumar, K., Tsaftaris, S.A., Schreiber, L.M., Yang, M., Liu, G., Xia, Y., Wang, G., Escalera, S., & Zhuang, X. (2022). **MyoPS: A Benchmark of Myocardial Pathology Segmentation Combining Three-Sequence Cardiac Magnetic Resonance Images**. *Medical image analysis*, 87, 102808. [Li et al., 2022a]

4. **Arega, T.W.**, Bricq, S., & Mériaudeau, F. (2022). **Automatic Quality Assessment of Cardiac MR Images with Motion Artefacts Using Multi-task Learning and K-Space Motion Artefact Augmentation**. *In International Workshop on Statistical Atlases and Computational Models of the Heart (pp. 418-428) STACOM@MICCAI 2022*. Cham: Springer Nature Switzerland. [Arega et al., 2022a]

5. **Arega, T. W.**, Bricq, S., & Meriaudeau, F. (2022). **Using Polynomial Loss and Uncertainty Information for Robust Left Atrial and Scar Quantification and Segmentation**. *In Challenge on Left Atrial and Scar Quantification and Segmentation (pp. 133-144), MICCAI 2022*. Cham: Springer Nature Switzerland. [Arega et al., 2022b]

6. Brahim, K., **Arega, T.W.**, Boucher, A., Bricq, S., Sakly, A., & Mériaudeau, F. (2022). **An Improved 3D Deep Learning-Based Segmentation of Left Ventricular My-**

ocardial Diseases from Delayed-Enhancement MRI with Inclusion and Classification Prior Information U-Net (ICPIU-Net). *Sensors* (Basel, Switzerland), 22. [Brahim et al., 2022]

7. **Arega, T. W.**, Legrand, F., Bricq, S., & Meriaudeau, F. (2021, September). **Using MRI-specific data augmentation to enhance the segmentation of right ventricle in multi-disease, multi-center, and multi-view cardiac MRI**. *In International Workshop on Statistical Atlases and Computational Models of the Heart (pp. 250-258), MICCAI 2021*. Cham: Springer International Publishing. [Arega et al., 2021b]

8. **Arega, T.W.**, Bricq, S., & Mériaudeau, F. (2023). **A Simple Uncertainty-based Quality Control for Cardiac MR Images Segmentation**. *Colloque Français d'Intelligence Artificielle en Imagerie Biomédicale (IABM 2023).*

2

Background: Clinical Context

This chapter provides relevant clinical background on cardiac anatomy, major cardiovascular diseases, diagnostic approaches, and cardiac imaging modalities. First, normal cardiac anatomy is reviewed, including heart walls, chambers, valves, arteries, and electrical conduction. Next, various cardiovascular diseases, such as myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, Takotsubo cardiomyopathy, and myocarditis are introduced. Standard techniques for diagnosing these cardiac conditions are then discussed. Finally, key cardiac imaging modalities are described, with a focus on magnetic resonance imaging and its applications in cine imaging, late gadolinium enhancement, and T1 mapping. Public datasets in cardiac MRI are also briefly summarized. Together, this background establishes core knowledge of the clinical context relevant to the main contributions of the thesis which are presented in later chapters.

2.1/ Cardiac Anatomy

The heart is a muscular organ that circulates blood in the body through the vascular or circulatory system. It is responsible for pumping blood and ensuring the circulation of oxygen, nutrients, hormones, and other essential substances throughout the body. The human heart is located within the thoracic cavity, medially between the lungs in the space known as the mediastinum, and it is surrounded by the pericardium, a two-layered protective sac [Foundation, 2017]. The heart's anatomy can be divided into different parts, including the walls, chambers, valves, coronary arteries, and electrical conduction system, as shown in Figure 2.1 (A).

2.1.1/ Heart Walls

The heart walls are the muscles that contract and relax to send blood throughout the body. The heart walls consist of three layers, namely the endocardium, the myocardium,

and the epicardium, as shown in Figure 2.1 (B). The endocardium is an inner layer of thin cells that line the heart's chambers and valves. The myocardium is the thick, muscular layer that contracts to pump blood throughout the body. The epicardium is the outermost layer, covering the surface of the heart [Foundation, 2017, Schülke, 1968].

The heart is divided into left and right sides by a layer of muscular tissue called the septum. The interatrial septum separates the atria, while the interventricular septum separates the ventricles. This separation prevents oxygenated and deoxygenated blood from mixing [Guo, 2017].

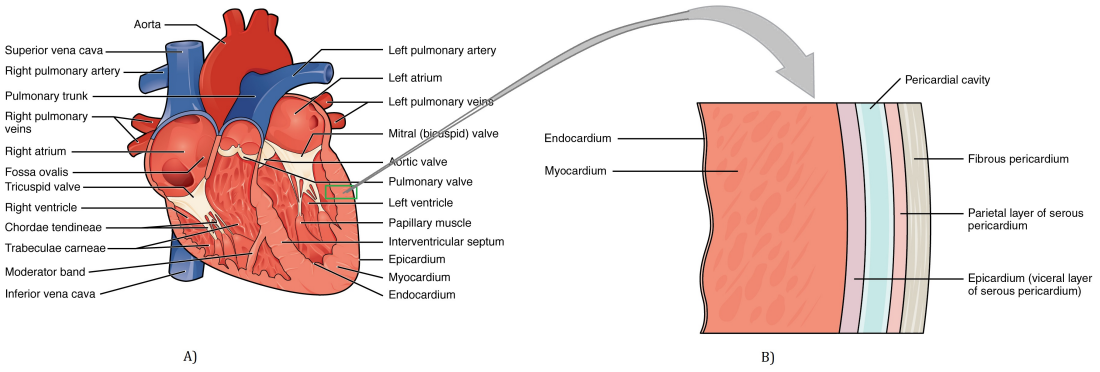

Figure 2.1: A) Internal structures of the heart, showing an anterior view of the four chambers, major vessels, and their branches, along with the valves. B) Pericardial membranes and layers of the heart wall. Image adapted from [Schülke, 1968].

2.1.2/ Heart Chambers

The heart is comprised of four chambers. It has two chambers on the top (atria) and two on the bottom (ventricles), one on each side of the heart, as shown in Figure 2.1 (A). The heart has two atria, the right atrium, and the left atrium. The atria receive blood returning to the heart. The right atrium receives deoxygenated blood from the body through the superior and inferior vena cava. The left atrium receives oxygenated blood from the lungs through the pulmonary veins. The heart also has two ventricles, the right ventricle, and the left ventricle. The ventricles pump blood out of the heart. The right ventricle pumps deoxygenated blood to the lungs through the pulmonary artery, while the left ventricle pumps oxygenated blood to the rest of the body through the aorta [Callie Tayrien, 2023].

2.1.3/ Heart Valves

Heart valves are structures that help regulate the flow of blood through the heart. There are four heart valves in the human heart: the tricuspid valve, the pulmonary valve, the mitral valve, and the aortic valve. The tricuspid valve is located between the right atrium and right ventricle, while the mitral (or bicuspid) valve is situated between the left atrium

and left ventricle. They prevent the backflow of blood into the atria when the ventricles contract. The pulmonary valve is on the right side of the heart, at the entrance to the pulmonary artery, preventing the backflow of blood from the artery into the right ventricle. The aortic valve is on the left side of the heart, at the entrance to the aorta, preventing the backflow of blood from the aorta into the left ventricle [Callie Tayrien, 2023].

2.1.4/ Coronary Artery

Coronary arteries are blood vessels that supply oxygenated blood to the heart muscle (myocardium). A constant supply of oxygen and nutrients is required for the proper functioning of the heart muscle, and this vital task is carried out by the coronary arteries. These essential substances are delivered to the heart muscle through two main branches, namely the left coronary artery and the right coronary artery, which originate from the aorta [Callie Tayrien, 2023].

2.1.5/ Electrical conduction system

The heart has its own electrical conduction system that coordinates and regulates its rhythm. It ensures synchronized contractions of the atria and ventricles. The sinoatrial (SA) node initiates the electrical impulses, causing the atria to contract. The impulses then pass through the atrioventricular (AV) node, bundle of His, and Purkinje fibers, leading to the contraction of the ventricles [Clinic, 2021].

2.2/ Cardiovascular Diseases

Having a thorough understanding of the heart's anatomy is crucial for comprehending the intricate interplay of its structures and functions. However, despite the remarkable design of this vital organ, it is vulnerable to numerous cardiovascular diseases that can hinder its regular operation. These conditions can impair the heart's ability to pump blood effectively, lead to structural abnormalities, or disrupt the electrical conduction system.

Cardiovascular diseases (CVDs) are a set of medical conditions that are defined by a range of diseases involving the heart or blood vessels. CVDs are the leading cause of death globally, with more people dying annually from cardiovascular disease than any other cause, according to World Health Organization (WHO) statistics [WHO, 2017]. An estimated 17.9 million people died from CVDs in 2019, which represents around 31% of all global deaths. Of these deaths, heart attacks and strokes account for 85%. CVDs can be broadly classified as coronary heart diseases (such as angina and heart attacks),

cerebrovascular diseases (such as stroke), and peripheral arterial disease (which affects blood flow to the limbs)[Lopez et al., 2022].

The causes of CVDs are numerous and complex, with multiple factors involved in the development and progression of cardiovascular diseases. These factors include lifestyle factors like tobacco use, physical inactivity, an unhealthy diet, and excessive alcohol consumption, as well as medical conditions like high blood pressure, diabetes, high blood cholesterol, obesity, and a family history of cardiovascular diseases. Age is the most important risk factor, with the risk of developing heart disease increasing with each decade of life. As a result, older adults are at greater risk of developing cardiovascular disease than younger adults [Rodgers et al., 2019].

There are many cardiovascular conditions that can affect the heart's ability to function properly. The most common ones are arrhythmia, atrial fibrillation, cardiomyopathy, congestive heart failure, coronary artery disease, heart attack (myocardial infarction), and pericarditis. Arrhythmia refers to any abnormal heart rhythm that occurs when the electrical impulses that regulate the heartbeat are disrupted. Atrial fibrillation is a type of arrhythmia in which the heart's upper chambers (atria) beat irregularly and out of sync with the lower chambers (ventricles). Cardiomyopathy refers to the thickening, enlargement, or stiffening of the heart muscle, which can impede its proper functioning. Congestive heart failure, on the other hand, happens when the heart is either too weak or stiff to pump blood throughout the body efficiently. Coronary artery disease, meanwhile, is characterized by the narrowing of the coronary arteries due to plaque buildup. A heart attack, also known as a myocardial infarction, occurs when a sudden blockage in the coronary artery cuts off oxygen supply to a part of the heart muscle. Pericarditis, which refers to an inflammation in the heart's lining or pericardium, can also impact the heart's function [NHS, 2022].

In this thesis, the focus will be directed towards CVDs that mainly impact the heart muscle.

2.2.1/ Myocardial Infarction

Myocardial Infarction (MI), commonly known as a heart attack, is caused by decreased or complete cessation of blood flow to a part of the heart muscle, which can lead to damage or death of the heart muscle cells. Most myocardial infarctions are due to underlying coronary artery disease. This happens when there is plaque or blood clots in the coronary artery (Figure 2.2), which is responsible for the supply of blood and oxygen to heart muscles. As the cells are deprived of oxygen, cellular injury occurs, which leads to the infarction or death of the cells. This condition can be asymptomatic or can cause severe hemodynamic problems and sudden death [Ojha et al., 2022].

The no-reflow phenomenon is a condition where the blood flow is restricted or blocked

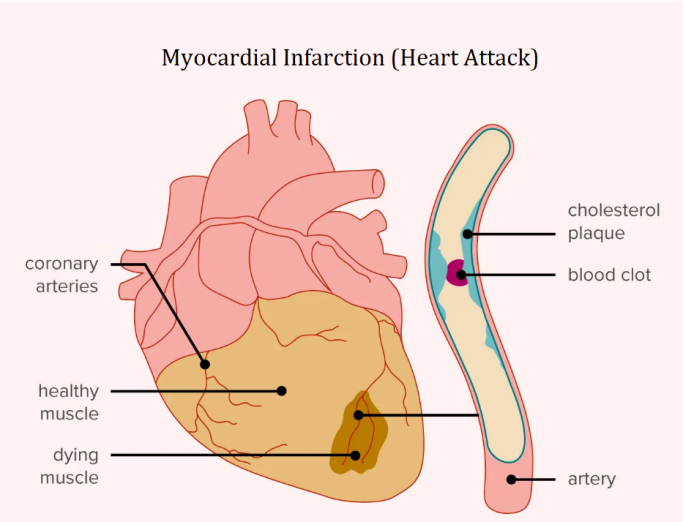

Figure 2.2: Myocardial Infarction. Image adapted from [Tricia Kinman, 2022].

even after the occlusion has been treated. This occurs because the small blood vessels, known as microvessels, that supply blood to the heart muscle are damaged during the heart attack and do not function properly, leading to reduced blood flow and oxygen delivery to the heart muscle. This incident usually appears in a proportion of patients with acute myocardial infarction following re-perfusion therapy of an occluded coronary artery. It can lead to inadequate myocardial perfusion, reperfusion injury, and poor healing of the infarct [Caiazzo et al., 2020].

2.2.2/ Dilated Cardiomyopathy

Dilated cardiomyopathy (DCM) is a non-ischaemic heart muscle disease where the heart muscle becomes weak and enlarged (Figure 2.3 (A)), which can make it harder for the heart to pump blood properly. It is characterized by left ventricular or bi-ventricular dilation and impaired contraction in the absence of coronary artery disease, hypertension, valvular disease, or congenital heart disease. DCM can be caused by a variety of factors, including genetics, infections, alcohol or drug abuse, and certain medications [Schultheiss et al., 2019].

2.2.3/ Hypertrophic Cardiomyopathy

Hypertrophic cardiomyopathy (HCM) is a genetic heart condition that is characterized by an abnormal thickening of the heart muscle, specifically of the left ventricle, as shown in Figure 2.3 (B). It is commonly caused by changes or mutations in genes that control heart muscle growth and function. These gene mutations lead to the thickening of the heart muscle. This thickening can cause stiffness of the heart muscle, which can lead to

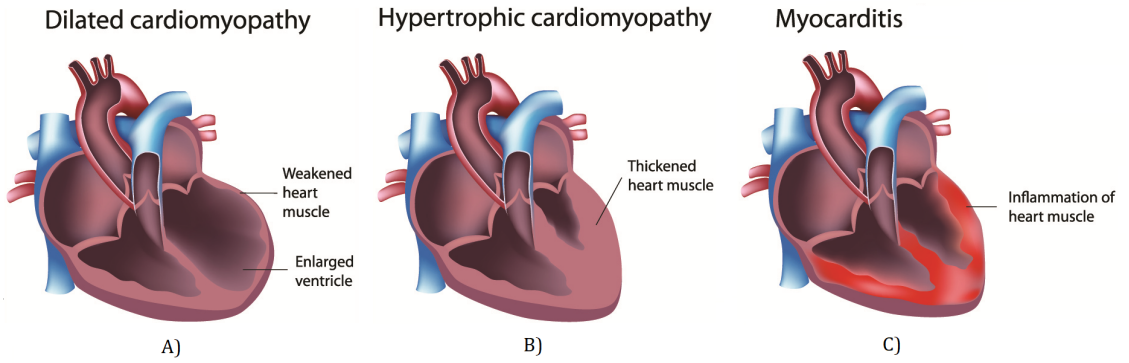

Figure 2.3: A) Dilated cardiomyopathy B) Hypertrophic cardiomyopathy C) Myocarditis. Image adapted from [Dhua, 2020].

problems with relaxation and filling of the heart, and can also obstruct the flow of blood out of the heart.

2.2.4/ MYOCARDITIS

Myocarditis is an inflammation of the heart muscle (myocardium), which can have a range of serious effects on the heart's ability to function properly (Figure 2.3 (C)). This inflammation typically occurs due to viral or bacterial infections or as a result of autoimmune disorders. The inflammation caused by myocarditis can weaken the heart muscle and affect the heart's electrical system, which can lead to cardiomyopathy and arrhythmia [Gilotra, 2023].

Myocarditis can be categorized into acute and chronic. Acute myocarditis is commonly caused by a viral infection and has a relatively fast onset. Symptoms of acute myocarditis may develop rapidly and can resolve quickly as well. In contrast, chronic myocarditis occurs when the disease takes longer to treat or when symptoms reappear after initial treatment. Chronic myocarditis may be attributed to systemic inflammatory conditions like autoimmune disorders, where the immune system attacks the body's healthy cells and tissues [Gilotra, 2023].

2.2.5/ TAKOTSUBO CARDIOMYOPATHY

Takotsubo cardiomyopathy or Takotsubo syndrome (TTS) is a non-ischemic form of cardiomyopathy that is characterized by sudden weakness or temporary weakening of the muscular portion of the heart. It causes the heart's main blood-plumping chamber (the left ventricle) to change shape and get larger. This condition is usually precipitated by a significant physical or emotional stressor. Some physical stressors that can cause TTS include sepsis, shock, subarachnoid hemorrhage, and pheochromocytoma, whereas

emotional stressors can include bereavement, divorce, or financial losses. TTS is also known as stress cardiomyopathy, broken heart syndrome, apical ballooning syndrome, and tako-tsubo syndrome, which takes its name from the Japanese word for ”octopus trap” due to the shape of the heart during the acute phase of the disease ((Figure 2.4) [Ghadri et al., 2018].

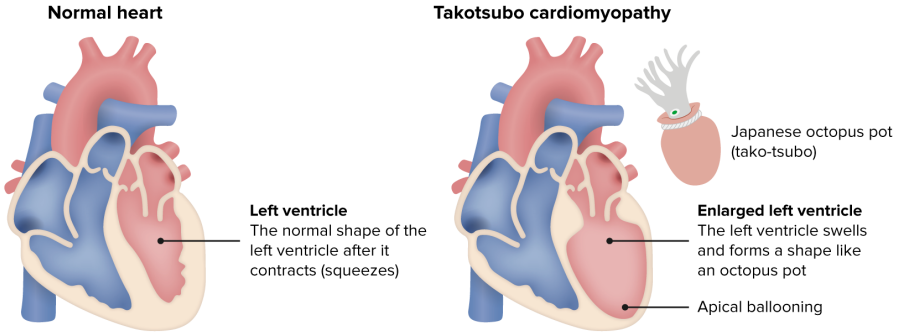

Figure 2.4: Takotsubo cardiomyopathy. Image adapted from [Stanley Oiseth, 2023].

2.3/ DIAGNOSING CVDS

The diagnosis of myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, myocarditis, and tako-tsubo cardiomyopathy usually involves a combination of medical history, physical examination, and diagnostic tests.

To aid in the diagnosis of myocardial infarction, an electrocardiogram (ECG) and blood test measuring cardiac biomarkers like troponin are commonly employed. Furthermore, imaging techniques such as echocardiograms and cardiac MRI scans are utilized to assess the structure and function of the heart [Ojha et al., 2022]. Similarly, for dilated cardiomyopathy and hypertrophic cardiomyopathy, echocardiograms and cardiac MRI scans are performed to evaluate cardiac structure and function. Additionally, genetic testing might be conducted to identify specific genetic mutations associated with these conditions. In the case of tako-tsubo cardiomyopathy, echocardiograms, cardiac MRI scans, or cardiac catheterization procedures are employed to assess the heart’s structure and function. These tests are instrumental in distinguishing tako-tsubo cardiomyopathy from other conditions that present similar symptoms, such as myocardial infarction [Ghadri et al., 2018]. For myocarditis, blood tests are conducted to detect signs of infection or inflammation, echocardiograms are employed to evaluate cardiac structure and function, and cardiac MRI scans or biopsies are used to identify any inflammation or damage to the heart muscle [Gilotra, 2023].

Preventing cardiovascular disease involves a range of lifestyle and medical interventions. Lifestyle changes like quitting smoking, increasing physical activity, and adopting

a healthy diet can help improve overall cardiovascular health. Medical interventions like controlling high blood pressure, cholesterol, and diabetes can also help reduce the risk of developing heart disease. In addition, early diagnosis and treatment are also critical for reducing the impact of cardiovascular disease on health outcomes [Harvard, 2022].

2.4/ Cardiac Imaging

2.4.1/ Cardiac magnetic resonance imaging (CMR)

Magnetic Resonance Imaging (MRI) is a non-invasive imaging technology that uses a powerful magnetic field and radio waves to create detailed three-dimensional anatomical images and is frequently used for disease detection, diagnosis, and treatment monitoring. It works by detecting the alteration in the rotational axis of protons that exist in the water that make up living tissues. MRI machines use powerful magnets that create a strong magnetic field, causing protons in the body to align with that field, and when radiofrequency current is added to the patient's body, the protons are excited, spinning out of balance and resisting the magnetic field. The MRI sensors detect the energy released as the protons realign with the magnetic field after the radio frequency field is turned off. The length of time and amount of energy released depends on the environment and chemical molecules in a patient's body. The emitted signals are measured after a certain period and converted into intensity levels using Fourier transformation which are displayed as shades of gray in a matrix arrangement of pixels. Different types of images can be created by varying the sequence of radiofrequency pulses applied and collected. [National Institutes of Health, 2022].

T1 and T2 are important parameters that describe the relaxation of protons back to their equilibrium state once the RF pulse is removed. T1 relaxation time is the time taken for longitudinal magnetization to recover by about 63% after an RF pulse is withdrawn, and T2 relaxation time is the time taken for transverse magnetization to decay by about 63% in excited tissues. T1 relaxation time increases with magnetic field strength, while T2 relaxation time is less dependent on magnetic field strength. T2 values are generally shorter than T1 values in biological tissues, and different types of tissues have different T1 and T2 values. For instance, fat has short T1 and T2 relaxation times, fluids have long T1 and T2 relaxation times, and non-fatty soft tissues like the myocardium have long T1 and short T2 relaxation times [Ripley et al., 2016].

During an MRI image acquisition, the patient is positioned inside a large magnet and must remain still to avoid blurring the image. Contrast agents with Gadolinium may be given to the patient intravenously before or during the MRI, increasing the pace at which protons realign with the magnetic field and brighter the image. MRI machines are excellent

for imaging non-bony or soft tissues without using X-rays' dangerous ionizing radiation. MRI machines do not emit harmful ionizing radiation, unlike X-ray and CT imaging, making them a safer option. The strength of the magnetic field used in MRI machines can range from 0.2 Tesla to 3 Tesla or higher, depending on the purpose of the scan. The higher the strength of the magnetic field, the more detailed the images that can be obtained. However, the magnetic field used in MRI machines can be very strong and may exert powerful forces on certain objects, such as those made of iron or other magnetizable materials. This means that patients with certain medical conditions, such as those with pacemakers or other implanted devices, may not be able to undergo an MRI scan [National Institutes of Health, 2022].

A cardiac MRI is a non-invasive imaging test that uses MRI to create detailed images of the heart and its structures, including the heart muscle, chambers, valves, and blood vessels. It can provide important information about the size and function of the heart, as well as the presence of any abnormalities, such as inflammation or scarring of the heart muscle. Cardiac MRI is a valuable diagnostic tool for a wide range of heart conditions, including myocardial infarction, myocarditis, hypertrophic cardiomyopathy, and dilated cardiomyopathy. It is also used to monitor the progression of heart disease over time and to assess the effectiveness of treatment.

There are two primary coordinate systems utilized in cardiac MRI, namely the body (scanner) planes and the cardiac planes. The body planes used in cardiac MRI are positioned perpendicular to the body's long axis and include the axial, sagittal, and coronal planes, as shown in Figure 2.5. These planes are useful for providing a general overview of the heart's anatomy. The axial plane is a horizontal plane that divides the body into upper and lower parts. It is obtained by imaging the body from the feet to the head or vice versa. The axial plane is used to visualize the heart from a cross-sectional view, with the images obtained perpendicular to the long axis of the body. The coronal plane is a vertical plane that runs from front to back, dividing the body into anterior (front) and posterior (back) portions. The sagittal plane is a vertical plane that divides the body into left and right portions. Sagittal images are obtained by imaging the body from one side to the other. In cardiac MRI, the axial plane can capture images of all four chambers of the heart and the pericardium simultaneously, while the sagittal plane can display the great vessels extending from the ventricles. The coronal plane is useful for assessing the left ventricular outflow tract, the left atrium, and the pulmonary veins [Ginat et al., 2011].

The standard cardiac imaging planes include the short-axis, horizontal long-axis (four-chamber view), and vertical long-axis (two-chamber view) planes, as shown in Figure 2.6. These planes are defined along a line that extends from the heart's apex to the center of the mitral valve (the heart's long axis) using images of the body taken from the axial plane. The short-axis plane is positioned at the level of the middle of the left ventricle

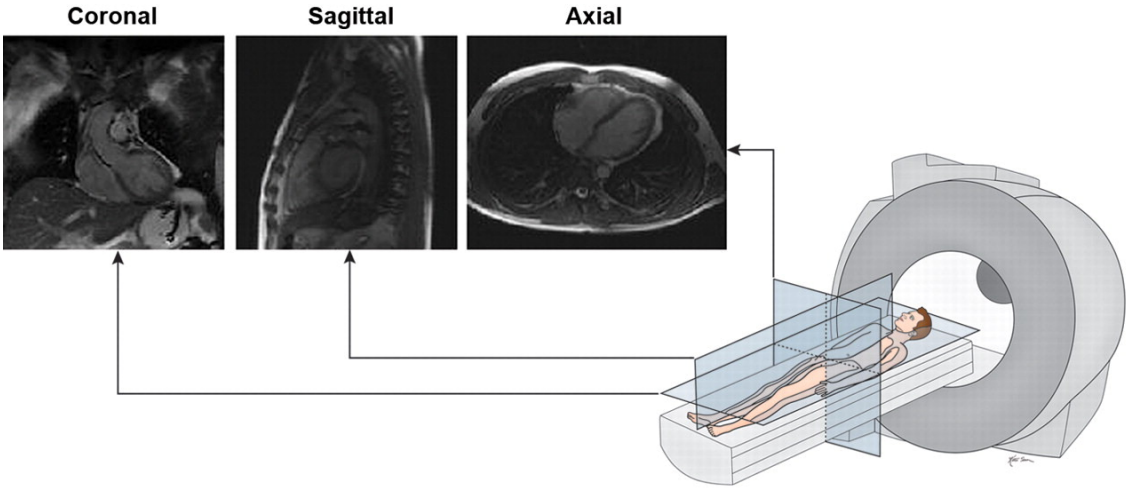

Figure 2.5: The anatomical planes of MRI: axial plane, coronal plane, and sagittal plane. Image adapted from [Ginat et al., 2011].

and is perpendicular to the true long axis of the heart. It allows for the measurement of ventricular volumes, ejection fraction, and wall thickness. Figure 2.7 shows short axis (SAX) images of the cardiac structures at basal, middle, and apical slices.

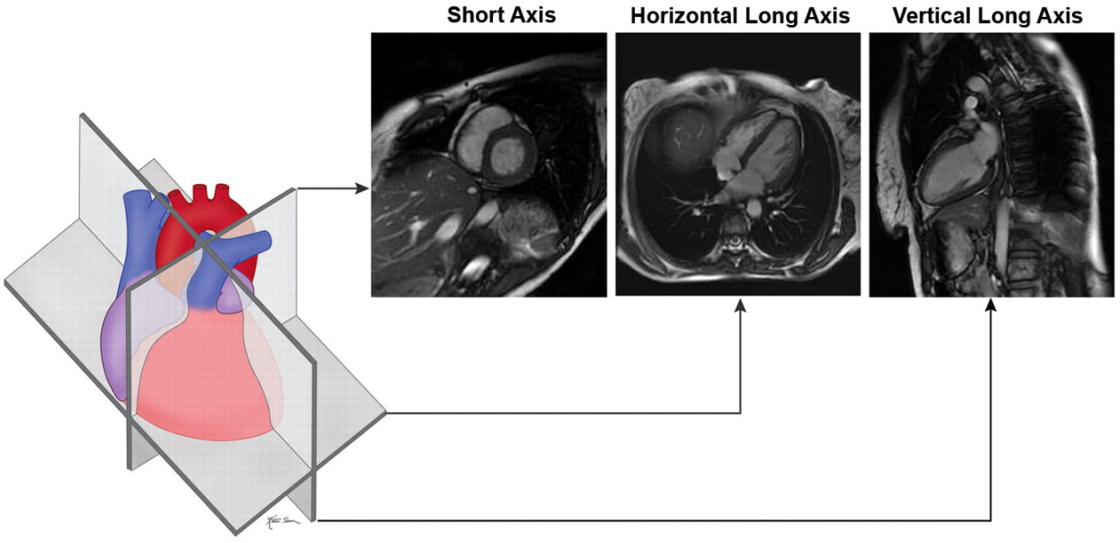

Figure 2.6: The orientation of main cardiac planes with respect to heart: short axis, horizontal long axis, and vertical long axis views. Image adapted from [Ginat et al., 2011].

The horizontal long axis is created by selecting a horizontal plane that is at a right angle to the short axis, while the vertical long axis is established by using a vertical plane that is at a right angle to the short axis plane. The horizontal long-axis plane is a view that shows both the left and right ventricles, the interventricular septum, the mitral and tricuspid valves, and the atria. It is useful for assessing the function of the ventricles and for detecting any abnormalities in the valves. The vertical long-axis plane is a view that

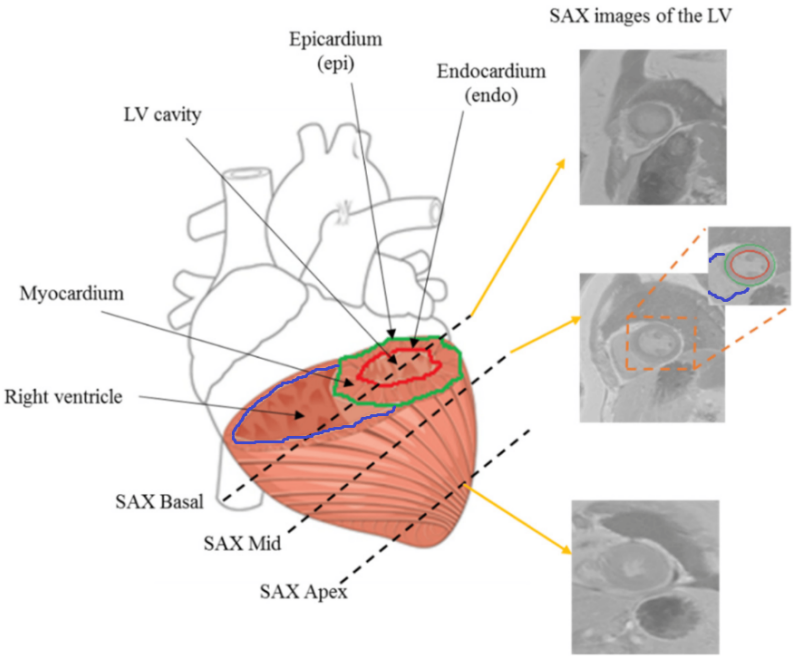

Figure 2.7: Short Axis (SAX) cardiac MRI. Image adapted from [Shaaf et al., 2022].

shows the left ventricle, the left atrium, and the aortic and mitral valves. It is useful for evaluating the left ventricular function and detecting any abnormalities in the mitral valve [Ginat et al., 2011].

The length of a CMR scan typically varies between 30 minutes to an hour, depending on the complexity of the referral questions being asked. During the scan, patients are required to hold their breath for most image acquisitions, which can be as short as a few seconds with modern fast scanners and adjusted based on the patient's ability. To prevent image distortion due to cardiac motion, vectorcardiogram (similar to ECG) triggering and gating are utilized, with cine images obtained during the entire cardiac cycle (prospective triggering or retrospective gating) and static images taken during diastole (prospective triggering). Most images are acquired over several cardiac cycles (segmented imaging) to mitigate the effects of arrhythmias and poor breath holding, which can degrade the quality of the image. However, in most cases, the use of arrhythmia rejection algorithms, non-breath holding (free breathing), or single-shot acquisition can still provide diagnostic quality information [Ripley et al., 2016].

There are many cardiac MRI modalities. In this thesis, we will focus only on three modalities: cine CMR, LGE CMR, and T1 mapping CMR. Figure 2.8 shows Cine (bSSFP), LGE, native T1, post-contrast T1, and ECV mapping images of one patient with a CVD.

2.4.1.1/ Cine CMR

Cine CMR is one type of cardiac MRI modality that is used to capture cardiac motion. It provides both a structural and functional assessment of the heart and is useful in the diagnosis and management of various cardiovascular diseases. Cine cardiac MRI can be performed using various imaging sequences, including balanced Steady-State Free Precession (bSSFP). bSSFP is a type of MRI sequence that uses a short repetition time and a low flip angle gradient echo sequence to achieve a steady state of magnetizations in which the signal from the MRI is maximized. This allows for a high signal-to-noise ratio and rapid image acquisition. It produces images with high spatial and temporal resolution, allowing for detailed images of the heart's anatomy and function. bSSFP protocols have different names among different MRI manufacturers, including TrueFISP (Siemens), FIESTA (General Electric), and balanced FFE (Philips) [AD Elster, 2023].

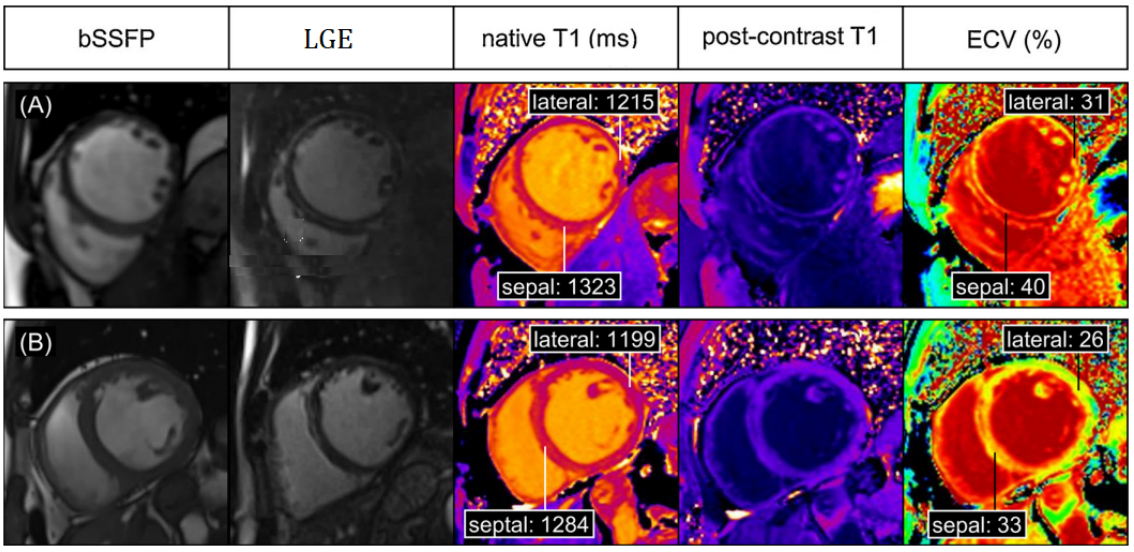

Figure 2.8: Cine (bSSFP), LGE, native T1, post-contrast T1 and ECV mapping for a patient with dilated cardiomyopathy. Septal myocardial native T1 and ECV values are elevated compared to the lateral. LGE, late gadolinium enhancement CMR. bSSFP, balanced steady-state free precession CMR. Image adapted from [Reiter et al., 2018].

2.4.1.2/ Late Gadolinium Enhancement (LGE) CMR

LGE CMR is a medical imaging technique that uses gadolinium-based contrast agents to visualize areas of scarring and fibrosis in the heart muscle. It works by injecting a contrast agent into the patient's bloodstream, which is taken up by the heart muscle tissue. Areas of healthy myocardium do not retain the contrast agent, while areas of damaged or scarred myocardium retain the contrast agent, leading to increased signal intensity on the CMR images.

To perform late gadolinium enhancement (LGE), medical professionals typically use 2D segmented inversion recovery gradient echo (GRE), or Phase-Sensitive Inversion-Recovery (PSIR) methods. In some cases, 3D sequences are preferred for patients who can hold their breath satisfactorily and if the signal-to-noise ratio (SNR) is sufficient. It is necessary to wait for at least 10 minutes after injecting the gadolinium-based contrast agent before proceeding. However, if lower gadolinium doses are used, and the blood pool signal falls below that of the late enhanced myocardium, the delay can be less than 10 minutes. Typically, images are obtained during a diastolic standstill. The inversion time (TI) is adjusted to nullify the signal from normal myocardium, making it appear black on the image. This is achieved by using a special inversion recovery pulse that flips the magnetization of the normal myocardium by 180 degrees, while leaving the scar tissue magnetization unaffected. The sequence is timed so that the normal myocardium magnetization is inverted and then allowed to recover to its steady-state (longitudinal magnetization) by the time the image is acquired, resulting in a nulled signal. In contrast, the scar tissue magnetization continues to generate a signal, resulting in a bright area on the final image [Kramer et al., 2020].

The resulting LGE CMR image shows the distribution and extent of scar tissue in the myocardium, which can be used to diagnose and monitor various cardiac conditions, such as myocardial infarction, myocarditis, and cardiomyopathies. The technique has a high spatial resolution that can accurately identify the location, extent, and distribution of myocardial scarring, which is important for guiding therapeutic decision-making [Vöhringer et al., 2007].

2.4.1.3/ T1 MAPPING CMR

T1 mapping CMR is a medical imaging technique that provides quantitative measurements of the T1 relaxation time of myocardial tissue. T1 relaxation time is a fundamental parameter that reflects the rate at which protons in tissue return to their equilibrium magnetization after an external magnetic field is applied.

LGE imaging is regarded as the gold standard for measuring focal myocardial fibrosis. However, in certain cardiomyopathies, fibrosis is frequently diffuse and not quantifiable on LGE images. Identifying interstitial fibrosis using conventional T1-weighted imaging has also been challenging due to the widespread and diffuse nature of structural changes. To address this issue, T1 mapping techniques have been developed to measure diffuse myocardial fibrosis and characterize the tissues [Arega et al., 2023a].T1 mapping refers to pixel-wise illustrations of absolute T1 relaxation times on a map [Haaf et al., 2016]. They have made it possible to accurately measure T1 in the heart and create color-coded T1 maps. These T1 maps use pixel values to represent T1 in each voxel, rather than

arbitrary signal intensity units, and can effectively illustrate even minor variations in T1 within the heart muscle to identify any tissue pathology [Taylor et al., 2016].

T1 mapping CMR refers to the process of acquiring a set of T1 weighted images with varying time intervals between preparation pulse and image acquisition. This generates different T1 weightings that allow for the estimation of the T1 value at each pixel. The estimation is done by fitting an exponential relaxation curve to the pixel intensities of the acquired images. The resulting T1 values can be displayed as a T1 map, where each pixel's intensity encodes an estimate of the T1 value [Moon et al., 2013, Fahmy et al., 2019a].

Modified Look-Locker Inversion Recovery (MOLLI) and Shortened Modified Look-Locker Inversion Recovery (ShMOLLI) are the most commonly used CMR imaging techniques for T1 mapping. MOLLI is a technique that acquires a series of images with different inversion times to measure T1 values of the myocardium, as shown in Figure 2.9. MOLLI sequences typically use three to five heartbeats to acquire the images, allowing for faster acquisition times than other T1 mapping techniques. ShMOLLI is a modification of the MOLLI technique that uses a shorter acquisition time by acquiring images during a single breath-hold. This allows for more accurate T1 measurements by reducing the effects of cardiac motion and respiratory motion. ShMOLLI has been shown to be more accurate and reproducible than MOLLI, especially at higher heart rates and in patients with arrhythmias [Taylor et al., 2016].

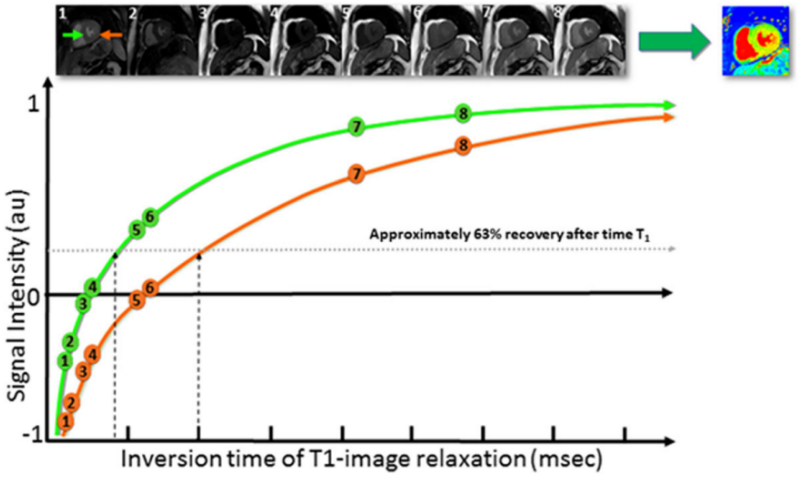

Figure 2.9: The MOLLI scheme is used for the T1 mapping in the heart. It involves two inversions to obtain eight images over 11 heartbeats. The area of myocardial infarction and elevated native T1 values is indicated by the orange arrow and relaxation curve, while the area of normal septal myocardium and normal native T1 values is indicated by the green arrow and relaxation curve. The images are sorted based on inversion times [Haaf et al., 2016]. Image adapted from [Haaf et al., 2016].

The use of native (non-contrast) T1 and extracellular volume (ECV) presents an opportunity to monitor significant biological changes in the myocardium. Without the need for

gadolinium-based contrast agents (GBCA), native T1 can provide insight into myocardial disease affecting both the myocyte and interstitium. ECV, which measures the size of the extracellular space and reflects interstitial disease, can categorize the myocardium into its cellular and interstitial components with the help of GBCA [Moon et al., 2013]. Some typical T1 mapping and ECV values of different CVDs are depicted in Figure 2.10.

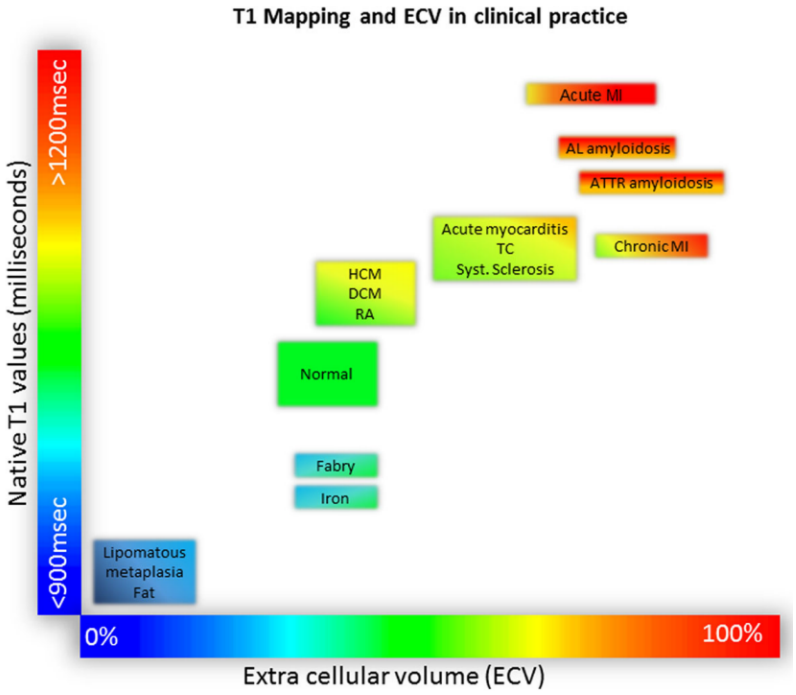

Figure 2.10: Tissue characterization using native T1 and extracellular volume (ECV) of different cardiovascular diseases (The native T1 and ECV are computed using 1.5 T scanners). Image adapted from [Haaf et al., 2016].

Native T1 measurements of the myocardium allow for the non-invasive detection of biologically important processes, which have the potential to improve disease diagnosis, severity assessment, and prognosis. Native T1 changes can detect pathologically significant processes related to excess water in edema, protein deposition, and other substances that alter T1, such as lipid or iron, without requiring the use of a contrast agent. Furthermore, native T1 techniques do not exclude patients with severe renal dysfunction. Changes in myocardial native T1 can indicate various cardiac and systemic diseases, such as acute coronary syndromes, infarction, myocarditis, cardiac amyloid, Anderson-Fabry disease, and siderosis. Evidence suggests that when used in combination with clinical scan protocols, native T1 mapping can reveal previously unknown pathologies, areas at risk in acute coronary syndromes, and preclinical disease or unsuspected cardiac involvement [Moon et al., 2013].

ECV is calculated by combining native and post-contrast (contrast-enhanced) T1 maps of blood and myocardium, Eq. 2.1. It represents the ratio of the extracellular space to

the total myocardial volume and reflects the proportion of myocardial tissue that is not comprised of cells (i.e., the interstitium). ECV serves as a biomarker for the extracellular space in the myocardium and can be used to monitor changes in interstitial disease. It is expressed as a percentage, with higher values indicating greater expansion of the extracellular space and a higher probability of myocardial fibrosis or other pathologies. ECV measurement is particularly useful for evaluating myocardial disease involving both the myocyte and interstitium without the need for a tissue biopsy or an invasive procedure [Lee et al., 2011, Moon et al., 2013].

$$ECV = (1 - hematocrit) \times \frac{(\frac{1}{T1_{myo\_post}} - \frac{1}{T1_{myo\_native}})}{(\frac{1}{T1_{blood\_post}} - \frac{1}{T1_{blood\_native}})} \tag{2.1}$$

Where $T1_{myo\_post}$ and $T1_{myo\_native}$ are the mean T1 value of myocardium computed from post-contrast and native T1 mapping images, respectively. $T1_{blood\_post}$ and $T1_{blood\_native}$ are the mean T1 value of the blood pool computed from post-contrast and native T1 mapping images, respectively.

2.4.2/ Cardiac Ultrasound (Echocardiography)

Echocardiography is a non-invasive medical imaging technique that uses high-frequency sound waves, or ultrasound, to create real-time images (echocardiogram) of the heart (Figure 2.11). This tool provides a quantification of the heart's size, shape, internal chamber size, pumping capacity, location, and extent of tissue damage, as well as an assessment of the valves. In addition, echocardiography can estimate other aspects of heart function, such as cardiac output, ejection fraction, and diastolic function. It is widely used in the diagnosis, management, and follow-up of patients with suspected or known heart diseases, including cardiomyopathies, myocardial infarction, heart failure, and valvular regurgitation or stenosis, among others [Otto, 2013, Cleve et al., 2018].

To perform an echocardiogram, a technician places a transducer on the chest, which emits sound waves that bounce off the heart and create images on a screen. These images reveal the size, shape, and movement of the heart's chambers and valves, as well as the blood flow through the heart. Overall, echocardiography is a valuable and painless tool for assessing heart health and guiding treatment decisions [Institute, 2019].

2.5/ Public datasets in cardiac MRI

Public datasets in cardiac MRI have been instrumental in advancing research and development in the field. These datasets provide a valuable resource for benchmarking al-

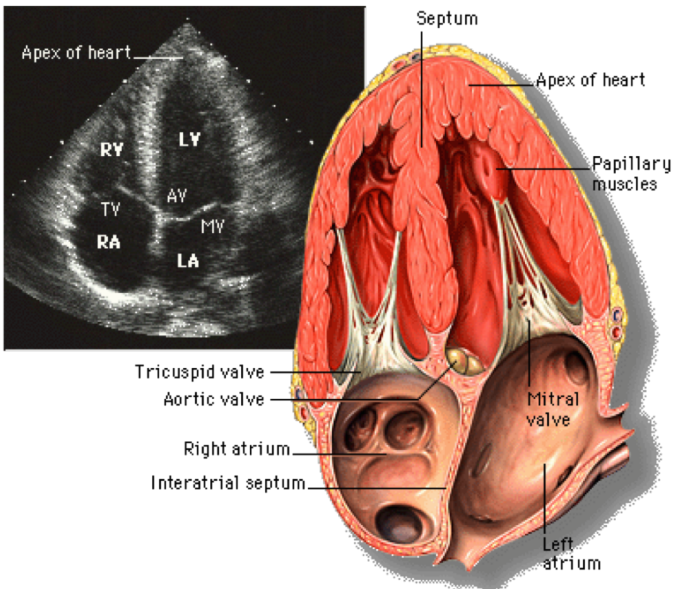

Figure 2.11: Apical 4-chamber view of the heart showing the left ventricle, left atrium, mitral valve, right ventricle, right atrium, tricuspid valve, and often aortic valve. Image adapted from [Institute, 2019].

gorithms, and comparing methodologies. Over the years, several challenges have been organized, resulting in the creation of diverse datasets that cover a wide range of cardiac pathologies and imaging modalities. Table 2.1 presents a comprehensive summary of the recent challenges and public datasets in cardiac MRI analysis [Li et al., 2022a].

Starting from the early SCD (Sunnybrook Cardiac Data) challenge dataset [Radau et al., 2009] in 2009, these datasets have evolved to encompass a wide range of cardiac pathologies. Challenges such as LVSC (Left Ventricle Segmentation Challenge) [Suinesiaputra et al., 2012], RVSC (Right Ventricle Segmentation Challenge) [Petitjean et al., 2015], and LiVScar (Left Ventricle Infarct Segmentation Challenge) [Karim et al., 2016] focused on segmenting specific cardiac structures and scar tissue from cine and LGE CMR.

Other challenges, like LASC (Left Atrium Segmentation Challenge) [Tobon-Gomez et al., 2015], introduced multi-modality datasets combining CT and cine CMR to address segmentation tasks in cases of the left atrium (LA) cavity. cDEMRIS (Cardiac Delayed Enhancement Segmentation Challenge) [Karim et al., 2013] used single-modality (SM) and multi-center/multi-vendor datasets for segmenting scars in the left atrium (LA) with atrial fibrillation (AF) from LGE images. SLAWT (Segmentation of LA Wall Thickness) [Karim et al., 2018] further expanded the scope by incorporating CT and 3D Flash CMR data to segment LA wall thickness in various pathologies.

The ACDC (Automated Cardiac Diagnosis Challenge) [Bernard et al., 2018] provided a comprehensive dataset for segmenting the left ventricle (LV), right ventricle (RV), and

myocardium (Myo), while also classifying pathologies such as MI, dilated cardiomyopathy (DCM), HCM, and abnormal right ventricle (ARV). Similarly, MM-WHS (Multi-Modality Whole Heart Segmentation) [Zhuang et al., 2019] aimed to segment multiple cardiac structures in normal and pathological cases, combining CT and cine CMR data.

These challenges and datasets have continued to evolve, covering more complex tasks and diverse cardiac pathologies. Challenges like Atrial Segmentation Challenge [Xiong et al., 2021], MS-CMRSeg (Multi-sequence Cardiac MR Segmentation Challenge) [Zhuang et al., 2022], and M&Ms (Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge) [Campello et al., 2021] have focused on segmenting specific cardiac structures or addressing multi-disease scenarios, further expanding the applicability of the datasets.

Recent additions to the CMR datasets include MyOps2020 (Myocardial pathology Segmentation combining multi-sequence CMR) [Li et al., 2022a], EMIDEC (Automatic Eval of Myocardial Infarction from DE Cardiac MRI) [Lalande et al., 2020], M&Ms2 (Multi-Disease, Multi-View & Multi-Center RV Segmentation) [Martín-Isla et al., 2023], LAScarQS (Left Atrial and Scar Quantification & Segmentation Challenge) [1], CMRxMotion (Extreme Cardiac MRI Analysis Challenge under Respiratory Motion) [2] and MYOSAIQ (Myocardial Segmentation with Automated Infarct Quantification) [3]. These challenges have increased the diversity of CMR datasets.

Among these public datasets, as part of the thesis, we utilized several cardiac MRI datasets, including ACDC, EMIDEC, LAScarQS, M&Ms, M&Ms2, and CMRxMotion, to evaluate our proposed methods. Additionally, we also used a proprietary T1 mapping CMR dataset collected from Dijon Hospital and other clinical centers in France. This proprietary dataset was obtained as part of the French National Research Agency (ANR) project (reference ANR-19-CE45-0001-01-ACCECIT), which served as the funding source for the Ph.D.

2.6/ CONCLUSION

In conclusion, this chapter has provided essential background on cardiac anatomy, major cardiovascular diseases, diagnostic approaches, and cardiac imaging modalities relevant to the clinical focus of this thesis. Key concepts covered include normal cardiac structure and function, common conditions like myocardial infarction and cardiomyopathies, standard diagnostic tests, and cardiac imaging techniques particularly magnetic resonance imaging and its various applications. Public datasets were also highlighted. This overview

[1] https://zmic.fudan.edu.cn/lascarqs22
[2] http://cmr.miccai.cloud/
[3] https://www.creatis.insa-lyon.fr/Challenge/myosaiq/

of the clinical context and core medical imaging domains establishes the necessary foundation for the technical contributions to advancing cardiac MRI analysis through deep learning and uncertainty estimation that will be presented in the subsequent chapters.

Table 2.1: Public datasets and existing challenges in cardiac MRI. SM: single modality; MM: multi-modality for the same patient; MM†: multi-modality for different patients; Multi-C: Multi-center; Multi-V: Multi-vendor; RV: Right Ventricle; LV: Left Ventricle; Myo: Myocardium; Seg: Segmentation; Classif: Classification; Eval: Evaluation; Patho: Pathology; MI: Myocardial Infarction; LA: left atrium; AF: Atrial fibrillation; DCM: Dilated cardiomyopathy; HCM: Hypertrophic cardiomyopathy; LVNC: left ventricle non-compaction; TriReg: Tricuspid Regurgitation; TF: Tetralogy of Fallot; ARV:Abnormal Right Ventricle; DRV:Dilated Right Ventricle; IAC: Inter-atrial Communication

Challenge	**Year**	**Source**	**Data**	**Objective**	**Pathologies**
Sunnybrook Cardiac Data (SCD)	2009	SM	45 cine CMR	Seg LV, Myo	Normal, HCM, MI
LVSC (LV Seg Challenge)	2011	SM	200 cine CMR	Seg LV, Myo	MI
RVSC (RV Seg Challenge)	2012	SM	48 cine CMR	Seg RV	Different patho
LiVScar (LV Infarct Seg Challenge	2012	SM Multi-C/V	30 LGE (15 humans + 15 pigs)	Seg LV scar	MI
cDEMRIS (Cardiac Delayed Enhancement Seg Challenge)	2012	SM Multi-C/V	60 LGE LA (30 pre + 30 post catheter ablation)	Seg scar LA	AF
LASC (Left Atrium Seg Challenge)	2013	MM†	30 CT + 30 cine CMR	Seg LA	AF
SLAWT (Seg of LA Wall for Thickness)	2016	MM†	10 CT + 10 3D Flash CMR	Seg LA wall	CT : Different patho MRI : normal
ACDC (Automated Cardiac Diagnosis Challenge)	2017	SM	150 cine CMR	Seg LV, RV, Myo Classif pathologies	Normal, MI, DCM, HCM, ARV
MM-WHS (Multi-Modality Whole Heart Seg)	2017	MM† Multi-V	60 CT + 60 cine CMR	Seg whole heart 7 structures	Normal, AF, MI, DCM, TF, TriReg, others
Atrial Seg Challenge	2018	SM	154 LGE	Seg LA	AF
MS-CMRSeg (Multi-sequence Cardiac MR Seg Challenge)	2019	MM	45 cine CMR + LGE +T2	Seg LV, RV, Myo	Cardiomyopathies
M&Ms (Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Seg Challenge)	2020	SM Multi-C/V	375 cine CMR	Seg LV, RV, Myo	Normal, HCM, DCM, ARV, LVNC, myocardites, others
MyOps2020 (Myocardial pathology seg combining multi-sequence CMR)	2020	MM	45 CMR cine, LGE, T2	Seg LV, RV, Myo scar, edema	MI
EMIDEC (Automatic Eval of Myocardial Infarction from DE Cardiac MRI)	2020	SM	150 LGE + clinical data	Seg Myo, scar, no-reflow Classif normal/patho	MI, Normal
M& Ms2 (Multi-Disease, Multi-View & Multi-Center RV Seg)	2021	SM Multi-C/V	360 cine CMR	Seg RV (+ LV and Myo)	Normal, HCM, DCM, DRV, TF, TriReg, IAC
LAScarQS (Left Atrial and Scar Quantification & Seg Challenge)	2022	SM Multi-C	194 LGE (post-ablation)	Seg LA cavity, LA scar	AF
CMRxMotion (Extreme Cardiac MRI Analysis Challenge under Respiratory Motion)	2022	SM	cine CMR 45 volunteers, 4 types of breathing	Seg LV, Myo, RV Eval image quality	Not Known (volunteers)
MYOSAIQ (Myocardial Seg with Automated Infarct Quantification)	2023	SM	467 LGE	Seg Myo,scar	MI

3

Background: Fundamentals of Deep Learning and the State-of-the-Art in Cardiac MRI Analysis

This chapter provides a comprehensive exploration of the fundamental concepts and state-of-the-art techniques in the field of cardiac MR analysis. It begins by delving into the realm of machine learning and deep learning, introducing the principles, methodologies, and applications within this domain. It covers a range of machine learning algorithms like Support-Vector Machine (SVM), random forests, and artificial neural networks. Additionally, deep learning models such as convolutional neural networks (CNNs) and vision transformers are explored in detail. Furthermore, the chapter emphasizes training neural networks by explaining vital aspects such as loss functions, regularization techniques, and evaluation metrics. These components are crucial for optimizing the performance of neural networks in various applications. Lastly, the chapter explores uncertainty estimation methods and their relevance in medical image analysis. It sheds light on techniques such as variational inference, Monte Carlo dropout, and deep ensemble methods. The application of these methods in medical image analysis is discussed, highlighting their significance in quantifying uncertainty and improving decision-making processes. The chapter provides a comprehensive foundation of knowledge necessary for the subsequent chapters of the thesis.

3.1/ Machine Learning and Deep Learning

3.1.1/ Introduction

Artificial intelligence (AI) is the development of software or machines that can perform tasks that would normally require human intelligence to complete. This includes tasks such as recognizing speech, understanding natural language, making decisions, and recognizing patterns in data [Ed Burns, 2023]. Machine learning (ML) is a subset of AI that focuses on designing algorithms that can learn from data and make predictions or decisions based on that data [Kotsiantis et al., 2006]. Deep learning is a subset of machine learning that uses neural networks with multiple layers to learn and extract features from data.

Depending on the type of data available and the task at hand, machine learning algorithms can be supervised, unsupervised, or semi-supervised. Supervised learning is the most common type of machine learning where the algorithm is trained using labeled data. In supervised learning, the algorithm tries to learn a mapping function that maps the input data to the correct output based on the labeled examples provided during training. This mapping function can then be used to predict the output for new data points [Russell et al., 1995]. For example, an algorithm might be trained to recognize handwritten digits by being shown labeled examples of digits from 0-9. Unsupervised learning, on the other hand, is a type of machine learning where the algorithm is trained using unlabeled data. In unsupervised learning, the goal is to find patterns or structures in the data without any specific guidance or labels. For example, an unsupervised learning algorithm might try to group similar images together without being told what the images represent [Janiesch et al., 2021]. Semi-supervised learning is a type of machine learning that combines both labeled and unlabeled data to train an algorithm. The goal of semi-supervised learning is to use the labeled data to guide the learning process while also leveraging the unlabeled data to improve the accuracy and robustness of the model [van Engelen et al., 2019].

In machine learning, experts with domain knowledge of the problem typically handcraft features, which can be a time-consuming and labor-intensive process, and identifying the most crucial features for a given problem is also challenging. Conversely, deep learning methods can automatically learn features from data using neural networks that mimic human brain functioning. Neural networks can recognize patterns in data that would be difficult or impossible for humans to detect. This automatic feature learning capability is advantageous for problems with large amounts of data and where hand-crafting features is very difficult [Hosny et al., 2018]. Deep learning methods have achieved state-of-the-art results in computer vision and medical image analysis tasks.

3.1.2/ Machine Learning

This section discusses some of the machine learning methods that have been commonly employed in computer vision and medical image analysis. These include the support vector machine, random forests, and artificial neural networks.

3.1.2.1/ Support vector machine (SVM)

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for classification and regression tasks. It is a flexible algorithm that can handle both linear and non-linear data. SVM works by finding the best possible hyperplane that separates the data into different classes [Boser et al., 1992, Cortes et al., 1995]. Hyperplane is a line or plane in n-dimensional space that divides the data into two regions. The hyperplane is chosen in such a way that it maximizes the margin between the two classes. The margin is defined as the distance between the hyperplane and the nearest data points from each class. The SVM algorithm tries to find the hyperplane that has the largest margin, which is considered to be the most robust and accurate solution. To help illustrate this concept, Figure 3.1 shows a diagram of an SVM in 2D space. The diagram depicts the hyperplane as a straight line that divides the data points into two classes, represented by blue and green dots. Support vectors are the data points or vectors that are the closest to the hyperplane. They influence the position and orientation of the hyperplane [JavaTpoint, 2023]. The margin is shown as the space between the dotted lines that run parallel to the hyperplane and is maximized by positioning the hyperplane equidistant from the closest support vectors. In regression tasks, the SVM algorithm tries to find a hyperplane that minimizes the mean squared error between the predicted output and the actual output.

In the case of linearly separable data, the SVM algorithm finds the hyperplane that separates the data into two classes with the maximum margin. This hyperplane is known as the optimal separating hyperplane [JavaTpoint, 2023]. However, in some cases, the data points may not be linearly separable, so the SVM uses a technique called kernel trick to map the input data into a higher-dimensional space where the data points can be separated by a hyperplane [Boser et al., 1992, Cortes et al., 1995].

The kernel trick is a mathematical technique that allows the SVM algorithm to transform non-linearly separable data into a linearly separable form. It does this by applying a non-linear function to the input data, which transforms it into a higher-dimensional space where it can be linearly separated. The kernel function calculates the similarity or distance between data points in the original feature space and maps them to a new, high-dimensional feature space. The most commonly used kernel functions are polynomial, Gaussian Radial Basis Function (RBF), and sigmoid kernels. Once the data is mapped

to a higher-dimensional space, the SVM algorithm finds the optimal hyperplane that separates the classes with the maximum margin. This hyperplane is then used to predict the class of new data points [Jakkula, 2006].

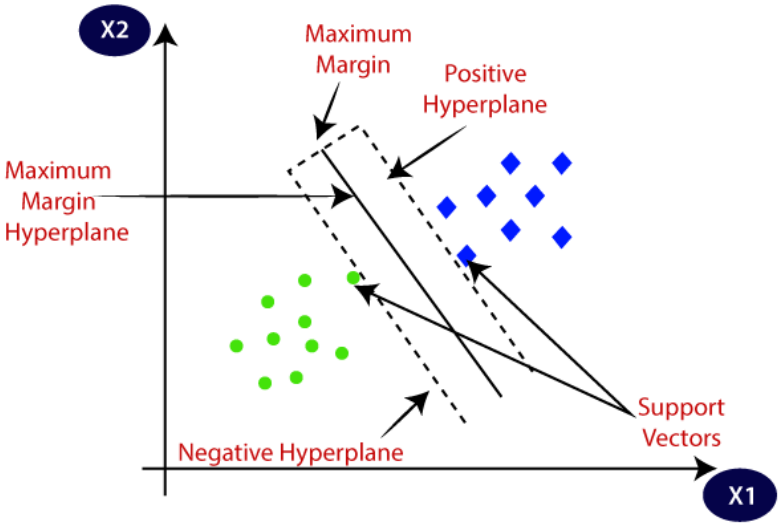

Figure 3.1: Support vector machine (SVM). Image adapted from [JavaTpoint, 2023].

3.1.2.2/ Random Forests

Random Forest is an ensemble machine learning algorithm that is used for both classification and regression problems. It utilizes multiple decision trees to make predictions, resulting in better accuracy than using a single decision tree. A decision tree is a non-parametric supervised learning algorithm that uses a hierarchical tree-like structure to make decisions based on input features. The structure of a decision tree consists of a root node, branches, internal nodes, and leaf nodes. The internal nodes represent the input features, the branches represent the decision rules, and the leaf nodes represent the final decision of the algorithm. The decision tree algorithm recursively splits the training data into smaller subsets based on the selected feature that best separates the data into the target classes. The selection of the best attribute (feature) is based on a metric such as entropy or Gini impurity. These metrics measure the level of impurity or randomness in the subsets, and the goal is to find the feature that maximizes the information gain or the reduction in impurity after the split [Bshyamanth, 2023].

Random forest uses an ensemble of individual decision trees to make predictions. In the Random Forest algorithm, each decision tree is trained on a random subset of the original training data and a random subset of the original features. This process is known as bootstrapping and feature bagging, respectively. The Random Forest algorithm uses both bootstrapping and feature bagging to create an uncorrelated forest of decision trees. Feature bagging generates a random subset of features for each decision tree, ensuring that

each tree has a different set of features. This technique reduces overfitting and variance in the model by increasing the diversity of the decision trees in the forest [IBM, 2023]. By using a random subset of features for each decision tree, the Random Forest algorithm can capture the most important features for making accurate predictions while reducing the impact of noisy or irrelevant features. This approach also ensures that the decision trees are not highly correlated, enabling the model to generalize well to new data.

When making a prediction with a random forest, each decision tree in the ensemble makes a prediction, and the final prediction is obtained by averaging or taking the mode of all the individual predictions for a classification task, as shown in Figure 3.2 [Dimitriadis et al., 2018]. For regression problems, the predicted value is the average of the predictions from the decision trees. Random Forests can handle noisy data, missing values, and high-dimensional data, making it a versatile algorithm in machine learning.

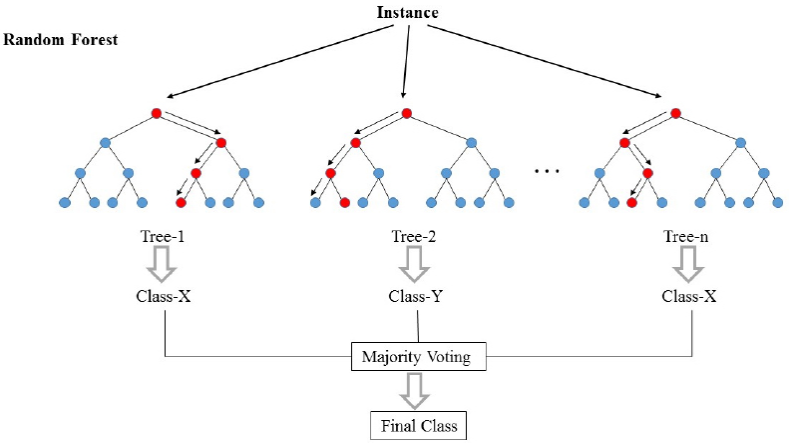

Figure 3.2: An informative illustration showcasing the structure of a random forest, which consists of multiple decision trees. Image adapted from [Dimitriadis et al., 2018].

3.1.2.3/ ARTIFICIAL NEURAL NETWORKS

An Artificial Neural Network (ANN) is a computational model inspired by the structure and function of the human brain. It consists of interconnected processing units called neurons, which work together to perform a specific task. Each neuron receives input signals from other neurons and processes the information using an activation function, and produces an output signal that is transmitted to other neurons in the network.

A single-layer perceptron is a type of artificial neural network that consists of a single layer of neurons, as shown in Figure 3.3. Each neuron receives input from the data and computes a weighted sum of those inputs. This weighted sum is then passed through an activation function, which determines whether the neuron should fire or not (i.e., whether it should output a 1 or a 0). The purpose of the activation function is to introduce non-linearity into the output of the perceptron, which allows it to model more complex rela-

tionships between the input and output data. The final output of the neuron $(\hat{y})$ can be computed as:

$$\hat{y} = g(\sum_{i=1}^{m} w_i x_i + b) = g(\mathbf{w}^T \mathbf{x} + b), \tag{3.1}$$

where w, b, x, m and $g()$ represent connection weights, neuron bias, input values, number of inputs to the perceptron, and activation function respectively. The bias is an additional input that is added to the weighted sum of inputs for each neuron. The bias allows the neuron to adjust its output independently of its inputs, and it plays a crucial role in the learning process of the perceptron. There are several different types of activation functions that can be used in a single-layer perceptron, including the step function, the sigmoid function, the ReLU function, and the tanh function, among others. While single-

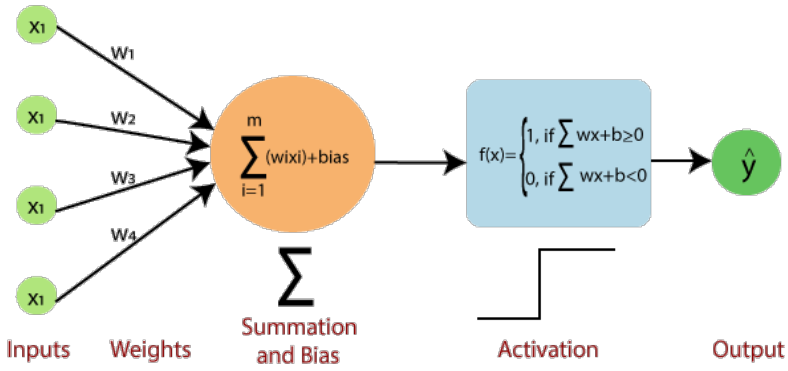

Figure 3.3: A Visual Representation of Single layer perceptron. Image adapted from [JavaTpoint, 2022].

layer perceptrons are effective at classifying linearly separable data, they have limitations when it comes to handling more complex, non-linear data. To address this, multi-layer perceptrons, or MLPs, were developed to allow for more sophisticated modeling of complex relationships between input and output data. A multi-layer perceptron (MLP) is a type of artificial neural network that consists of multiple layers of neurons, including an input layer, one or more hidden layers, and an output layer, as presented in Figure 3.4. Each neuron in the MLP receives input from the neurons in the previous layer, computes a weighted sum of those inputs, and then applies an activation function to produce its output. The output of each neuron in the previous layer serves as input to the next layer, and this process is repeated until the output layer produces the final output of the network.

Training an ANN involves initializing the weights, forwarding the input through the network, computing the loss, backpropagating the error to adjust the weights and biases, and repeating these steps for multiple steps until the loss function converges to a minimum value.

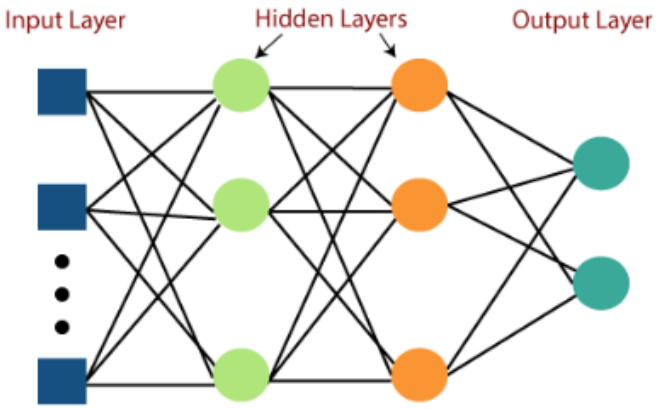

Figure 3.4: A Multi-layer Perceptron, comprising an input layer, two hidden layers, and an output layer. Image adapted from [JavaTpoint, 2022].

3.1.3/ Deep Learning

Deep Learning (DL) is a subfield of machine learning that involves training artificial neural networks with multiple layers to learn complex representations of data. Unlike traditional machine learning models, which rely on hand-crafted features, deep learning models learn features automatically from raw data. This makes deep learning particularly well-suited for tasks like image recognition, speech recognition, and natural language processing, where the input data is high-dimensional and complex [Hosny et al., 2018].

In recent years, deep learning has emerged as the leading technique for many image analysis tasks, delivering state-of-the-art performance on image classification, object detection, semantic segmentation, and image captioning. Convolutional neural networks (CNNs) are the most widely used deep neural network architecture for image analysis [Alzubaidi et al., 2021b]. They have demonstrated impressive results across various computer vision and medical image analysis tasks. More recently, vision transformer-based models have matched or exceeded the performance of CNNs on some computer vision and medical image analysis tasks [Willemink et al., 2022, Khan et al., 2023]. In the following section, we will explore CNN and vision transformer-based approaches for medical image analysis.

3.1.3.1/ Convolutional Neural Networks (CNNs)

While MLPs have been successful in many applications, they have limitations when it comes to image processing tasks. MLPs treat each pixel in the image as a separate input feature, which can lead to a large number of parameters and slow training times. Additionally, MLPs do not take into account the spatial relationships between pixels in an image, while CNN architectures make the explicit assumption that the inputs are images.

This allows to encode of certain properties into the architecture and makes the processing more efficient by reducing the number of parameters in the network [Chen et al., 2021b].

CNN architectures are made up of multiple building blocks, such as convolutional layers, pooling layers, and fully-connected layers, as depicted in Figure 3.5.

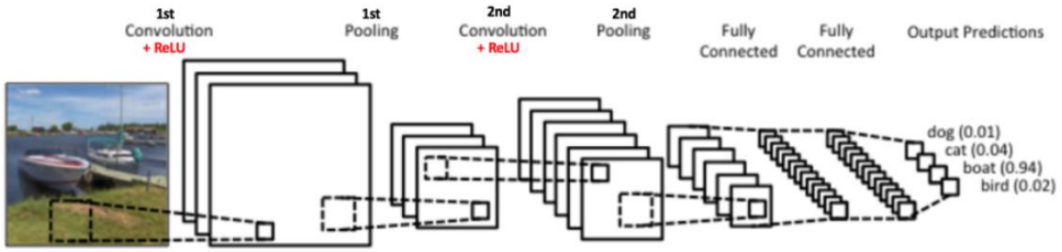

Figure 3.5: CNN architecture. Image adapted from [Jiang, 2019].

A convolutional layer is a fundamental building block of a convolutional neural network (CNN). It is responsible for extracting important features from the input image by performing a convolution operation between the input image and a set of learnable filters (kernels). A convolution operation involves sliding a small filter over the input image and computing the dot product between the filter and the corresponding region of the input image (local receptive field), as shown in Figure 3.6. This produces a single value, which is then placed in the output feature map at the corresponding location. The filter is then moved to the next location and the process is repeated until the entire input image has been covered. The size of the filter and the amount by which it is shifted over the input image are hyperparameters that are typically chosen based on the characteristics of the input image and the desired properties of the output feature map [Karpathy, 2016, Chen et al., 2021b].

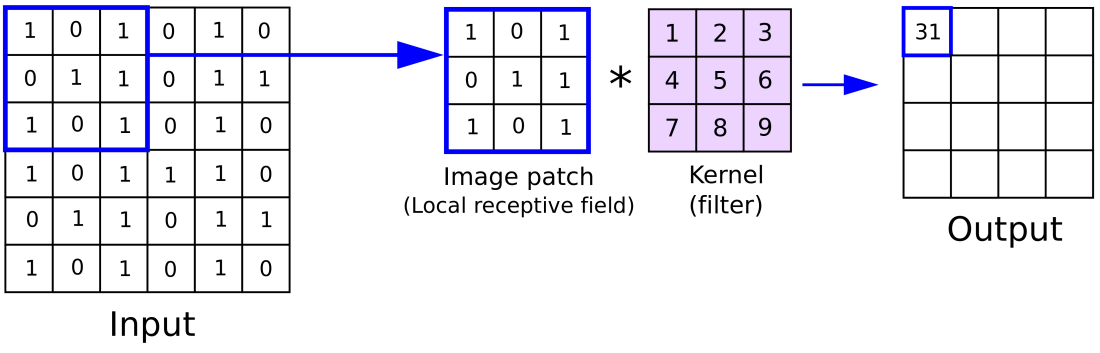

Figure 3.6: Convolution operation in CNN. Image adapted from [Reynolds, 2019].

A convolutional layer may have multiple filters, each of which learns to detect a different feature. The output feature map of the convolutional layer is obtained by stacking the output maps produced by each filter. The output feature map has a smaller spatial dimension than the input image, due to the loss of pixels around the edges of the image

during the convolution operation. To prevent the output feature map from becoming too small, it is common to use padding around the edges of the input image before performing the convolution operation. This ensures that the output feature map has the same spatial dimensions as the input image [Karpathy, 2016].

Some of the core benefits of convolutional layers are weight sharing, local connectivity, translation invariance, and hierarchical representation learning. Weight sharing involves using the same set of weights (or filters) across the entire input image, which reduces the number of parameters needed and improves model efficiency. Furthermore, convolutional layers only process a small portion of the input image at a time, allowing them to capture the local structure and patterns in the image. Their ability to be translation invariant ensures that the same features can be detected regardless of their position in the input image, which is particularly valuable for image recognition tasks. In addition, convolutional layers can learn hierarchical representations of input data, with each layer capturing increasingly more complex features, which leads to better representation of the input data and improved performance [Alzubaidi et al., 2021a].

In CNNs, **a pooling layer** is a type of layer that is typically added after convolutional layers to reduce the spatial size of the input feature maps while retaining important information. This is achieved by dividing the input image into non-overlapping regions and computing a summary statistic, such as the maximum or average value, for each region [O'Shea et al., 2015]. There are different types of pooling layers commonly used in CNNs, such as max pooling, average pooling, and global average pooling (GAP), as depicted in Figure 3.7. Max pooling selects the maximum activation in each region of the feature map, whereas average pooling computes the average activation in each region of the feature map. Global average pooling reduces the spatial dimensions of feature maps to a single value by taking the average value of each feature map across all spatial locations [Alzubaidi et al., 2021a]. It is often used in the final layers of CNNs to produce a fixed-length output. Pooling layers have several benefits in CNNs, such as reducing the number of parameters and computation needed while preserving important features. They also help to increase the model's robustness to small changes in the input by providing translation invariance [Goodfellow et al., 2016].

A fully connected layer, also known as a dense layer, is a type of layer commonly used in CNNs to perform classification or regression tasks. Fully connected layers are typically placed at the end of a CNN architecture and receive input from the output of the preceding convolutional and pooling layers. In a fully connected layer, each neuron is connected to every neuron in the previous layer. This means that the output of each neuron in the previous layer is fed into each neuron in the current layer. The purpose of fully connected layers in CNNs is to perform high-level reasoning or decision-making based on the features learned by the convolutional and pooling layers [Gu et al., 2018]. These

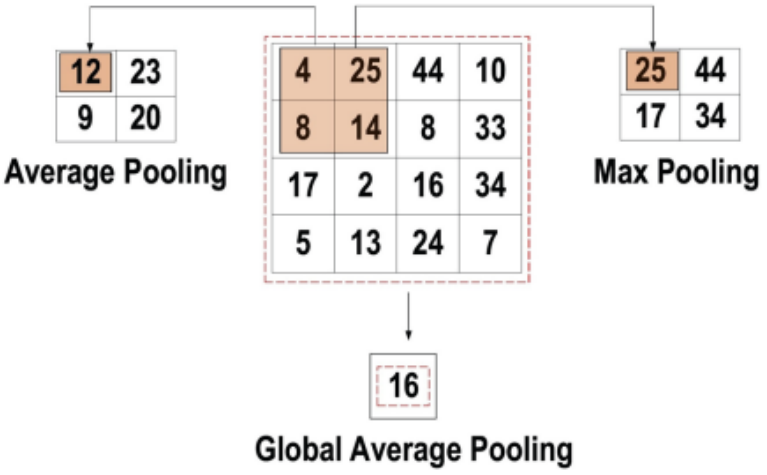

Figure 3.7: Different types of pooling layers. Image adapted from [Alzubaidi et al., 2021a].

layers take the output of the preceding layers and produce a vector of class probabilities or regression values [Yamashita et al., 2018].

At the end of convolutional and fully connected layers in CNN architecture, **activation functions** are employed to introduce non-linearity to the network, allowing it to learn complex features and make accurate predictions. The non-linear transformation is necessary to capture the complex relationships between the input and output variables and to model complex patterns in the data [Montesinos López et al., 2022]. There are several activation functions that are commonly used in CNNs, including Sigmoid, ReLU (Rectified Linear Unit), Leaky ReLU, and ELU (Exponential Linear Units), as shown in Figure 3.8.

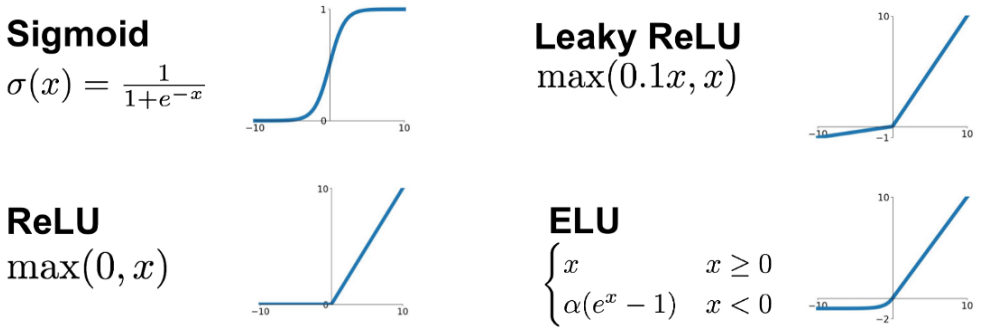

Figure 3.8: Different types of non-linear activation functions. Image adapted from [Jayawardana et al., 2021].

ReLU is one of the most widely used activation functions in CNNs. It is defined as $f(x) = max(0, x)$, which means that it returns 0 for all negative inputs and returns the input value for all positive inputs. ReLU is computationally efficient and has demonstrated excellent performance in practice.

Leaky ReLU is a modification of the ReLU function that addresses some of its limitations. It is defined as $f(x) = max(ax, x)$, where a is a small positive constant. Leaky ReLU allows

for a small non-zero gradient for negative inputs, which can help to prevent the ”dying ReLU” problem where some neurons permanently output 0 [Alzubaidi et al., 2021a].

ELU is also a variant of ReLU that modifies the slope of the negative part of the function by using a log curve to define the negative values, unlike the leaky ReLU. Mathematically, it is defined as $f(x) = x$ if $x >= 0$ and $f(x) = a(exp(x) - 1)$ if $x < 0$, where a is a small positive constant.

The sigmoid function is commonly used in the output layer of binary classification problems. It is defined as $f(x) = 1/(1 + exp(-x))$, which returns a value between 0 and 1. However, the sigmoid function suffers from the vanishing gradient problem.

CNN architectures have undergone significant evolution and improvement since their inception in the 1990s, with various variants emerging to address different challenges in image processing tasks. LeNet-5 was one of the first CNN architectures, developed by [LeCun et al., 2015] in the 1990s. It consists of several convolutional and pooling layers, followed by two fully connected layers. LeNet-5 was primarily used for handwritten digit recognition.

AlexNet, proposed by Krizhevsky et al. in 2012 [Krizhevsky et al., 2012], is a more complex CNN architecture that won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. It used GPUs (Graphics Processing Units) and introduced the concept of rectified linear units (ReLUs), dropout, and local response normalization (LRN) layers, which improved the network's performance. AlexNet consists of five convolutional layers and three fully connected layers.

VGGNet [Simonyan et al., 2014] is another CNN architecture that is deeper than AlexNet, with 16 or 19 layers. VGGNet's architecture is characterized by its use of small 3x3 convolutional filters, which are stacked on top of each other to form deeper representations. It was a runner-up in the ILSVRC challenge in 2014, and it showed that the depth of the network plays an important role in achieving good performance. VGGNet also introduced the use of batch normalization, which helps to improve the stability and convergence of the network.

The Google Inception Network (GoogLeNet) [Szegedy et al., 2015] is a variant of CNN that utilizes multiple filter sizes and aspect ratios in a single layer. It is characterized by its use of ”inception modules,” which are parallel branches of convolutional layers with different filter sizes (Figure 3.9). This allows it to capture features at multiple scales. Inception achieves high accuracy while minimizing the number of parameters by using 1x1 convolutions to reduce the dimensionality of feature maps before applying larger convolutions, as shown in Figure 3.9. It achieved state-of-the-art results on the ILSVRC 2014.

With the increasing depth of CNNs, a new challenge arises, vanishing gradients. Infor-

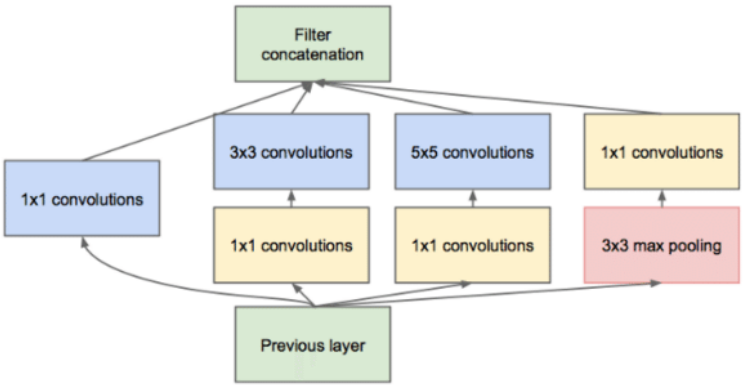

Figure 3.9: Inception module with dimension reductions in the Google Inception Network. Image adapted from [Szegedy et al., 2015].

mation about the input or gradient can diminish and disappear as it travels through many layers, making it difficult to reach the end (or start) of the network [Huang et al., 2017]. To address this issue, He et al. [He et al., 2016] proposed ResNet, a CNN architecture in 2016, that tackles the problem of vanishing gradients in deep networks. The key innovation in ResNet is the use of residual learning blocks where the input to each block is added to the output via skip connections, as depicted in Figure 3.10. This allows gradients to flow through identity mappings unimpeded. By retaining gradient magnitude throughout the network, ResNet's residual design enables extremely deep networks (e.g., ResNet-50, ResNet-101, ResNet-152) while still achieving excellent performance.

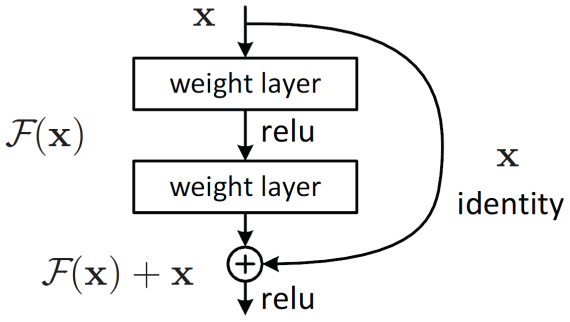

Figure 3.10: Residual block in ResNet architecture. Image adapted from [He et al., 2016].

Another CNN architecture that aims to enhance the feature extraction process and alleviates the vanishing-gradient problem is DenseNet, which was proposed by Huang et al. in 2017 [Huang et al., 2017]. DenseNet employs "dense blocks" that connect all layers to each other within a block, as shown in Figure 3.11. In each dense block, every layer obtains input from all preceding layers and passes its output to all subsequent layers, as shown in Figure 3.11. This results in a highly interconnected network where each layer can access the collective knowledge of all preceding layers. To accomplish this dense

connectivity, DenseNet concatenates the feature maps of all preceding layers. This enables feature reuse, mitigates the vanishing-gradient issue, and reduces the number of parameters in the network, leading to better generalization. Variants of DenseNet include DenseNet-121, DenseNet-169, and DenseNet-201 [Huang et al., 2017], which are deeper than ResNet.

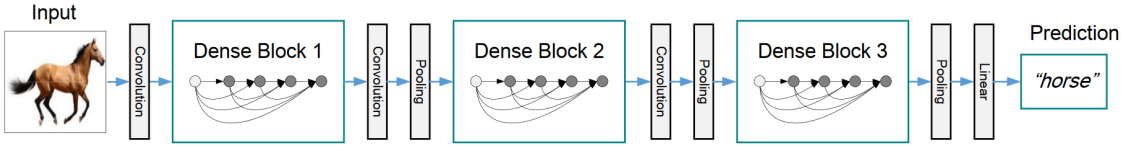

Figure 3.11: A deep DenseNet with three dense blocks, where transition layers between each block use convolution and pooling to modify feature-map sizes. Image adapted from [Huang et al., 2017].

Following DenseNet, several noteworthy CNN architectures were introduced, including ResNeXt [Xie et al., 2017b], SENet [Hu et al., 2017], and EfficientNet [Tan et al., 2019]. ResNeXt [Xie et al., 2017b] builds upon ResNet by incorporating ”split-transform-merge” blocks, which enhance cardinality and improve representation. On the other hand, SENet [Hu et al., 2017] employs ”squeeze-and-excitation” modules to enhance channel interdependencies and recalibrate features, whereas, EfficientNet [Tan et al., 2019] revolutionized CNN scaling by introducing compound coefficient scaling for depth, width, and resolution, resulting in more efficient models compared to previous approaches.

3.1.3.2/ Vision Transformers

The Transformer architecture [Vaswani et al., 2017], introduced in 2017, revolutionized the field of natural language processing (NLP) by providing a novel approach to machine translation tasks. As depicted in Figure 3.12, the Transformer model consists of an encoder and a decoder, both comprising multiple transformer blocks. The encoder processes input sequences and generates encodings, which are then passed to the decoder. The decoder utilizes the encodings, along with their inherent contextual information, to generate the output sequence. This process allows the Transformer to efficiently capture long-range dependencies and contextual relationships in the input data. Each transformer block is designed with a multi-head attention layer, a feed-forward neural network, a shortcut connection, and layer normalization [Han et al., 2020, Vaswani et al., 2017].

The core idea behind the transformer model is self-attention, also known as scaled dot-product attention. The self-attention layer is a crucial component in the Transformer architecture, enabling the model to attend to different parts of the input sequence simultaneously. In this layer, the input vector is first transformed into three distinct vectors: the query vector (q), the key vector (k), and the value vector (v) with dimension

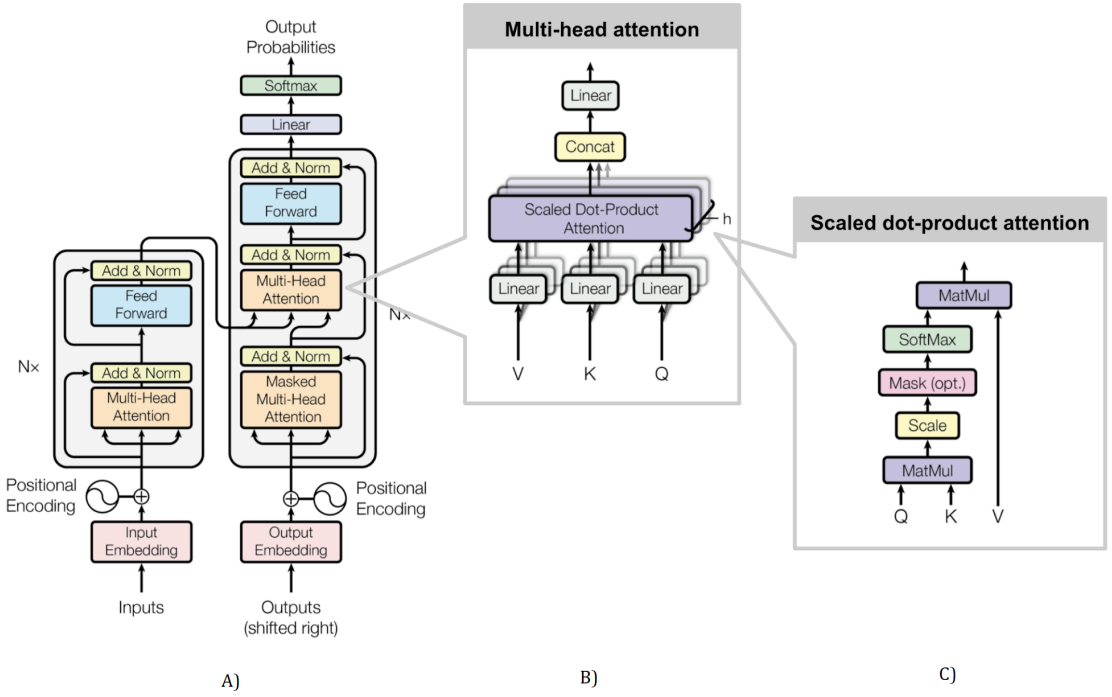

Figure 3.12: A) The original transformer architecture. B) Multi-head attention C) Self-attention. Image adapted from [Vaswani et al., 2017].

$d_q = d_k = d_v = d_{model} = 512$. The query, key, and value vectors are then organized into three matrices: Q, K, and V. The attention function is then computed in four steps. First, the scores between different input vectors are computed with $S = QK^T$. These scores measure the relevance of other words when encoding the current word. Then the scores are normalized for gradient stability, $S_n = \frac{S}{\sqrt{d_k}}$. After that, the scores are translated into probabilities using the softmax function, $P = softmax(S_n)$. Finally, each value vector is multiplied by the sum of the probabilities to give the weighted value matrix: $Z = VP$. Vectors with larger probabilities receive additional focus from the following layers. These steps can be combined into a single function, as shown below:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.2}$$

The scaled dot product attention is invariant to word order, so the self-attention layer cannot capture positional information. To incorporate positional information, a positional encoding of dimension d_{model} is added to the input embeddings before self-attention. This encoding allows the model to take into account the word's position in the sequence, enabling it to capture important contextual information. In addition to the fixed positional encoding in the vanilla transformer, there are other types of positional encoding such as learned positional encoding and relative positional encoding.

Multi-head attention is a variation of the self-attention mechanism that allows the model

to jointly attend to information from different representation subspaces at different positions. This is achieved by applying multiple attention mechanisms in parallel, each with its own set of learnable weights and activation functions, as depicted in Figure 3.15 (B). The outputs of these attention mechanisms are then combined to form the final output. The idea behind multi-head attention is to capture different types of relationships between the input elements. Each attention head can capture a different aspect of the input. By combining the outputs of multiple attention heads, the model can capture a richer representation of the input than a single attention head could [Vaswani et al., 2017, Han et al., 2020].

Mathematically, multi-head attention can be represented as:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O$$

Where Q, K, V are the query, key, and value matrices, respectively; h is the number of attention heads; $head_i$ is the output of the i^{th} attention head; and W^O is the learnable weight matrix that projects the concatenated output to the final output space.

The output of each attention head $head_i$ is computed as:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

where W_i^Q, W_i^K, W_i^V are the learnable weight matrices for the i^{th} attention head.

The input to each attention head is the concatenation of the query, key, and value matrices, which are linearly transformed using the learnable weight matrices W_i^Q, W_i^K, W_i^V. The output of each attention head is then computed using the attention mechanism, and the outputs of all attention heads are concatenated and linearly transformed using the learnable weight matrix W^O to produce the final output. Multi-head attention allows the model to capture different types of relationships between the input elements, which can improve the performance of the model [Vaswani et al., 2017].

The **feed-forward network** is a component within the transformer block that processes and transforms the information at each position in the sequence independently. It is applied after the self-attention layer in each transformer block, as shown in Figure 3.12 (A). It consists of two linear transformations with a non-linear activation function in between [Vaswani et al., 2017].

Residual connections, also known as skip connections, are employed in the transformer architecture to enhance information flow across the layers. They are added to each sub-layer in both the encoder and decoder, as shown in Figure 3.12 (A). The inclusion of residual connections strengthens the flow of information, leading to improved performance. Layer normalization is typically applied after the residual connection to normalize

the output [Han et al., 2020].

Taking inspiration from the successful scaling of Transformers in natural language processing (NLP), Dosovitskiy et al. [Dosovitskiy et al., 2020] proposed **Vision Transformer (ViT)** by directly applying a standard Transformer architecture to images with minimal modifications. Their approach involved dividing an image into patches and treating them as sequential inputs, similar to tokens (words) in NLP, as shown in Figure 3.13. The patches were then converted into linear embeddings and fed into the Transformer. The models were trained using supervised learning techniques for image classification tasks [Dosovitskiy et al., 2020].

Traditional convolutional neural networks (CNNs) have been the dominant architecture for image recognition tasks. CNNs operate on grid-like structures, such as images, by sliding convolutional filters across the input to extract local features. This sequential processing of the input limits the model's ability to capture global dependencies and relations between different parts of the image. In contrast, the ViT model introduces a new paradigm by treating images as sequences of patches and utilizing the Transformer's self-attention mechanism to capture long-range dependencies.

To handle 2D images, the image $X \in R^{H \times W \times C}$ is reshaped into a sequence of flattened 2D patches $x_p \in R^{N \times (P^2 C)}$ such that C is the number of channels, (H, W) is the resolution of the original image, and (P, P) is the resolution of each image patch. The effective sequence length for the transformer is $N = H \times W / P^2$. As the transformer uses constant latent vector size D in all of its layers, a trainable linear projection maps each vectorized patch to the model dimension D. The output of this projection is referred to as patch embeddings [Dosovitskiy et al., 2020, Han et al., 2020].

The learnable class token in ViTs [Dosovitskiy et al., 2020] functions similarly to the [class] token in BERT [Devlin et al., 2019]. It is an embedding applied to the sequence of embedding patches. The state of this embedding represents the image representation. Both during pretraining and fine-tuning, classification heads are attached to this embedding, maintaining the same size. Additionally, 1D position embeddings are added to the patch embeddings to preserve positional information. It is important to note that ViT uses the vanilla transformer's [Vaswani et al., 2017] encoder. The encoder consists of a series of layers that alternate between multiheaded self-attention (MSA) and multilayer perceptron (MLP) blocks. Layer normalization (LN) is applied before each block, and residual connections are used after each block. The MLP block typically consists of two layers with a Gaussian Error Linear Unit (GELU) non-linearity activation function. ViT is typically pre-trained on large datasets and subsequently fine-tuned on smaller datasets for specific tasks [Dosovitskiy et al., 2020, Han et al., 2020]. The Vision Transformer (ViT) has demonstrated impressive results on various computer vision tasks like image classification, object detection, and semantic segmentation. A key advantage of

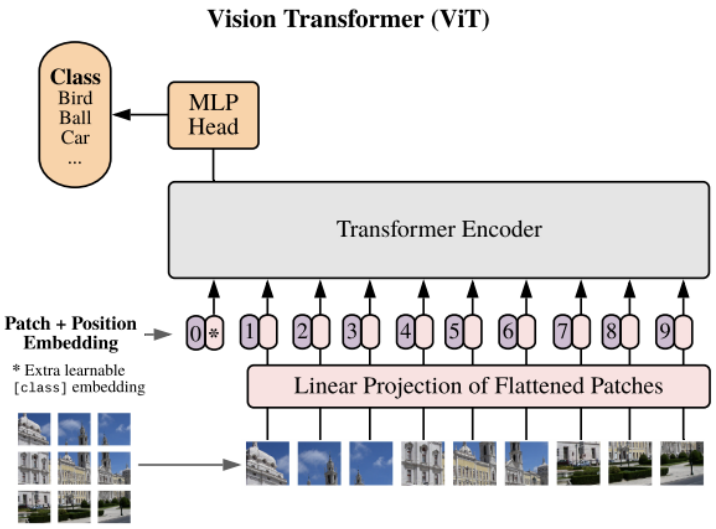

Figure 3.13: Vision Transformer architecture. Image adapted from [Dosovitskiy et al., 2020].

ViT is that the self-attention mechanism enables highly parallelizable computation across image patches. However, relying on image patches may cause ViT to miss fine-grained details that convolutional neural networks (CNNs) can capture. Additionally, ViT struggles with more general vision tasks like detection and segmentation compared to CNNs. This is due to ViT having fewer built-in inductive biases that are beneficial for images, such as locality and translation equivariance. Furthermore, the global self-attention used in ViT leads to quadratic complexity relative to input size, making it inefficient for high-resolution images.

In vanilla Transformer models [Vaswani et al., 2017, Dosovitskiy et al., 2020], tokens have a fixed scale, which is not ideal for vision applications. Images have a much higher pixel resolution compared to words in text passages. Tasks like semantic segmentation require dense prediction at the pixel level, but this is challenging for Transformers on high-resolution images due to the quadratic computational complexity of self-attention. To address these issues, Liu et al. [Liu et al., 2021] proposed **Swin Transformer**, a versatile Transformer backbone. Swin Transformer constructs hierarchical feature maps and achieves linear computational complexity with respect to image size. It starts with small patches (outlined in gray) and progressively merges neighboring patches in deeper layers, as depicted in Figure 3.14. This hierarchical representation allows Swin Transformer to leverage advanced techniques like feature pyramid networks (FPN) or U-Net for dense prediction. The linear computational complexity is achieved by computing self-attention locally within non-overlapping windows that divide the image (outlined in red). Swin Transformer is suitable for various vision tasks, unlike previous Transformer archi-

tectures [Dosovitskiy et al., 2020] that produce feature maps of a single resolution and have quadratic complexity due to computation of self-attention globally.

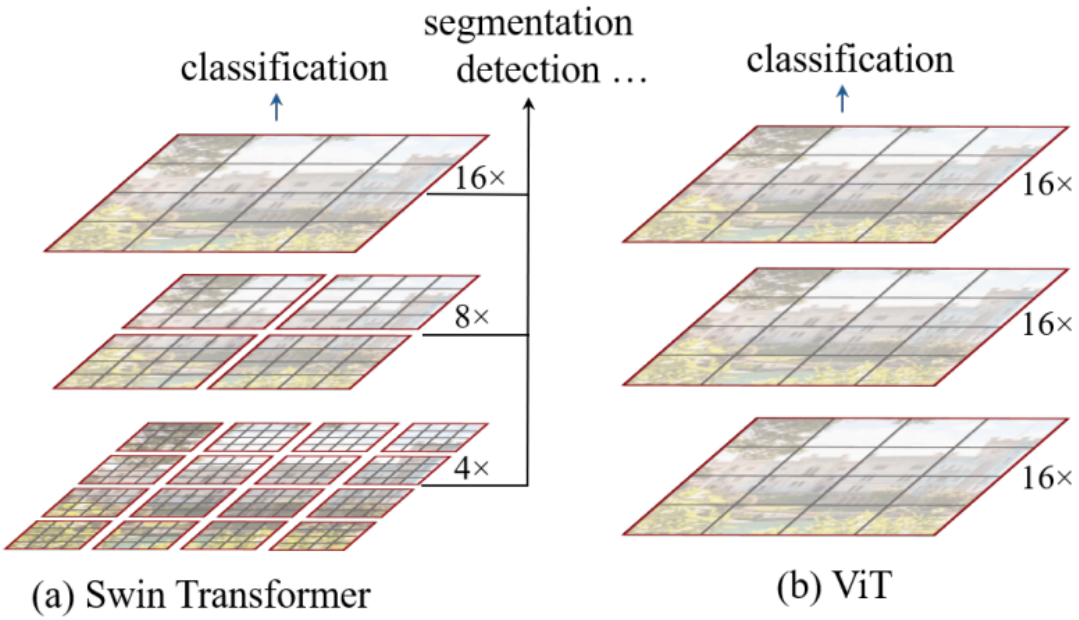

Figure 3.14: (a) The Swin Transformer architecture produces hierarchical feature maps by merging image patches, represented in gray, in deeper layers. (b) In contrast, vision Transformers (ViTs) generate feature maps with a single low resolution. Image adapted from [Liu et al., 2021].

In Figure 3.15, an overview of the Swin Transformer architecture is presented. The input RGB image is split into non-overlapping patches using a patch-splitting module, similar to ViT. Each patch is treated as a "token," and its feature is obtained by concatenating the raw pixel RGB values. In their implementation, [Liu et al., 2021] used a patch size of 4×4, resulting in a feature dimension of 48 $(4 \times 4 \times 3)$. A linear embedding layer is then applied to project this raw-valued feature to an arbitrary dimension denoted as C. These patch tokens, along with the Swin Transformer blocks, which have a modified self-attention computation, and linear embedding, are referred to as "Stage 1" and maintain the number of tokens $(H/4 \times W/4)$.

To create a hierarchical representation, patch merging layers are employed to reduce the number of tokens as the network progresses. The features of each group of 2 × 2 neighboring patches are concatenated by the first patch merging layer, and a linear layer is applied to the concatenated 4C-dimensional features. This downsamples the resolution by a factor of 2×2 = 4 and sets the output dimension to 2C. Subsequently, Swin Transformer blocks are applied for feature transformation while keeping the resolution at $(H/8 \times W/8)$. This initial stage of patch merging and feature transformation is denoted as "Stage 2". The process is repeated twice more, resulting in "Stage 3" and "Stage 4" with output resolutions of $(H/16 \times W/16)$ and $(H/32 \times W/32)$, respectively. A hierarchical rep-

resentation is jointly produced by these stages, having the same feature map resolutions as conventional convolutional networks such as VGG and ResNet. As a result, the proposed architecture can conveniently replace the backbone networks in existing methods for various vision tasks [Liu et al., 2021].

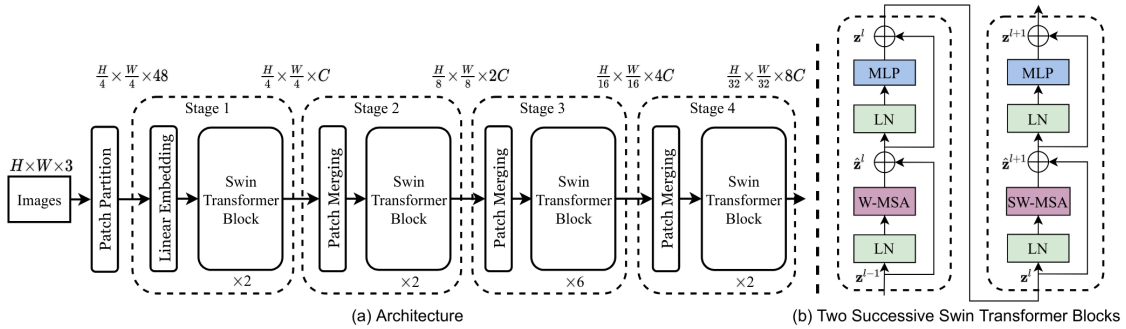

Figure 3.15: (a) Swin Transformer architecture (b) two successive Swin Transformer Blocks. Image adapted from [Liu et al., 2021].

The **Swin Transformer block** is constructed by replacing the standard multi-head self-attention (MSA) module in the vanilla Transformer block with a module based on shifted windows, while keeping other layers unchanged. In Figure 3.15(b), a Swin Transformer block consists of a shifted window-based MSA module followed by a 2-layer MLP with GELU nonlinearity in between, with LayerNorm (LN) applied before each module and residual connections after [Liu et al., 2021].

To enable efficient modeling, self-attention is computed within local windows. These windows evenly partition the image in a non-overlapping manner. If each window contains $M \times M$ patches, the computational complexity of a global MSA module and a window-based one on an image with $h \times w$ patches can be computed as follows:

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C,$$

$$\Omega(W - MSA) = 4hwC^2 + 2M^2hwC,$$

Where the former exhibits quadratic complexity with respect to the number of patches hw, while the latter has linear complexity when M remains fixed (default value of 7). Global self-attention computation becomes impractical for large hw, whereas window-based self-attention remains scalable.

However, the window-based self-attention module lacks connections across windows, limiting its modeling capability. To address this and maintain efficient computation with non-overlapping windows, [Liu et al., 2021] proposed a shifted window partitioning approach. This approach alternates between two partitioning configurations in consecutive Swin Transformer blocks. As shown in Figure 3.16, the first module employs a regular

window partitioning strategy, starting from the top-left pixel. The 8×8 feature map is evenly divided into 2×2 windows of size $4 \times 4 (M = 4)$. Subsequently, the next module adopts a shifted window configuration, displacing the windows by $(M/2, M/2)$ pixels from the regularly partitioned windows in the preceding layer [Liu et al., 2021]. The computation of consecutive Swin Transformer blocks is performed using the shifted window partitioning approach, which can be summarized as follows:

$$\begin{aligned} \hat{z}^l &= WMSA(LN(z^{l-1})) + z^{l-1}, \\ z^l &= MLP(LN(\hat{z}^l)) + \hat{z}^l, \\ \hat{z}^{l+1} &= SWMSA(LN(z^l)) + z^l, \\ z^{l+1} &= MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1}, \end{aligned} \tag{3.3}$$

Where $\hat{z}^l$ and z^l represent the output features of the (S)WMSA module and the MLP module for block l, respectively. W-MSA and SW-MSA denote window-based multi-head self-attention using regular and shifted window partitioning configurations, respectively. By incorporating connections between neighboring non-overlapping windows from the previous layer, the shifted window partitioning approach has proven to be successful in tasks like image classification, object detection, and semantic segmentation [Liu et al., 2021].

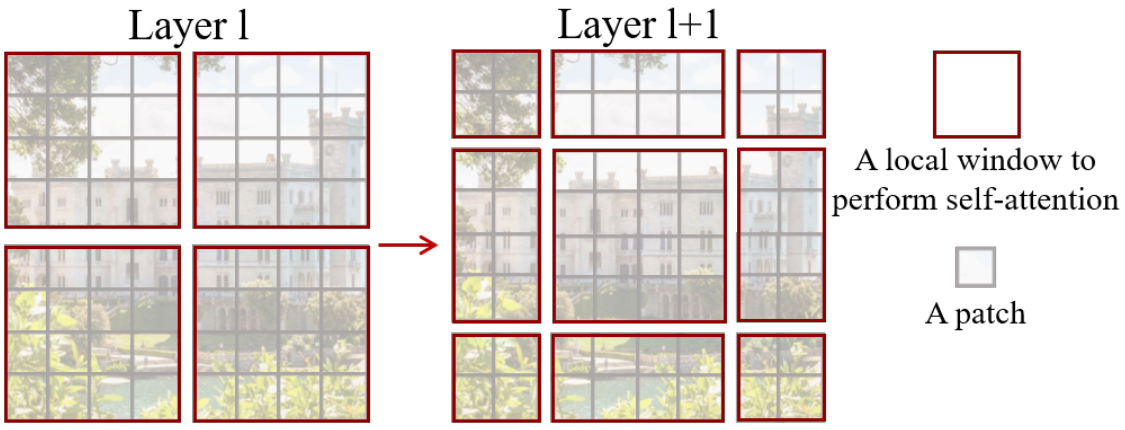

Figure 3.16: An illustration of the shifted window approach for computing self-attention in the Swin Transformer architecture. Image adapted from [Liu et al., 2021].

Positional encoding is essential in transformers and Vision Transformers (ViTs) as these models do not inherently model the order or position of tokens or image patches in their input sequences. Unlike recurrent neural networks (RNNs) and convolutional neural networks (CNNs), which capture positional information through their sequential or spatial operations, transformers and ViTs process their inputs in parallel without regard to position. This lack of positional encoding is problematic for tasks relying on element order or position. Natural language processing tasks like machine translation and language understanding depend heavily on word order. Similarly, in computer vision, the spatial layout of

image patches is crucial for capturing fine details and overall structure [Wu et al., 2021]. Therefore, positional representations are essential in transformers to overcome their sequence order invariance and enable modeling of structured data. There are two main approaches used to encode positional information for transformers: Absolute Position Encoding and Relative Position Encoding.

Absolute position encoding is a method to represent the absolute position of each element in a sequence. It allows the model to differentiate between elements based on their positions. In the vanilla Transformer model [Vaswani et al., 2017], absolute position encoding is achieved by adding positional embeddings to the input token embedding. The positional embeddings are added element-wise to the input embeddings of the sequence, enabling the model to attend to different positions based on the added positional information. There are several choices of absolute positional encodings, such as the fixed encodings by sine and cosine functions with different frequencies and the learnable encodings through training parameters [Wu et al., 2021].

Relative position encoding is an alternative approach that aims to capture positional relationships between elements in a sequence by calculating the relative distance between input elements and by learning the pairwise relationships between tokens. It allows the model to focus on relative distances and patterns rather than relying solely on absolute positions. Relative position encoding is achieved by incorporating relative positional biases into the attention mechanism. Instead of using absolute positional embeddings, the attention scores are adjusted based on the relative positions between the query and key elements. More specifically, they encode the relative position between the input elements into query vector (Q_p), the key vector (K_p), and the value vector (V_p), which have the same dimension as d_{model} (in Eq. 3.2). The encoding vectors are embedded into the self-attention module by re-formulating the self-attention (Eq. 3.2) as follows [Wu et al., 2021]:

$$Attention = softmax(\frac{(Q + Q_p)(K + K_p)^T}{\sqrt{dk}})(V + V_p) \tag{3.4}$$

In this way, the pairwise positional relation is learned during transformer training. Relative position encoding is particularly useful in tasks where the absolute position of elements is less important, but the relationships or dependencies between elements matter more. For example, in tasks like object detection or semantic segmentation, the relative positions of objects or regions within an image are more relevant than their absolute positions.

3.1.4/ TRAINING NEURAL NETWORKS

Neural networks need to be trained before they can make accurate predictions. Training involves optimizing the weights and biases of the network to minimize the difference be-

tween the network's predictions and the true target values. Network training is an iterative process done in two main steps: forward propagation and backward propagation.

During forward propagation, the input data is fed through the network layer by layer until a prediction is generated. Neurons within each layer receive inputs from the preceding layer, perform weighted summations, and apply an activation function to produce an output. This output becomes the input for the subsequent layer, continuing until the output layer yields a prediction.

After making a prediction, the network compares the predicted value to the true target value. It then calculates the loss or the error. The loss function quantifies the discrepancy between the predicted and actual outputs. During backward propagation, this error is propagated back through the network in the reverse direction, layer by layer. Leveraging the chain rule of calculus, the derivatives of the error with respect to the weights and biases are calculated. These derivatives, or gradients, indicate the contribution of each weight to the error. The weights and biases are then adjusted to minimize the error. The magnitude of these adjustments is determined by the optimization algorithm [Bushaev, 2017].

Optimizers play a critical role in the training process of neural networks by guiding the updates to the network's weights and biases based on the computed gradients. Various optimization algorithms are employed during the backpropagation to enhance the stability and efficiency of the training process. One commonly used algorithm is gradient descent, which updates the network's parameters in the direction opposite to the computed gradients. The size of the weight updates is determined by the learning rate, which scales the negative gradients.

Stochastic gradient descent (SGD), a variant of gradient descent, updates the weights for each sample individually, while mini-batch gradient descent applies weight updates for a subset of the training set, resulting in smoother parameter adjustments. Another popular optimizer is Adam (Adaptive Moment Estimation), which combines the advantages of adaptive learning rates and momentum methods. Adam maintains separate learning rates for each parameter, adapting them based on the gradients. It also incorporates momentum, facilitating faster convergence by accumulating past gradients [Bushaev, 2017].

3.1.4.1/ Loss Functions

In a neural network, a loss function, also known as a cost function or objective function, quantifies the discrepancy between the predicted output of the network and the true target values. It serves as a measure of how well the network is performing on a given task [Pandit, 2023]. The choice of a suitable loss function depends on the nature of the problem at hand, such as regression, classification, or segmentation.

Regression problems involve predicting continuous numerical values. Mean Squared Error (MSE) is one of the most widely used loss functions for regression. It is computed as the mean of the squared differences between each prediction and target value. Mean Absolute Error (MAE) is an alternative to MSE that measures the average absolute difference between the predicted and true values. It is less sensitive to outliers compared to MSE and provides a more robust measure of error [Pandit, 2023].

Classification problems involve assigning input data to predefined classes or categories. Binary cross-entropy loss (BCE) is commonly used for binary classification tasks, where there are only two classes. It measures the dissimilarity between the true class labels and the predicted probabilities as shown in Eq. 3.5, where y is the ground truth label (0 or 1) and $\hat{y}$ is the predicted value. The loss function penalizes the model more for incorrect predictions with higher confidence [Pandit, 2023].

$$L_{BCE} = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \tag{3.5}$$

Categorical cross-entropy loss (CE) is used for multi-class classification problems. It calculates the average cross-entropy loss over all classes, as shown in Eq. 3.6, where y is the ground truth label vector (a one-hot vector), $\hat{y}$ is the predicted probability vector and C is the number of classes [Pandit, 2023]. The loss function encourages the model to assign high probabilities to the correct class and low probabilities to the other classes.

$$L_{CE} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{3.6}$$

Sparse categorical cross-entropy loss is similar to categorical cross-entropy but is used when the true class labels are provided as integers rather than one-hot encoded vectors. It avoids the need for one-hot encoding by directly comparing the true labels with the predicted probabilities.

In image segmentation tasks, binary cross-entropy (BCE) and categorical cross-entropy (CE) loss functions can also be employed, treating each pixel as a binary or multi-class classification problem. For example, for multi-class segmentation, the CE loss can be modified by incorporating a pixel-wise summation, as depicted in Eq. 3.7, where M represents the number of pixels in the corresponding image.

$$L_{CE(seg)} = -\frac{1}{M} \sum_{j=1}^{M} \sum_{c=1}^{C} y_j^c \log(\hat{y}_j^c) \tag{3.7}$$

Weighted cross-entropy (WCE) is a frequently used extension of CE, which is used to address potential class imbalance issues encountered in medical image segmentation

tasks by penalizing majority classes [Ma et al., 2021]. It is computed as shown in Eq. 3.8, where w_c denotes the weight assigned to each class. The weights w_c are typically inversely proportional to the class frequencies.

$$L_{WCE(seg)} = -\frac{1}{M}\sum_{j=1}^{M}\sum_{c=1}^{C} w_c y_j^c \log(\hat{y}_j^c) \tag{3.8}$$

Region-based loss functions are also utilized in image segmentation to minimize the mismatch or maximize the overlap regions between the ground truth and predicted segmentation [Ma et al., 2021]. One of the most prevalent region-based loss functions is the Dice loss [Milletarì et al., 2016], which quantifies the similarity between the predicted and true segmentation masks. There are two common variants [Isensee et al., 2021]: one employing squared terms in the denominator [Milletarì et al., 2016], and the other without squared terms [Drozdzal et al., 2016] [Ma et al., 2021]. The squared version, depicted in Eq. 3.9, calculates the Dice loss by summing the product of predicted probabilities and ground truth labels and dividing it by the sum of the squares of predicted probabilities and ground truth labels for each class.

$$L_{Dice} = 1 - \frac{2\sum_{j=1}^{M}\sum_{c=1}^{C} y_j^c \hat{y}_j^c}{\sum_{j=1}^{M}\sum_{c=1}^{C}(y_j^c)^2 + \sum_{j=1}^{M}\sum_{c=1}^{C}(\hat{y}_j^c)^2} \tag{3.9}$$

Compound loss functions are often used in segmentation methods, combining multiple loss functions [Ma et al., 2021]. A widely used compound loss for segmentation is the combination of Dice loss and cross-entropy loss, as expressed in Eq. 3.10.

$$L_{DiceCE} = L_{Dice} + L_{CE} \tag{3.10}$$

3.1.4.2/ Regularization

Regularization refers to a set of techniques that aim to prevent overfitting, where the model becomes overly specialized to the training data and performs poorly on unseen data [Jain, 2018]. This is particularly important in medical image analysis, where the availability of labeled data is often limited and the models need to generalize well to unseen patient cases.

Here are some common regularization techniques used in deep learning [Jain, 2018]:

- **Weight regularization**: Weight regularization, also known as weight decay, is a widely used technique to control the complexity of neural networks. It introduces a penalty term to the loss function, discouraging the model from relying excessively on individual weights. L1 regularization encourages sparsity in the weights by adding

the sum of the absolute weights, leading to some weights being set to zero, while L2 regularization promotes smaller weights by adding the sum of squared weights to the loss function.

- **Dropout**: Dropout is a regularization technique that randomly deactivates a fraction of neurons during training. This prevents neurons from relying too heavily on specific input features and encourages them to learn more robust and generalizable representations. Dropout effectively acts as an ensemble of multiple models, reducing the risk of overfitting.
- **Data Augmentation**: Data augmentation involves applying a variety of transformations to the training images, such as rotations, translations, flips, zooms, or intensity variations. This artificially increases the size of the training dataset and introduces diversity, helping the model learn more invariant and robust features. Data augmentation is particularly useful in medical image analysis, where variations in patient positioning, imaging modalities, and acquisition protocols are common.
- **Early Stopping**: Early stopping is a straightforward yet effective regularization technique. It involves monitoring the model's performance on a validation set during training and stopping the training process when the performance begins to decline. This approach helps prevent overfitting by finding a balance between model complexity and generalization.
- **Transfer Learning**: Transfer learning involves leveraging pre-trained models that were trained on large-scale datasets, such as ImageNet, and adapting them to the medical image analysis task at hand. By utilizing the knowledge learned from these datasets, transfer learning allows models to generalize better with limited medical image data.

3.1.4.3/ Evaluation Metrics

Automated segmentation algorithms are commonly evaluated using two types of metrics: overlap-based metrics, such as the Dice Similarity Coefficient, and surface distance-based metrics, such as the Hausdorff distance.

Overlap-based metrics assess the similarity between a segmented region and a ground truth region in image segmentation tasks. These metrics provide a measure of segmentation accuracy by quantifying the degree of overlap between the segmented region and the ground truth. The Dice Similarity Coefficient [Dice, 1945] is a widely used metric for evaluating segmentation performance. It calculates the overlap between the segmented region (S) and the ground truth region (G) by considering the intersection of their pixel sets and the sum of their pixel counts, as shown in Eq. 3.11 [Taha et al., 2015]. The

equation also illustrates an alternative formulation of the Dice Similarity Coefficient using true positives (TP), false positives (FP), and false negatives (FN).

$$Dice = \frac{2|S \cap G|}{|S| + |G|} = \frac{2TP}{2TP + FP + FN} \tag{3.11}$$

where TP represents the number of pixels or voxels that are correctly identified as positive, FP denotes the number of pixels or voxels that are incorrectly identified as positive and FN represents the number of pixels or voxels that are incorrectly identified as negative, as depicted in Figure 3.17.

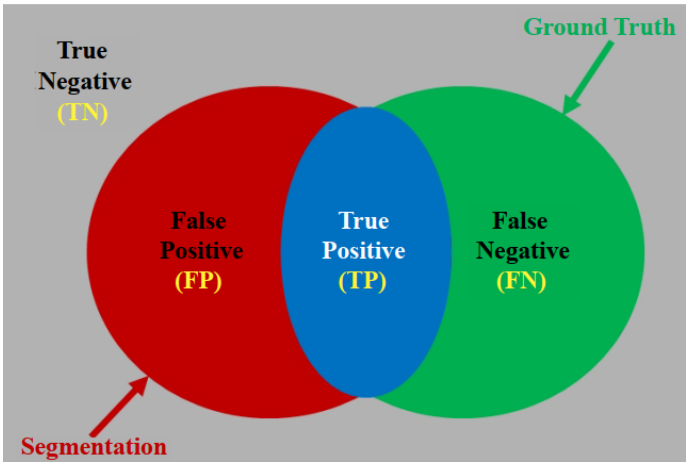

Figure 3.17: Schematic representation depicting the measurement of segmentation errors for the computation of the Dice similarity coefficient. Image adapted from [McClure et al., 2014].

The Dice value varies between 0 and 1, with a value of 1 indicating a perfect overlap between the segmentation and ground truth. A higher Dice value indicates a higher degree of similarity between the segmentation and the ground truth, reflecting a more accurate and precise segmentation result.

Surface distance-based metrics play a crucial role in evaluating image segmentation quality by measuring dissimilarity. The Hausdorff distance [Huttenlocher et al., 1993] is a well-known surface distance-based metric commonly used for this purpose. It quantifies the dissimilarity between two sets of points, typically representing the boundaries of a segmented region and a ground truth region. The Hausdorff Distance (HD) between two finite point sets A and B is defined by Eq. 3.12, where $h(A, B)$ represents the directed Hausdorff distance [Taha et al., 2015].

$$HD(A, B) = \max(h(A, B), h(B, A)) \tag{3.12}$$

The directed Hausdorff distance, denoted as $h(A, B)$, is calculated as the maximum distance between a point in set A and its closest point in set B. The Hausdorff Distance is

computed by taking the maximum value between $h(A, B)$ and $h(B, A)$, which accounts for dissimilarities in both directions.

The Hausdorff Distance metric evaluates the maximum discrepancy or dissimilarity between the boundaries of the segmented region and the ground truth region. Typically, sets A and B correspond to the boundary points or contours of the segmented region and the ground truth region, respectively. A smaller Hausdorff Distance indicates a higher level of similarity between the two sets, reflecting a better alignment between the segmentation and the ground truth boundaries.

The Hausdorff distance metric is known to be sensitive to outliers, which can be problematic in medical segmentations where noise and outliers are common. Consequently, directly using the HD metric is sometimes not recommended [Gerig et al., 2001]. However, an alternative approach to address this issue is the quantile method proposed by Huttenlocher et al. [Huttenlocher et al., 1993]. The Hausdorff quantile method suggests redefining the HD as the q^{th} quantile of distances instead of the maximum. By doing so, possible outliers are excluded from the computation, providing a more robust measurement. The specific value of q is selected based on the particular application and the characteristics of the point sets being measured [Taha et al., 2015].

3.1.5/ Deep learning for cardiac MR image segmentation

In recent years, deep learning has emerged as a powerful technique in various medical imaging applications, including cardiac MR image segmentation. As mentioned in Section 2.4.1, cardiac MR imaging is crucial in diagnosing and assessing various cardiovascular diseases, providing detailed anatomical and functional information about the heart. Accurate segmentation of cardiac structures and pathological tissues such as scars from cardiac MR images is essential for quantitative analysis, surgical planning, and disease monitoring. While traditional image segmentation methods have been employed with some success, they often require manual intervention, extensive preprocessing and may struggle with complex anatomical variations, noise, and imaging artifacts. On the other hand, deep learning techniques have shown great potential in automating the cardiac MR image segmentation process, offering improved accuracy, efficiency, and robustness.

Several studies have been conducted to explore the application of deep learning algorithms for cardiac MR image segmentation. These studies have demonstrated promising results and paved the way for the development of advanced segmentation approaches in cardiac imaging. The use of convolutional neural networks (CNNs), in particular, has gained significant attention due to their ability to learn hierarchical representations of features from images automatically. Several studies have employed CNN-based architectures, such as Fully Convolutional Network (FCN) [Shelhamer et al., 2014] and U-Net

[Ronneberger et al., 2015], for cardiac MR image segmentation.

FCN [Shelhamer et al., 2014] are a type of convolutional neural network that can take an input of arbitrary size and produce a correspondingly sized output. Unlike standard CNNs, which have fully-connected layers at the end that require fixed-size inputs, FCNs replace the fully-connected layers with convolutional layers that preserve spatial information. FCNs are designed with an encoder-decoder structure, as depicted in Figure 3.18 (A), allowing them to process input images of any size and generate output maps of the same size. The encoder component of FCN transforms the input image into high-level feature representations, while the decoder interprets these feature maps and recovers spatial details through operations such as transposed convolutions. Transposed convolutions are commonly used to upscale the feature maps by a factor of 2, although alternative approaches like unpooling and upsampling layers can also be employed. However, the simple encoder-decoder structure of FCNs may fail to capture intricate context information due to the elimination of some features by pooling layers in the encoder [Chen et al., 2019b].

U-Net [Ronneberger et al., 2015] is a specific type of FCN architecture that has gained substantial popularity in medical image segmentation tasks, including cardiac structure segmentation. U-Net is characterized by its U-shaped architecture, which consists of an encoder pathway followed by a decoder pathway. The encoder pathway captures context and abstract features through a series of down-sampling operations, while the decoder pathway performs up-sampling and concatenation of features from the encoder to recover spatial resolution. U-Net incorporates skip connections between the encoder and decoder, as illustrated in Figure 3.18 (B). These skip connections enable the recovery of spatial context lost during the down-sampling process, leading to more precise segmentation results. The U-Net, along with its 3D variants such as the 3D U-Net [Çiçek et al., 2016] and the 3D V-Net [Milletarì et al., 2016], has been widely adopted as the backbone network in state-of-the-art cardiac image segmentation methods. These methods have demonstrated promising segmentation accuracy in various cardiac segmentation tasks [Chen et al., 2019b]. Recently, nnU-Net [Isensee et al., 2021] has been proven a successful adaptive framework that automatically configures itself, including preprocessing, network architecture, training and post-processing for automatic segmentation of different types of medical images. It has been shown to be effective for a variety of medical image segmentation tasks and winning multiple medical image segmentation challenges [Ma, 2021].

Recently, a notable advancement in deep learning approaches for medical image segmentation has been the utilization of vision transformer-based models. Vision transformers [Dosovitskiy et al., 2020], originally introduced for natural image classification tasks, have shown remarkable capabilities in capturing long-range dependencies and modeling

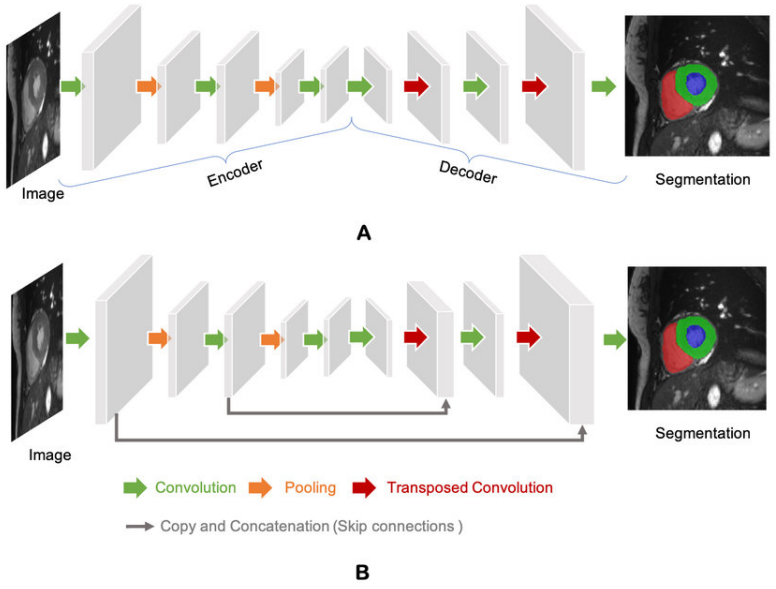

Figure 3.18: A) FCN and B) U-Net segmentation architectures. Image adapted from [Chen et al., 2019b].

complex spatial relationships, as described in detail in Section 3.1.3.2. This has motivated researchers to adapt vision transformers for medical image analysis, including cardiac MR image segmentation. This section aims to provide an overview of the existing literature on deep learning approaches for cardiac MR image segmentation, highlighting their methodologies, achievements, and potential limitations.

3.1.5.1/ DEEP LEARNING FOR CARDIAC STRUCTURES SEGMENTATION

In cardiac magnetic resonance (CMR) imaging, various anatomical structures of the heart can be visualized and segmented. These cardiac structures are essential for understanding the cardiac anatomy, function, and pathology. In this thesis, we are mainly focusing on the segmentation of left ventricular cavity or blood pool (LV), right ventricular blood pool (RV) and left ventricular myocardium (MYO) from CMR images.

Tran (2016) [Tran, 2016] pioneered the use of fully convolutional networks (FCNs) to directly segment the left ventricle, myocardium, and right ventricle in short-axis cardiac MR images. Their end-to-end FCN approach demonstrated faster and more accurate segmentation compared to traditional methods. Since the introduction of the Automated Cardiac Diagnosis Challenge (ACDC) [Bernard et al., 2018] in 2017, numerous deep learning approaches have been proposed to further improve cardiac segmentation performance. Isensee et al. [Isensee et al., 2017] used an ensemble of 2D and 3D U-Nets, while Khened et al. [Khened et al., 2018] developed a dense U-Net with inception modules to combine multiscale features for robust segmentation across vari-

able anatomy. Other works have explored different loss functions like weighted cross-entropy, Dice loss, deep supervision loss, and focal loss to boost segmentation accuracy [Jang et al., 2017, Yang et al., 2017, Sander et al., 2019, Chen et al., 2019c]. Most methods utilize 2D networks rather than 3D due to the typically low through-plane resolution and motion artifacts in cardiac MR scans, which limit the utility of 3D networks [Baumgartner et al., 2017, Chen et al., 2019b].

Li et al. [Li et al., 2019] proposed a two-stage method using FCNs for CMR image segmentation. In the first stage, they localized the heart region by identifying a region of interest (ROI). Subsequently, the localized region was used to segment the left ventricular blood pool, myocardium, and right ventricular blood pool. This two-stage approach allowed for accurate segmentation of the cardiac structures within the CMR image.

One limitation of 2D networks in cardiac segmentation is that they operate slice-by-slice without leveraging inter-slice dependencies. Consequently, 2D networks can fail to locate and segment the heart on challenging apical and basal slices where ventricular contours are poorly defined. To provide additional contextual guidance to the 2D segmentation networks, some approaches have incorporated shape priors learned from labels or multi-view images [Zotti et al., 2017, Chen et al., 2019a]. Others extract spatial information from neighboring slices using recurrent units (RNNs) or multi-slice networks (2.5D) to aid segmentation [Chen et al., 2019b].

A limitation of 2D and 3D FCNs trained with pixel-wise loss functions is that they may not learn features representing underlying anatomy. To improve prediction accuracy and robustness, some approaches incorporate anatomical constraints as regularization terms during training. These constraints account for topology [Clough et al., 2019], contour/region information [Chen et al., 2019a] or shape [Oktay et al., 2017] to encourage anatomically plausible segmentations. Along with network regularization during training, Painchaud et al. [Painchaud et al., 2019] proposed a variational autoencoder to post-process and correct inaccurate segmentations.

As part of the M&Ms (Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge) [Campello et al., 2021] challenge, [Full et al., 2020, Ma, 2020b] addressed the challenge of domain shift or distribution shift in CMR image segmentation by employing data augmentation-based solutions. These solutions included techniques such as histogram matching, contrast modification, and image synthesis. By utilizing these methods, they tried to reduce the differences in data distribution between training and testing sets, improving the robustness and generalizability of the segmentation models.

Looking at the performance of transformer-based methods for cardiac image segmentation, Chen et al. [Chen et al., 2021a] introduced TransUNet as a novel method for cardiac structure segmentation, leveraging the strengths of both convolutional neural networks (CNNs) and Transformers. TransUNet follows a two-step approach, where high-resolution

spatial features are initially extracted using a CNN. Subsequently, global context information is encoded using a Transformer, which incorporates self-attentive features. These encoded features from the Transformer are then upsampled and merged with features from multiple scales extracted through skip connections in the encoding path. This fusion of features enables precise localization in the segmentation process. TransUNet demonstrated superior performance compared to other models like V-Net [Milletarì et al., 2016], Attention U-net [Oktay et al., 2017], and ViT [Dosovitskiy et al., 2020] for multi-organ and cardiac segmentation tasks [He et al., 2022].

In a similar way, Xu et al. [Xu et al., 2021] proposed LeViT-UNet, which integrates a Light Vision Transformer (LeViT) [Yang et al., 2022] Transformer module into the U-Net architecture for efficient and accurate segmentation of cardiac MR images. In LeViT-UNet, the LeViT Transformer serves as the encoder, providing a better balance between accuracy and efficiency in the Transformer block. Moreover, the authors incorporated multi-scale feature maps from both transformer blocks and convolutional blocks of LeViT into the decoder through skip connections. This integration allows for the effective utilization of spatial information present in the feature maps, enhancing the segmentation performance [He et al., 2022].

Gao et al. [Gao et al., 2021] introduced UTNet, a model that incorporates self-attention modules in both the encoder and decoder blocks to capture long-range dependencies at multiple scales while maintaining computational efficiency. They proposed an efficient implementation of self-attention and relative position encoding techniques, reducing complexity without sacrificing performance [He et al., 2022]. This allowed UTNet to effectively model global interactions in the input data.

On the other hand, Cao et al. [Cao et al., 2021] proposed Swin-Unet, a Transformer-based architecture designed specifically for cardiac MR image segmentation and other medical imaging tasks. Swin-Unet adopts a U-shaped encoder-decoder architecture with skip connections to facilitate the learning of local and global semantic features. The encoder component utilizes a hierarchical Swin Transformer with shifted windows to extract contextual features, enabling effective context modeling. Meanwhile, the decoder, based on a symmetric Swin Transformer, performs up-sampling operations to restore the spatial resolution of the feature maps, aiding in accurate segmentation of CMR images [He et al., 2022].

3.1.5.2/ DEEP LEARNING FOR CARDIAC SCAR TISSUE SEGMENTATION

Cardiac scar tissue segmentation from LGE CMR images is of great importance in the evaluation and management of patients with various cardiac pathologies, including myocardial infarction, cardiomyopathies, and arrhythmias. LGE CMR imaging provides high-

resolution images that highlight regions of myocardial fibrosis or scar, which are crucial in assessing the extent and location of damaged tissue.

Approaches to scar segmentation can be divided into non-deep learning and deep learning methods. Non-deep learning techniques rely on thresholding and clustering. Thresholding exploits the enhanced intensity of infarcted myocardium compared to healthy myocardium. Full width at half maximum (FHFW) defines the threshold as half the maximum infarcted intensity [Amado et al., 2004]. The nSD method uses a threshold n standard deviations above healthy mean intensity, with n between 2-6 [Kim et al., 1999]. While simple, these methods require a manual region of interest selection to determine thresholds. Clustering methods like Gaussian mixture modeling [Hennemuth et al., 2012] and Fuzzy C-means [Baron et al., 2013] classify myocardial intensities to segment scars but still need some manual intervention.

Recent deep learning methods use semi-automatic or fully automatic approaches. Zabihollahy et al. (2018) manually segmented the myocardium and then applied a 2D FCN for scar segmentation [Zabihollahy et al., 2018]. Moccia et al. (2019) proposed a semi-automatic and fully automatic scar segmentation method [Moccia et al., 2018]. The former approach outperformed the latter approach due to the mediocre segmentation performance of the fully automatic method on the myocardium. De la Rosa et al. (2019) used a fully automatic pipeline of 2D U-net myocardium segmentation, top-hat transform coarse scar segmentation, and final voxel classification [de la Rosa et al., 2019]. In addition, Fahmy et al. (2019) demonstrated the effectiveness of a deep learning approach using a 3D CNN for segmenting left ventricular scar in patients with hypertrophic cardiomyopathy (HCM), outperforming 2D CNNs, and demonstrating comparable performance in a multi-center and multi-vendor setting [Fahmy et al., 2019b].

More recently, the use of deep learning methods for myocardial scar segmentation from LGE MRI has gained some attention, especially following the EMIDEC (automatic Evaluation of Myocardial Infarction from Delayed Enhancement Cardiac MRI) [Lalande et al., 2020] challenge at MICCAI 2020. Numerous approaches have been proposed for this task, typically adopting a two-stage cascaded framework. These approaches commonly involve delineating the myocardium as a Region of Interest (ROI) in the first stage and subsequently segmenting the different myocardial tissues within the ROI using another model in the second stage. Alternatively, some studies have proposed one-stage models that aim to achieve end-to-end segmentation of all the target tissues [Lalande et al., 2021].

For instance, Zhang 2020 proposed a cascaded 2D-3D framework that utilizes a 2D U-Net for initial segmentation, focusing on intra-slice information [Zhang, 2020]. This is followed by a 3D U-Net that incorporates both the original volume and the 2D segmentation information to refine the segmentation. Similarly, Brahim et al. (2020) introduced

a two-stage deep learning framework for enhanced segmentation of myocardial diseases [Brahim et al., 2020]. In the first step, they employed an encoder-decoder segmentation network to generate segmentations of the myocardium and cavity from the entire volume. Then, a 3D U-Net incorporating shape priors was used to identify the segmentation of myocardial infarction based on the predictions from the first network.

Feng et al. 2020 proposed an automatic LGE-MRI segmentation model that addressed image orientation dependency by incorporating rotation-based augmentation [Feng et al., 2020]. They utilized a dilated 2D U-Net to enhance the network's robustness and employed weighted cross-entropy and soft-Dice loss functions to handle class imbalance. Another approach by Yang et al. 2020 introduced a hybrid U-Net network for simultaneous segmentation of various regions in LGE-MRI [Yang et al., 2020]. Their architecture incorporated the squeeze-and-excitation residual (SE-Res) module in the encoder to capture dependencies among feature channels and used a selective kernel block in the decoder to adaptively adjust the receptive field size for gathering multi-scale feature information.

In a recent study, Abdelhamed et al. (2023) proposed NesT-UNet, a 2D segmentation network that leverages the Nested Hierarchical Transformer (NesT) [Zhang et al., 2022b] architecture [Abdelhamed et al., 2023]. Their approach employed the NesT architecture in both the encoder and decoder and utilized self-supervised pre-training for improved segmentation of myocardium and scar. The method achieved results comparable to the state-of-the-art on the EMIDEC dataset.

3.2/ UNCERTAINTY ESTIMATION METHODS

In many machine learning systems, it is crucial to have an understanding of what a model does not know. Deep learning algorithms have made significant advancements in learning complex representations that can effectively map high-dimensional data to various outputs. As a result, deep learning has become a key component in numerous modern applications, enabling state-of-the-art performance. However, most deep learning models lack the ability to represent uncertainty [Kendall et al., 2017b].

To address this limitation, Bayesian deep learning approaches provide a practical framework for capturing and comprehending uncertainty in deep learning models. By incorporating Bayesian principles into deep learning, these approaches enable the modeling of uncertainty, offering insights into the reliability and confidence of predictions. Bayesian deep learning serves as a valuable tool for understanding the limits of a model's knowledge and provides a means to quantify and interpret uncertainty in deep learning models [Gal, 2016, Kendall et al., 2017b].

In Bayesian modeling, there exist two primary types of uncertainty that can be effectively modeled. The first type, known as aleatoric uncertainty, captures the inherent noise present in the observations. This noise can originate from various sources, such as sensor inaccuracies or motion disturbances. Aleatoric uncertainty represents the portion of uncertainty that remains unchanged, regardless of the amount of data collected. It encompasses the irreducible uncertainty associated with the noise in the observations. Within the category of aleatoric uncertainty, there are further classifications. Homoscedastic uncertainty refers to uncertainty that remains constant for different inputs, indicating a consistent level of noise across the entire dataset. On the other hand, heteroscedastic uncertainty depends on the inputs to the model. It accounts for the possibility of certain inputs having noisier outputs compared to others, resulting in varying levels of uncertainty across the dataset [Kendall et al., 2017b].

The second type, epistemic uncertainty, also known as model uncertainty, stems from insufficient knowledge about the true underlying data distribution or limitations in the model structure. It can be attributed to factors such as limited data availability or model misspecification. Unlike aleatoric uncertainty, epistemic uncertainty can potentially be reduced by collecting more data, allowing the model to refine its understanding and provide more accurate predictions [Kendall et al., 2017b].

Bayesian neural networks (BNNs) differ from regular neural networks in that their weights are assigned a probability distribution instead of a single value or point estimate, as shown in Figure 3.19. These distributions represent the uncertainty associated with the weights and allow for the estimation of uncertainty in predictions [Blundell et al., 2015]. By employing a Bayesian approach, we can quantify the epistemic uncertainty inherent in our beliefs using probability distributions. Given training inputs $X = x_1, x_2, ..., x_N$ and their corresponding outputs $Y = y_1, y_2, ..., y_N$, the task is to find a function that maps the input data-points X to the output labels Y. If a Bayesian model, parameterized by the model parameters ω drawn from the initial prior distribution $P(\omega)$, is used to learn this mapping function. The prior distribution ($P(\omega)$) represents our initial belief as to which parameters are likely to have generated our data before we observe any data points. The goal of the learning process is to modify the prior distribution based on the training data in such a way that the model is transformed into an ideal or near-ideal mapping function between X and Y. During the training, Bayes' theorem is used to update the prior distribution of the weights $P(\omega)$ to give the posterior $P(\omega|X, Y)$ based on the data likelihood $P(Y|X, \omega)$ [Gal, 2016, Manivannan, 2020] as shown in the equation below:

$$P(\omega|X, Y) = \frac{P(Y|X, \omega)P(\omega)}{P(Y|X)} \tag{3.13}$$

Where $P(\omega)$ denotes the prior distribution, and $P(\omega|X, Y)$ is the likelihood. The likeli-

hood function captures the probability of observing the data given the model parameters. It describes how well the training data can be modeled with the available model parameters. $P(Y|X)$ is called model evidence. It serves as a normalization factor and describes the principal probability over the training data independent of the parameter choice [Gal, 2016, Wu et al., 2023], and it is computed as:

$$P(Y|X) = \int P(Y|X,\omega)P(\omega)d\omega \tag{3.14}$$

Given a new input point x^*, the output y^* can be inferred by integrating over all values of the posterior probability distribution of ω:

$$P(y^*|x^*,X,Y) = \int P(y^*|x^*,\omega)P(\omega|X,Y)\,d\omega \tag{3.15}$$

These equations highlight the advantage of using a Bayesian model compared to classical non-Bayesian neural networks, which typically provide a single-point estimate. In contrast, a Bayesian model can provide a distributional output for both the model weights and predictions. Distributions are valuable because they allow for easy extraction of uncertainty estimates, such as computing the variance, and point estimates are obtained using various statistical measures like mean, median, or mode. However, obtaining predictive posterior distributions in deep neural networks using simple Bayesian modeling tools is not feasible. The computation of the posterior distribution (Eq. 3.13) and the closed-form integral computation of the predictive distribution (Eq. 3.13), which involves integrating over the space of all possible model parameters, are intractable due to the complex data likelihood function and the presence of numerous weight parameters with countless combinations of values [Gal, 2016, Wu et al., 2023].

To address this issue, approximation methods are necessary to efficiently estimate the posterior distribution of the parameters in Bayesian neural networks (BNNs). These approximation methods provide practical ways to approximate the posterior distribution, making the computations more manageable. Hence, various techniques have been proposed to approximate the posterior. One category of methods involves Markov Chain Monte Carlo (MCMC) techniques [Neal, 1995], which generate samples from the posterior distribution. Another popular set of techniques involves variational inference [Graves, 2011, Gal et al., 2015], where the posterior is approximated by a variational distribution [Mukhoti et al., 2018]. In the following section, we will explore various variational inference-based techniques for approximating the posterior distribution of BNN parameters.

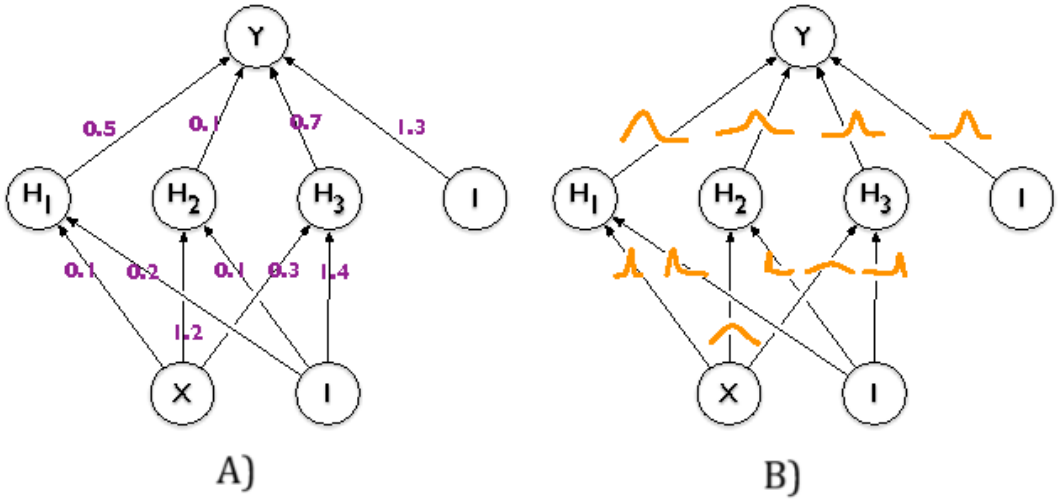

Figure 3.19: A comparison between deterministic Neural Networks (A) that have fixed parameter values, and Bayesian Neural Networks (B) which exhibit a distribution defined over their parameters. Image adapted from [Blundell et al., 2015].

3.2.1/ Variational Inference

In most cases, it is not possible to analytically evaluate the true posterior probability distribution, denoted as $P(\omega|X,Y)$. Instead, we employ an approximation method using a variational distribution, denoted as $q(\omega|\theta)$, which belongs to a known tractable family with a predefined functional form. This can be done by minimizing the Kullback-Leibler (KL) divergence between the variational distribution $q(\omega|\theta)$ and the true posterior $P(\omega|X,Y)$ [Gal, 2016]. Therefore, by finding the optimal parameters θ for the distribution $q(\omega|\theta)$, we aim to minimize the KL divergence between the variational distribution and the true Bayesian posterior distribution of the weights as follows:

$$\begin{aligned}\theta^* &= \underset{\theta}{argmin} \quad KL[q(\omega|\theta)||P(\omega|X,Y)] \\ &= \underset{\theta}{argmin} \int q(\omega|\theta)\log\frac{q(\omega|\theta)}{P(Y|X,\omega)P(\omega)}d\omega \\ &= \underset{\theta}{argmin} \quad KL[q(\omega|\theta)||P(\omega)] - E_{q(\omega|\theta)}[\log P(Y|X,\omega)]\end{aligned} \tag{3.16}$$

The cost function resulting from this approach is commonly referred to as the variational free energy. The negative variational free energy is also known as evidence lower bound (ELBO). It consists of two components: the expected log-likelihood of the data under the variational distribution and the negative KL divergence between the variational distribution and the prior. Maximizing the ELBO encourages the variational distribution to closely approximate the true posterior. The complete derivation of the cost function can be found

in [Blundell et al., 2015], but for simplicity, let's denote it as:

$$F(\theta, X, Y) = KL[q(\omega|\theta)||P(\omega)] - \mathbb{E}_{q(\omega|\theta)}[\log P(Y|X, \omega)] \tag{3.17}$$

where $\mathbb{E}_{q(\omega|\theta)}$ denotes expectation over the variational posterior $q(\omega|\theta)$. The first term in the equation (Eq. 3.17) is the KL divergence between the variational distribution $q(\omega|\theta)$ and the prior $P(\omega)$, often referred to as the complexity cost. The second term is the expected value of the likelihood with respect to the variational distribution and is known as the likelihood cost.

One way to represent the approximate posterior distribution $q(\omega|\theta)$ is by utilizing a fully factorized Gaussian, parameterized by $\theta = (\mu, \sigma)$ where μ is the mean vector of the distribution and σ the standard deviation vector. To ensure both a positive σ value and training stability, σ is parameterized with ρ using softplus function, expressed as $\sigma = \text{softplus}(\rho) = \ln(1 + e^{\rho})$. The prior distribution $P(\omega)$ is typically selected as a fully factorized Gaussian with a mean $\mu_{\text{prior}}I$ and covariance $\sigma_{\text{prior}}I$, where I denotes an identity matrix [Ng et al., 2020].

During a training iteration, two main steps are involved: the forward pass and the backward pass. In the forward pass, a single sample is randomly drawn from the variational posterior distribution. This sample is then utilized to evaluate the cost function as defined in Eq. 3.17. On the other hand, during the backward pass, the gradients of μ and ρ are computed using backpropagation so that their values can be updated by an optimizer. As the forward pass incorporates a stochastic sampling process, it becomes necessary to employ the "re-parameterization trick" to enable successful backpropagation. This training technique is referred to as Bayes by Backprop [Blundell et al., 2015], and its steps are briefly summarized below [Ng et al., 2020]:

(1) For each weight θ, generate a sample ϵ from a standard normal distribution ($\mathcal{N}(0, 1)$) and set $\theta = \mu + \text{softplus}(\rho) \cdot \epsilon$.

(2) Compute the loss according to Eq. 3.17, which involves the negative logarithm of the likelihood $P(Y|X, \omega)$, and include a regularization term represented by $\text{KL}[q(\omega|\theta)||\mathcal{N}(\mu_{\text{prior}}I, \sigma_{\text{prior}}I)]$.

(3) Perform gradient descent to update the values of μ and ρ.

After the optimization process is complete, the trained BNN can be utilized for making predictions and uncertainty estimates. Unlike a traditional neural network that provides a single-point estimate, the BNN generates a distribution of potential outputs. This distribution allows for uncertainty estimation by calculating statistics such as variance or entropy, and the final predictions are obtained using the mean. It is important to note that the effective number of trainable parameters in a BNN is doubled compared to a regular neural

network, as BNNs are parameterized with both μ and σ.

3.2.2/ MONTE CARLO-DROPOUT

Monte Carlo Dropout (MC Dropout) [Gal et al., 2015] is a technique that combines the ideas of dropout regularization [Srivastava et al., 2014] and Monte Carlo sampling [Shapiro, 2003] to approximate the posterior distribution in Bayesian neural networks (BNNs). It can be seen as a form of approximate variational inference.

In standard dropout regularization [Srivastava et al., 2014], during training, random subsets of the neural network's units (neurons) are dropped out or set to zero with a certain probability. This helps prevent overfitting and encourages model robustness. During inference, the dropout is turned off, and the model's predictions are obtained using the full network. In MC-dropout [Gal et al., 2015], a neural network with dropout is trained, and during test time (inference), the dropout is activated to generate Monte-Carlo samples of the prediction. From these Monte-Carlo samples, the uncertainty associated with the output is estimated.

MC Dropout operates based on the principles of variational inference. In MC-dropout, the variational distribution is assumed to be a Bernoulli distribution, and positioning this distribution over the layer's weights with parameter p is considered to be identical to adding a dropout on that layer with a dropout rate of p. To perform approximate inference, it is necessary to train a network with dropout. However, unlike common practice, these dropout layers remain active even during the testing phase. The objective is to obtain samples from the posterior distribution. As the dropout layers introduce a Bernoulli distribution over the network weights, performing a stochastic forward pass through a trained network can be interpreted as generating a Monte Carlo sample from the posterior distribution. Consequently, multiple forward passes using the same input yield multiple Monte Carlo samples, the average of which can be utilized as the network's prediction. The variance can be interpreted as an estimate of uncertainty [Gal et al., 2015, Mukhoti et al., 2018].

3.2.3/ DEEP ENSEMBLE

Deep ensemble [Lakshminarayanan et al., 2016] is a technique used in deep learning that involves training and combining multiple neural networks to improve prediction performance and estimate uncertainty. It is based on the principle of ensemble learning, where multiple models are trained independently, and their predictions are combined to make final decisions.

In deep ensemble, multiple deterministic neural networks with the same architecture are trained using the same data (or different subsets of the same data) with different random

initialization [Lakshminarayanan et al., 2016, Ng et al., 2020]. Each network follows a different training trajectory due to the inherent randomness in the training process. As a result, each network in the ensemble learns a slightly different representation of the data.

During inference, each network in the ensemble independently makes predictions on the input data. The final prediction is obtained by aggregating the individual predictions of all the networks. This aggregation can be done by averaging the predicted probabilities or taking the majority vote, depending on the type of task. To measure the uncertainty, the variance of the predictions made by the individual models are used [Lakshminarayanan et al., 2016]. While deep ensembles are typically not categorized as Bayesian approaches [Lakshminarayanan et al., 2016, Ng et al., 2020] since they do not explicitly model the posterior distribution over network weights like BNNs, there is an increasing body of research [Wilson et al., 2020, Hoffmann et al., 2021] arguing that they can be seen as an approximate Bayesian method. Figure 3.20 shows the difference between MC Dropout and Deep Ensemble. In the case of Deep Ensemble, all the networks in the ensemble possess the same architecture but have been randomly initialized differently for training.

Deep ensemble offers several advantages, including its simplicity of implementation, easy parallelization, minimal hyperparameter tuning, and the ability to generate high-quality predictive uncertainty estimates [Lakshminarayanan et al., 2016]. However, it does come with certain drawbacks. One such drawback is the increased computational and storage demands compared to training a single neural network. Due to the need to train and store multiple networks, both training and inference times can be prolonged. Moreover, ensemble models necessitate additional memory to store the parameters of each individual network. Consequently, in some cases, deep ensembles can be prohibitively expensive to use.

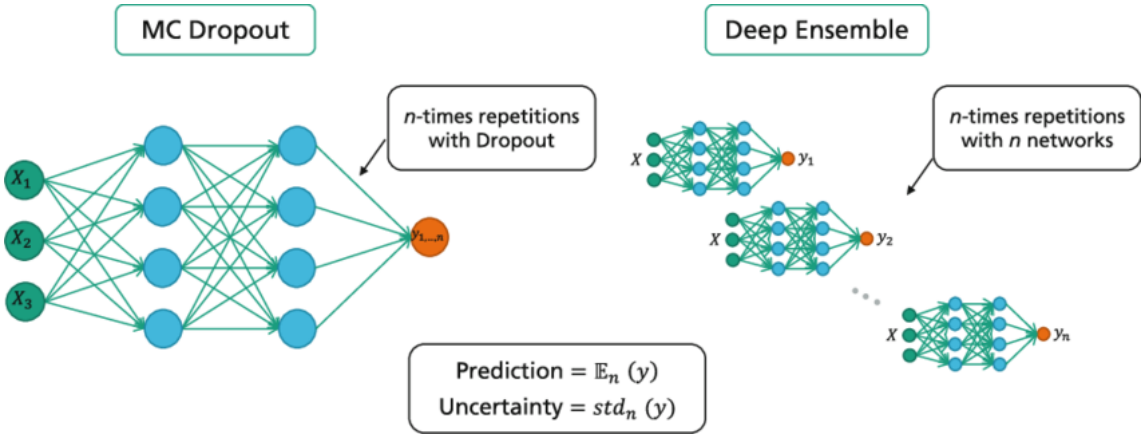

Figure 3.20: Comparison between Monte-Carlo Dropout (left) and Deep Ensemble (right) methods. Image adapted from [Wu et al., 2023].

3.2.4/ Bayesian Deep Learning in medical image analysis

Bayesian deep learning has emerged as a promising approach in the field of medical image analysis, offering significant advancements in both accuracy and uncertainty quantification. Traditional deep-learning methods excel at capturing complex patterns in medical images, but they often lack the ability to provide reliable uncertainty estimates for their predictions. This has become one of the main factors limiting the deployment of deep learning models in clinical practice [Zou et al., 2023]. Bayesian deep learning, on the other hand, combines the power of deep neural networks with probabilistic modeling, enabling a more comprehensive understanding of the underlying uncertainties in medical image analysis tasks. This section delves into the application of Bayesian deep learning in medical image analysis, more particularly in medical image classification and segmentation tasks.

In recent years, there has been a growing interest in incorporating uncertainty theory to establish trustworthy classification in medical images. Various studies have explored this approach in different domains, including fundus retinal, dermoscopic, histopathology, and MR images (a more detailed review can be found from Zou et al. [Zou et al., 2023] and Lambert et al. [Lambert et al., 2022b]).

Leibig et al. [Leibig et al., 2016] evaluated the use of dropout-based Bayesian uncertainty measures for deep learning in diagnosing diabetic retinopathy (DR) from fundus images. Their findings demonstrated that such measures effectively capture uncertainty, outperforming straightforward alternatives. Moreover, they highlighted that incorporating uncertainty-informed decision referrals can enhance diagnostic performance. Ayhan et al. [Ayhan et al., 2019] proposed an intuitive framework based on test-time data augmentation to quantify the diagnostic uncertainty of deep neural networks for diagnosing diabetic retinopathy. They showed that the derived measure of uncertainty is well-calibrated and that experienced physicians also find cases with uncertain diagnoses challenging to evaluate.

Araujo et al. [Araújo et al., 2019] introduced a deep learning-based CAD system for DR grading. Their system supports decision-making by providing a medically interpretable explanation and an estimation of the prediction uncertainty. This allows ophthalmologists to assess the level of trustworthiness associated with a particular decision. Their method adopts a Gaussian sampling approach within a multiple-instance learning framework, enabling the inference of image grades along with explanation maps and prediction uncertainties, even with image-wise labels during training. Filos et al. [Filos et al., 2019] conducted an assessment of different Bayesian deep learning models, including MC-Dropout, variational inference, and deep ensemble, in the context of DR tasks. Their focus was on leveraging model uncertainty for medical pre-screening, where patients are referred to experts when the model's diagnosis is uncertain.

Drawing inspiration from the actual practices of pathologists and the automatic Whole Slide Image (WSI) classification system, [Feng et al., 2022] developed a multi-scale classification framework that integrates predictions and uncertainty estimates from different magnification levels. They employed vision transformers to model both class predictions and uncertainty based on evidential theory. By combining evidence across scales, their approach aims to produce reliable predictions comparable to pathologists' analysis. Dolezal et al. [Dolezal et al., 2022] quantified uncertainty in whole slide classifiers using dropout sampling. By establishing thresholds on training data, they differentiated between low and high-confidence predictions which facilitated uncertainty-aware clinical decision-making. For digital pathology, Linmans et al. [Linmans et al., 2022] suggested the use of multi-head CNNs (multi-head ensembles) to efficiently estimate predictive uncertainty and identify out-of-distribution (OOD) images. Their approach focused on enhancing uncertainty estimation in digital pathology using a multi-head ensemble model.

In the context of skin lesion analysis, Molle et al. [Molle et al., 2019] highlighted the limitations of a variance-based uncertainty metric, which often yields small and difficult-to-interpret values. They proposed a new uncertainty measure that considered the overlap between distributions of different output classes. This metric provided a clear range between 0 (indicating high certainty) and 1 (indicating low certainty), making it easy to interpret. The effectiveness of these metrics was verified in skin lesion classification tasks.

In MR image analysis, Tousignant et al. [Tousignant et al., 2019] introduced an automatic end-to-end Bayesian deep learning framework for predicting future disability progression in patients with Multiple Sclerosis using multi-modal brain MRI. They demonstrated that uncertainty estimates, derived from Monte Carlo dropout sample variance, exhibited a correlation with the model's errors. By providing clinicians with the model's predictions and associated uncertainty estimates, they were able to determine which scans require further examination. Herzog et al. [Herzog et al., 2020] proposed a Bayesian convolutional neural network that predicts the probability of a stroke lesion on 2D MR images while providing uncertainty information about the reliability of the prediction. In the classification of Adamantinomatous Craniopharyngioma from preoperative MRI, Prince et al. [Prince et al., 2023] utilized Variational Inference by elliptical slice sampling to quantify uncertainty. By incorporating this uncertainty-aware deep learning approach, they aimed to provide more reliable and informative classification results for this specific medical condition [Zou et al., 2023].

Bayesian deep learning methods have also been used successfully for medical image segmentation. Nair et al. [Nair et al., 2020] were among the pioneers in exploring multiple uncertainty estimates based on MC dropout for lesion detection and segmentation in medical images. They investigated four voxel-based uncertainty measures and analyzed their performance in voxel-based segmentation and lesion-level detection. Their study re-

vealed that utilizing uncertainty measures consistently led to the selection of superior operating points compared to solely relying on the network's sigmoid output as a probability. In the context of semantic segmentation of polyps from colonoscopy images, Wickstrøm et al. [Wickstrøm et al., 2018] developed MC dropout in FCNs and focused on enhancing model interpretability. Their work aimed to improve the accuracy and reliability of polyp segmentation by incorporating uncertainty estimation. Yu et al. [Yu et al., 2019] introduced MC dropout within a semi-supervised framework and presented an uncertainty-aware model for left atrium segmentation from 3D MR images. Soberanis-Mukul et al. [Soberanis-Mukul et al., 2019] proposed a segmentation refinement method that leveraged uncertainty analysis and graph convolutional networks. They utilized the uncertainty levels obtained from a convolutional network on a specific input volume to formulate a semi-supervised graph learning problem. By training a graph convolutional network, they refined the segmentation results. The effectiveness of their method was validated using the medical segmentation decathlon dataset.

Baumgartner et al. [Baumgartner et al., 2019] presented a segmentation approach inspired by Probabilistic U-Net [Kohl et al., 2018] to capture uncertainty in medical image segmentation. A probabilistic U-Net is a generative segmentation model that combines a U-Net with a conditional variational autoencoder, enabling the generation of an unlimited number of segmentation samples. The final prediction is computed as the mean of these samples, while the uncertainty is determined by their variance. To enhance the diversity of segmentation samples, Baumgartner et al. [Baumgartner et al., 2019] introduced a hierarchical probabilistic model inspired by the Laplacian Pyramids. This model generates image-conditional segmentation samples by initially producing outputs at a low resolution and progressively refining the distribution of segmentations at higher resolutions. The authors evaluated their method, PHiSeg, on two segmentation tasks: thoracic CT images with lesions and prostate MRI. The results demonstrated that PHiSeg generated more realistic and diverse segmentations compared to other similar methods.

In another study, Garg et al. [Garg et al., 2018] proposed a method for exact Markov chain Monte Carlo (MCMC) sampling from generic Bayesian Markov random field (MRF) models. Their approach was built upon Fill's algorithm, a technique for sampling from a Markov chain with the desired distribution as its equilibrium distribution. By extending Fill's algorithm to generic MRF models, Garg et al. [Garg et al., 2018] introduced a novel bounding chain algorithm. The method was evaluated on both simulated data and clinical brain image segmentation tasks, demonstrating its ability to produce more accurate uncertainty estimates compared to other state-of-the-art methods.

An alternative approach to generating uncertainty in medical image segmentation involves utilizing ensembles of deep networks. Mehrtash et al. [Mehrtash et al., 2020] explored the estimation of predictive uncertainty by employing an ensemble of multiple

FCNs. They investigated the effectiveness of these multi-FCN ensembling methods on cardiac and prostate MR image segmentation. Building upon the concept of ensemble learning, Cao et al. [Cao et al., 2020] developed an uncertainty-aware model for semi-supervised breast ultrasound mass segmentation.

In the context of short-axis cardiac MRI segmentation, Guo et al. [Guo et al., 2022] devised a globally optimal label fusion algorithm based on ensemble learning. Kushibar et al. [Kushibar et al., 2022] proposed Layer Ensembles, an uncertainty estimation method that utilizes a single network by ensembling the predictions from different layers or segmentation heads of a deep learning model. They evaluated the effectiveness of this method on mass segmentation from mammogram images and cardiac structure segmentation from cardiac MRI. The results demonstrated competitive performance compared to state-of-the-art Deep Ensembles.

In another study, Zhao et al. [Zhao et al., 2022] introduced an uncertainty estimation method based on a posterior sampling of the weight space for nnU-Net [Isensee et al., 2021], a widely used deep learning model for medical image segmentation. Their method involved ensembling multiple snapshots of the nnU-Net model, saved at different stages during training. The predictions from the ensembled models were used to estimate the uncertainty of the model's predictions. The method was evaluated on two cardiac MRI segmentation datasets (ACDC [Bernard et al., 2018] and M&Ms [Campello et al., 2021]) and showcased improved uncertainty estimation compared to various baseline methods.

3.2.5/ USAGE OF UNCERTAINTY ESTIMATES IN MEDICAL IMAGE ANALYSIS

Uncertainty estimation is critical for building trust in AI systems used in healthcare, particularly in the field of medical image analysis, where early disease detection can be life-changing. Deep learning models for medical image analysis can produce highly accurate predictions, but their reliability is limited without proper uncertainty quantification [Zou et al., 2023]. By providing confidence measures along with predictions, uncertainty estimates can enable the identification of model limitations, quality control, active learning, and informed decision-making by medical professionals.

Incorporating uncertainty estimates into automated predictions enriches the process in multiple ways. Uncertainty estimates can serve as indicators of potential errors or limitations in the medical image analysis system. For example, high uncertainty could indicate anomalies within the input data, a factor vital for Quality Control (QC) and out-of-distribution (OOD) detection [Lambert et al., 2022b].

Additionally, uncertainty maps can direct attention to challenging or pathological regions in medical images that are prone to mistakes. For instance, a model segmenting tumors

in brain MR images may express higher uncertainty where tumors are small or irregularly shaped. By highlighting these uncertain areas, radiologists can focus their review and request additional imaging if needed. The uncertainty estimates thus allow clinicians to make more informed diagnostic decisions.

Active learning is another promising application, where uncertainty guides the selection of informative samples for labeling to improve model performance [Nath et al., 2020]. By prioritizing data with high uncertainty for annotation, the model can be retrained to produce more reliable predictions.

Among the various applications of uncertainty, this thesis specifically focuses on quality control and out-of-distribution detection. By utilizing uncertainty information, the aim is to improve the reliability and trustworthiness of deep learning models, as extensively detailed in Chapters 5 and 6.

3.3/ CONCLUSION

In summary, this chapter has provided a technical overview of machine learning and deep learning concepts, uncertainty modeling approaches, and the state-of-the-art in cardiac MRI analysis. The core machine and deep learning techniques were covered, including neural network architectures, training procedures, and evaluation metrics. Main methods for estimating uncertainty in deep learning models were reviewed, as well as their emerging use in medical image analysis. Recent literature applying deep learning to advance cardiac MRI segmentation and analysis was discussed. Collectively, this background establishes the foundation required to present this thesis' contributions around developing reliable cardiac MRI analysis through novel deep learning approaches and uncertainty-based techniques. The next chapters will build upon the concepts, methods, and state-of-the-art surveys presented here to introduce the proposed methods for robust and reliable cardiac MRI analysis.

II

CONTRIBUTIONS

4

Leveraging Uncertainty Estimates to Improve Segmentation Performance in Cardiac MR

In medical image segmentation, several studies have used Bayesian neural networks to segment and quantify the uncertainty of the images. These studies show that there might be an increased epistemic uncertainty in areas where there are semantically and visually challenging pixels. The uncertain areas of the image can be of great interest as they can possibly indicate the regions of incorrect segmentation. In this chapter, we propose a segmentation model that incorporates uncertainty into its learning process to leverage the uncertainty information. Firstly, we generate the uncertainty estimate (sample variance) using Monte-Carlo dropout during training. Then we incorporate it into the loss function to improve the segmentation accuracy and probability calibration. The proposed method is validated on the publicly available EMIDEC MICCAI 2020 dataset and LAScarQS MICCAI 2022 which mainly focuses on the segmentation of infarcted myocardium and left atrial (LA) scars from Late Gadolinium Enhancement (LGE) MRI.

4.1/ Introduction

Cardiac magnetic resonance imaging with late gadolinium enhancement (LGE-CMR) is the gold standard for quantifying myocardial infarction caused by interrupted coronary blood supply. LGE-CMR enables precise visualization and quantification of the infarcted tissue resulting from this irreversible myocardial damage [Kate Meier et al., 2009]. The no-reflow phenomenon is an incident that usually appears in a proportion of patients with acute myocardial infarction following re-perfusion therapy of an occluded coronary artery

[Abbas et al., 2015].

Recent deep learning methods for automatic myocardial scar segmentation from LGE images have utilized multi-stage cascaded frameworks. Approaches include Zabihollahy et al. [Zabihollahy et al., 2018] who used manual myocardium segmentation followed by a 2D FCN, and Zhang [Zhang, 2020], Ma [Ma, 2020a] and Girum *et al.* [Girum et al., 2020] who employed two-stage nnUNets to coarsely segment the myocardium and refine the scar segmentation. Arega *et al.* [Arega et al., 2020] also developed a three-network cascaded system. While these methods have demonstrated promising performance on datasets like EMIDEC, the main problem with these cascaded and complex methods is that they can be time-consuming and computationally expensive.

As part of the Left Atrial and Scar Quantification and Segmentation Challenge (LAScarQS 2022), several techniques were proposed by different challengers to segment the Left Atrial (LA) cavity and scar from LGE MRI. Tu et al. [Tu et al., 2022] adopted a self-pre-training paradigm, combining Mask Autoencoder (MAE) and Vision Transformers (ViT), to learn contextual information as priors from the LGE-MRI dataset before fine-tuning the segmentation task. Zhang et al. [Zhang et al., 2022a] introduced a TopK loss focused on boundary pixels for better LA cavity delineation and a distance map input to constrain scar locations. Mazher et al. [Mazher et al., 2022] developed a semi-supervised segmentation approach using pseudo labeling for improved left atrial and scar segmentation from LGE MRI. Their method involved generating pseudo labels using a 3D ResUNet model on training and validation data, which were then used alongside true labels to train the nnUNet model for the final segmentation. Liu et al. [Liu et al., 2022] proposed the UG-former framework, which integrates transformers, graph convolutional networks (GCN), and convolutional decoders for LA scar segmentation. Their approach employed enhanced transformers with deformable convolutions to capture irregular shapes and GCN bridges to improve generalization across images from different scanners.

Bayesian deep learning has been used in segmentation tasks to provide a prediction as well as quantify the uncertainty associated with each prediction. Recently, several studies have employed Monte Carlo Dropout to estimate uncertainty for medical image segmentation [Mehrtash et al., 2020, Nair et al., 2020, Ng et al., 2020, Roy et al., 2018, Sander et al., 2019]. Nair *et al.* [Nair et al., 2020] explored MC dropout-based uncertainty estimates for multiple sclerosis lesion detection and segmentation. They improved the segmentation results by filtering and excluding the most uncertain voxels. Similarly, Sander *et al.* [Sander et al., 2019] applied the MC Dropout based method for cardiac MRI segmentation and showed that the uncertainty maps are close to the reported segmentation errors and they improved the segmentation results by correcting the uncertain pixels. These previous studies [Nair et al., 2020, Roy et al., 2018, Jungo et al., 2019, Sander et al., 2019, Mehrtash et al., 2020] mostly focused on the correlations between

predictive uncertainty and the segmentation accuracy and how the uncertainty metrics can be used to improve the segmentation by filtering the most uncertain predictions. However, these methods did not leverage the uncertainty information during training to enhance the segmentation result.

In this work, we proposed a segmentation model that generates uncertainty estimates during training using MC dropout. Then it leverages these uncertainty estimates to improve the segmentation results by incorporating them into the loss function. Uncertainty information can possibly indicate the regions of incorrect segmentation [Sander et al., 2019, Wang et al., 2019]. We hypothesized that incorporating this information as part of the learning process can help the network to improve the segmentation results by correcting the segmentation errors that have high epistemic uncertainty. The proposed method was evaluated on the publicly available EMIDEC MICCAI 2020 [Lalande et al., 2020] and LAScarQS MICCAI 2022 [Li et al., 2021, Li et al., 2022b, Li et al., 2022c] datasets. It achieved state-of-the-art results outperforming the top-ranked methods of both challenges. The experimental results showed that the uncertainty information was indeed beneficial in enhancing the segmentation performance. We also observed that the improvements were more significant in the semantically and visually challenging images which have higher epistemic uncertainty. Assessing the probability calibration, we showed that the proposed method produced more calibrated probabilities than the baseline method.

4.2/ Materials and Methods

4.2.1/ Materials

4.2.1.1/ EMIDEC

The Automatic Evaluation of Myocardial Infarction from Delayed-Enhancement Cardiac MRI challenge (EMIDEC)[1] is a MICCAI 2020 challenge that focuses on cardiac MRI segmentation. The dataset consists of LGE images of 100 patients for training. Of these cases, 67 are pathological cases and the remaining 33 are normal cases. The testing set includes 50 patients of which 33 are pathological and 17 are normal cases. Each case has 5 to 10 short-axis slices covering the left ventricle from base to apex with the following characteristics: slice thickness of 8 mm, the distance between slices of 10 mm, and spatial resolution ranging from $1.25 \times 1.25\ mm^2$ to $2 \times 2\ mm^2$ [Lalande et al., 2020]. As a pre-processing step, we normalized the intensity of every patient image to have zero mean and unit-variance and we resampled all the volumes to have a voxel spacing of

[1] http://emidec.com/

$1.458mm \times 1.458mm \times 10.0mm$.

4.2.1.2/ LAScarQS 2022

The Left Atrial and Scar Quantification and Segmentation Challenge (LAScarQS 2022) [2] consists of 200 LGE MRIs acquired in a real clinical environment from patients suffering Atrial fibrillation (AF). All the LGE MRIs were collected from three different clinical centers. The images from the first center (University of Utah) were acquired using Siemens Avanto 1.5T or Vario 3T. The voxel resolution of the images was $1.25 \times 1.25 \times 2.5$ mm. The LGE MRIs from the second center (Beth Israel Deaconess Medical Center) were acquired with Philips Achieva 1.5T. The spatial resolution of the images was $1.4 \times 1.4 \times 1.4$ mm. Similar to the second center, the images from the third center (King's College London) were acquired with a Philips Achieva 1.5T. The spatial resolution of the LGE MRI scan was $1.3 \times 1.3 \times 4.0$ mm. The challenge focuses on the segmentation of left atrial blood pool and left atrial scar [Li et al., 2021, Li et al., 2022b, Li et al., 2022c].

4.2.2/ METHODS

Various Bayesian deep learning methods are used to estimate uncertainties in images. Among the most widely used Bayesian deep learning methods in medical images is Monte-Carlo dropout (MC-dropout). In MC-dropout, a network with dropout is trained, and then during testing the network is sampled N times in order to get N segmentation samples. From these N segmentation samples, the uncertainty measure (sample variance) is computed. In our method, we used MC dropout during training in order to get the uncertainty estimates. During training, the model is sampled N times and the mean of these samples is used as the final segmentation as can be seen from Figure 4.1. The uncertainty metric is computed from the N Monte-Carlo dropout samples. It can be calculated per pixel or per structure [Ng et al., 2020]. In this research, we used the pixel-wise uncertainty and image-level uncertainty. Pixel-wise uncertainty is computed per pixel. Sample variance is one of the pixel-wise uncertainty measures. It is calculated as the variance of the N Monte-Carlo prediction samples of a pixel. Each pixel i has N sigmoid predictions $(y_{i,1}...y_{i,N})$. From these predictions, the mean μ_i is computed (Eq. 4.1). In Eq. 4.2, σ_i^2 is the sample variance of each pixel i of the image [Nair et al., 2020]. In order to compute the image-level uncertainty, the per-pixel uncertainty is averaged over all pixels of the image as shown in Eq. 4.4. In this equation, I is the total number of pixels of the image.

[2] https://zmic.fudan.edu.cn/lascarqs22

$$\mu_i = \frac{1}{N}\sum_n (y_{i,n}) \tag{4.1}$$

$$\sigma_i^2 = \frac{1}{N}\sum_n (y_{i,n} - \mu_i)^2 \tag{4.2}$$

As stated by [Sander et al., 2019] and [Wang et al., 2019], uncertainty information indicates potential mis-segmentations and the most uncertain part of the segmentation results covers regions of incorrect segmentations. In order to leverage this uncertainty information, we proposed to include it as part of the loss function so that the network will learn to correct the possible mis-segmentations. Hence, the total loss is computed as a sum of the segmentation loss and uncertainty loss as can be seen from Figure 4.1. The segmentation loss is the weighted average of cross-entropy (CE) loss and Dice loss (Eq. 4.3). For the uncertainty loss, we first computed the image level uncertainty (Eq. 4.4). Then, it is added to the segmentation loss with a hyper-parameter value alpha (α) that controls the contribution of the uncertainty loss to the total loss (Eq. 4.5).

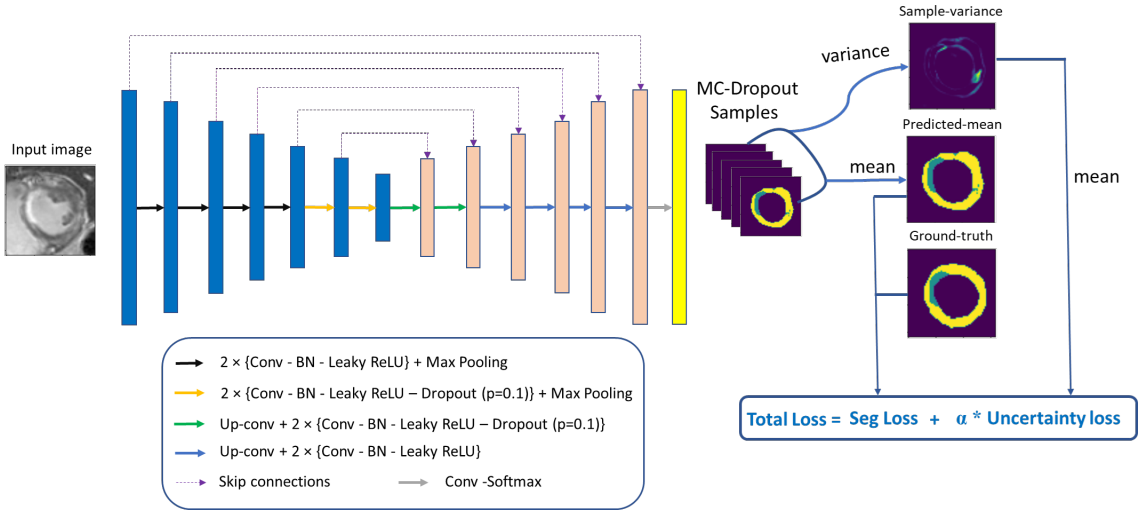

Figure 4.1: The proposed method

$$L_{Seg} = \lambda_{Dice} L_{Dice} + \lambda_{CE} L_{CE} \tag{4.3}$$

$$L_{Uncertainty} = \frac{1}{I}\sum_i (\sigma_i^2) \tag{4.4}$$

$$L_{Total} = L_{Seg} + \alpha \times L_{Uncertainty} \tag{4.5}$$

For the segmentation network, we used a 3D U-Net [Isensee et al., 2021] architecture with dropout placed at the middle layers of the network (Figure 4.1) as suggested by the literatures [Kendall et al., 2015, Fortunato et al., 2017, Blundell et al., 2015, Ng et al., 2020]. The dropout rate was set at 0.1. The U-Net's encoder and decoder con-

sists of 8 convolutional layers where each convolution is followed by batch normalization and Leaky ReLU (negative slope of 0.01) activation function.

4.2.3/ Implementation

The weights of the segmentation network are optimized using Stochastic gradient descent (SGD) with nesterov momentum ($\mu = 0.99$) with an initial learning rate of 0.01. The mini-batch size was 5 and the model was trained for 1000 epochs on a five-fold cross-validation scheme. For the segmentation loss, we set a weighting factor of 1.0 for Dice loss and 1.0 for CE loss as they provided the best results. In order to generate the segmentation uncertainty (sample variance), we used 5 Monte Carlo samples (the N value in Eq. 5.1). The weighting factor (α) for the uncertainty loss (in Eq. 4.5) is empirically selected to be 3.0 after experimenting with different weighting factors. The training was done on NVIDIA Tesla V100 GPUs using Pytorch deep learning framework based on nnU-Net implementation [Isensee et al., 2021].

4.3/ Results and Discussions

To evaluate the segmentation results, we used geometrical metrics such as the Dice coefficient (DSC) and Hausdorff distance (HD). In addition, for the EMIDEC dataset, we computed clinical metrics that are commonly used in cardiac clinical practice. These include the average volume error (VD) of the left ventricular myocardium (in cm^3), the volume (in cm^3) and percentage (PD) of infarction and no-reflow [Lalande et al., 2020].

To measure the probability calibration of the models, we used the Brier Score (BS). Brier score measures how close the predicted segmentation probabilities are to their corresponding ground truth probabilities (one-hot encoding of each class) by computing the mean square error of the two probabilities [Ng et al., 2020]. To compare image level uncertainties among the segmentation results, we utilized Dice agreement within MC samples (*DiceWithinSamples*) [Roy et al., 2018, Ng et al., 2020]. It is the average Dice score of the mean predicted segmentation and the individual N MC prediction samples. Note that *DiceWithinSamples* is inversely related to uncertainty.

4.3.1/ Ablation Study

To evaluate the effect of adding uncertainty information to the segmentation loss, we compared the model that uses only segmentation loss which is called *baseline* with the model that uses combined loss of segmentation loss and uncertainty loss which is referred to as *proposed*. Both networks have the same architecture and the comparison is done on the

Table 4.1: Comparison of myocardium and scar (infarction) segmentation performance of the baseline method and the proposed method in terms of geometrical and clinical metrics obtained on the EMIDEC test set (50 cases). The values mentioned are mean (standard deviation). The best results are in bold. VD is the volume error. For DSC, the higher the value the better whereas for HD, Brier score (BS), and VD the lower is the better.

Method	**Myocardium**			**Infarction**		
	DSC (%)	**HD** (mm)	**BS** (10^{-2})	**DSC** (%)	**VD** (cm^3)	**BS** (10^{-2})
Baseline	88.0 (2.63)	12.1(7.79)	4.03 (2.45)	65.0 (29.7)	3.04 (5.0)	1.19 (1.81)
Proposed	**88.2 (2.55)**	**11.8 (7.26)**	**3.86 (2.8)**	**67.6 (28.8)**	**2.99 (4.55)**	**1.18 (1.83)**

test dataset. For the ablation study, most of the comparisons are done on the main two classes which are healthy myocardium and infarction.

As can be seen from the table 4.1, the addition of uncertainty information into the segmentation loss enhanced the segmentation accuracy. It increased the DSC of scar (infarction) by 3% and that of myocardium by around 0.2%. It also improved the HD and the average volume error of both scar and myocardium. The segmentation enhancement is more significant on the scar than on the myocardium. This can be explained by the fact that the scar has more irregular shape, smaller area, and visually challenging pixels which may result in higher uncertainty compared to the myocardium (Figure 4.2 (b)).

The apical and basal slices of the left ventricle are more difficult to segment than mid-ventricular images even for human experts [Bernard et al., 2018, Petitjean et al., 2011]. Particularly at the apical slices, the MRI resolution is so low that it is even difficult to resolve the size of small structures (first row in Figure 4.3). Assessing the segmentation performance and uncertainties at different slice positions of the left ventricle, it can be observed that the apical slices have the highest epistemic uncertainty (lowest *DiceWithinSamples*) among the slices (Figure 4.2 (b)). Similarly, in the comparison of segmentation performance, most of the improvements due to the addition of uncertainty information (proposed method) are predominantly on the apical slices (Figure 4.2 (a)). The DSC increased by 2% for scar and by almost 1% for myocardium in the apical slices. While the segmentation performance of the proposed method at the mid and basal slices is similar or slightly better than the baseline method. This tells us that the addition of uncertainty information to the loss function is more advantageous to the semantically and visually challenging images which generate higher epistemic uncertainty. This confirms our initial assumption about the proposed method.

Figure 4.3 shows examples of the segmentation results of baseline and proposed method at apical, mid-ventricular, and basal slices. At the apical slice, one can see that the segmentation result of the baseline method has a lot of errors. In the generated uncertainties (sample variance), the incorrectly segmented regions have higher uncertainty. The pro-

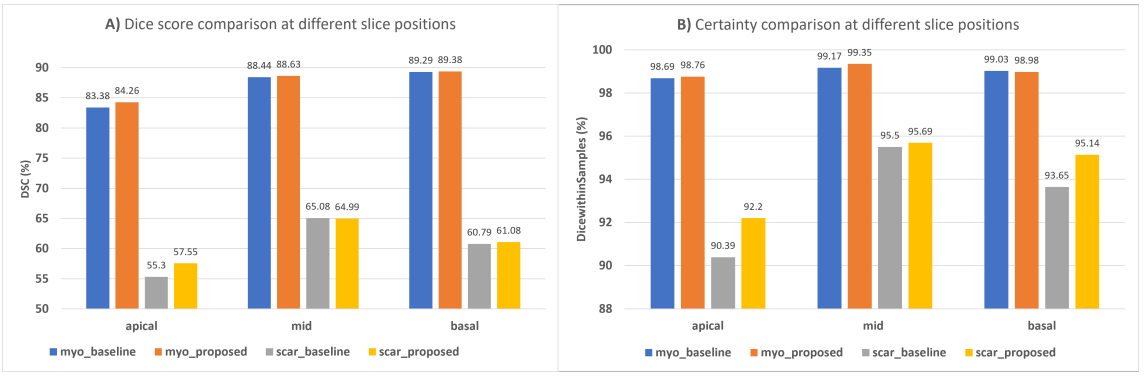

Figure 4.2: Dice score (A) and certainty (B) comparison of the baseline and proposed method at different slice locations. Myo_baseline and Scar_baseline refer to the myocardium and scar Dice score or certainty of the baseline method respectively. Similarly, Myo_proposed and Scar_proposed refer to myocardium and scar Dice score or certainty of the proposed method.

posed method, which utilizes the sample variance as part of the loss, minimized the segmentation errors of the baseline. Similarly, our proposed method produced more robust segmentation results at the mid and basal slices. From the results, we can say that the uncertainty captures relevant information that can be leveraged to improve the segmentation result.

Regarding probability calibration, the proposed method produced more calibrated probabilities than the baseline method on both the myocardium and scar as it yielded a lower Brier score. This suggests that using MC-dropout during training and the addition of uncertainty information to the loss can improve not only the segmentation accuracy but also the calibration of the probabilities.

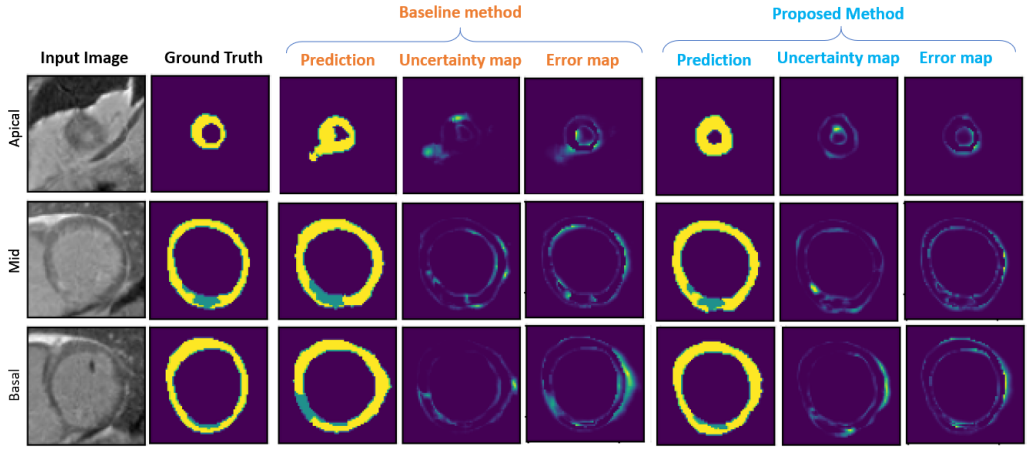

Figure 4.3: Qualitative results comparison of the proposed method with the baseline on a typical cardiac MRI. The generated uncertainty is the sample variance. Scar (green) and myocardium (yellow).

Table 4.2: Comparison of segmentation performance with state-of-the-art methods on EMIDEC challenge's test set (50 cases). Bold results are the best.

Methods	**Myocardium**			**Infarction**			**NoReflow**		
	DSC (%)	**VD** (cm^3)	**HD** (mm)	**DSC** (%)	**VD** (cm^3)	**PD** (%)	**DSC** (%)	**VD** (cm^3)	**PD** (%)
[Zhang, 2020]	87.86	9.26	13.01	**71.24**	3.12	2.38	78.51	0.635	0.38
[Ma, 2020a]	86.28	10.2	14.31	62.24	4.87	3.50	77.76	0.830	0.49
[Feng et al., 2020]	83.56	15.2	33.77	54.68	3.97	2.89	72.22	0.883	0.53
[Yang et al., 2020]	85.53	16.5	13.23	62.79	5.43	4.37	60.99	1.851	1.63
[Hüllebrand et al., 2020]	84.08	10.87	18.3	37.87	6.16	4.93	52.25	0.953	0.64
[Zhou et al., 2020]	82.46	13.29	83.42	37.77	6.10	4.71	51.98	0.879	0.54
[Girum et al., 2020]	80.26	11.81	51.48	34.00	11.52	8.58	78.00	0.891	0.51
Proposed	**88.22**	**7.75**	**12.87**	67.89	**2.59**	**2.06**	**81.25**	**0.487**	**0.32**

4.3.2/ COMPARISON WITH STATE-OF-THE-ART METHODS

Table 4.2 shows the comparison of the proposed method with state-of-the-art methods on the EMIDEC challenge. One can observe that the proposed method outperformed the state-of-the-art methods on most of the geometrical and clinical metrics. Our proposed method yielded much better results in all metrics than Feng *et al.* [Feng et al., 2020], which used a dilated 2D U-Net. Zhang [Zhang, 2020] and Ma [Ma, 2020a] employed a nnU-Net-based segmentation pipeline which is similar to the proposed method's pipeline. However, the proposed method, which utilizes a novel loss function that took into account the uncertainty generated during training, outperformed these two top-ranked methods. In the segmentation of infarction, the proposed method reduced the average volume error from 3.12 cm^3 to 2.99 cm^3 and the percentage from 2.38% to 2.29% compared to Zhang's [Zhang, 2020] method. In terms of the Dice score of infarction, Zhang's [Zhang, 2020] method achieved better results, however, this was obtained using a two-stage cascaded framework which is more computationally expensive framework.

Table 4.3 presents a performance comparison of our proposed method with state-of-the-art approaches on the LAScarQS challenge's test set. Our method outperformed all others and emerged as the winner of the challenge during MICCAI 2022. Despite the other methods [Mazher et al., 2022, Tu et al., 2022, Liu et al., 2022, Zhang et al., 2022a] employing more complex and cascaded techniques, such as semi-supervised segmentation with pseudo labeling and self-supervised pre-training, our relatively simple approach, incorporating uncertainty information into the loss function, achieved Dice scores of 59.5% for LA scar and 94.7% for cavity segmentation. These scores surpassed the state-of-the-art methods by a notable margin of 3%. A more detailed comparison of the proposed method on the LAScarQS challenge can be found in the Appendix A.

Table 4.3: Comparison of segmentation performance with state-of-the-art methods on the LAScarQS challenge's test set (25 cases). Bold results are the best.

Methods	**Dice LA Scar (%)**	**Dice LA Cavity (%)**
[Zhang et al., 2022a]	56.1	90.8
[Liu et al., 2022]	54.9	92.2
[Tu et al., 2022]	47.8	82.8
[Mazher et al., 2022]	56.7	89.4
Proposed	**59.5**	**94.7**

4.4/ CONCLUSION

In this work, we proposed a segmentation model that generates uncertainty estimates during training using the MC-dropout method and utilizes the uncertainty information to enhance the segmentation results by incorporating it into the loss function. The proposed method was evaluated on the publicly available EMIDEC and LAScarQS datasets. It achieved state-of-the-art results outperforming the top-ranked methods of both challenges. Assessing the segmentation performance of the proposed method at different slice positions, we observed that the Dice scores of the more challenging apical slices increased much more than the other slice positions. Furthermore, the improvements in the more difficult scar segmentation were higher than those of myocardium segmentation. In the quantitative and qualitative results, we demonstrated that the uncertainty information was indeed advantageous in enhancing the segmentation performance and the improvements were more significant at the semantically and visually challenging images which have higher epistemic uncertainty. In addition, the proposed method produced more calibrated segmentation probabilities.

The main limitation of our method is that it takes more time to train than the baseline method as it uses MC dropout during training to generate the uncertainty estimates. However, once it is trained, the inference time is exactly the same as the baseline method.

5

Uncertainty-based quality control in Cardiac MR image analysis

In Chapter 4, we proposed a segmentation model that incorporates uncertainty information to improve scar tissue delineation from late gadolinium enhancement MRI. Our method leverages uncertainty estimates from Bayesian neural networks to enhance segmentation performance. While this technique shows promise for identifying small, ambiguous regions like scars, it does not safeguard against outright segmentation failures that could occur on difficult or unusual cases. Before analyzing derived measures like myocardial tissue quantification, we must first ensure the segmentations are sufficiently accurate. This motivates the development of quality control techniques to detect flawed segmentations, as explored in this Chapter. By proposing an automated uncertainty-based quality control framework, we can identify inaccurate cardiac MR segmentation results and exclude them prior to the downstream tasks.

5.1/ Introduction

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, responsible for 31% of all global deaths according to the World Health Organization [WHO, 2017]. Cardiac magnetic resonance (CMR) imaging is increasingly used to assess CVDs. T1 mapping CMR quantifies diffuse myocardial fibrosis and characterizes tissues by assigning T1 relaxation time values to each pixel [Haaf et al., 2016]. Native T1 mapping without contrast is sensitive to disease processes that alter myocardial tissue composition. Changes in native T1 values can indicate both primary cardiac conditions like infarction and systemic diseases like amyloidosis [Moon et al., 2013, Haaf et al., 2016]. Contrast-enhanced T1 mapping reflects extracellular characteristics and is used to calculate extra-

cellular volume (ECV) [Messroghli et al., 2017], which measures extracellular space size, as described in detail in Section 2.4.1.

5.1.1/ Related work

To analyze the T1 mapping and ECV values of patients with CVDs, the regions of interest (ROIs) are manually drawn on native and post-contrast T1 images in the blood pool, septum, and free wall of the left ventricle [Nakamori et al., 2018, Ali et al., 2021, Thongsongsang et al., 2021]. Using manual segmentation of ROI, [Thongsongsang et al., 2021] computed native T1 and ECV for patients with distinct types of myocardial disease, including amyloidosis, dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), myocarditis and coronary artery disease (CAD) to determine the diagnostic yield and cut-off values of T1 and ECV to differentiate myocardial diseases and CAD from their control group. Similarly, [Nakamori et al., 2018] and [Ali et al., 2021] manually segmented the ROIs to evaluate the clinical application of native and post-contrast T1 mapping in assessing diffuse myocardial fibrosis in patients with HCM and DCM respectively. However, manual segmentation is tedious and likely to suffer from intra- and inter-observer variability, particularly when the images are very challenging.

Few works have focused on deep learning-based segmentation of cardiac structures from T1 mapping images. [Fahmy et al., 2019a] proposed to automatically segment the left ventricular myocardium from T1 mapping images. Similarly, [Hann et al., 2021] presented quality-controlled segmentation of cardiac MRI T1 mapping using deep ensembles. However, their work did not include the segmentation and analysis of myocardial T1 and ECV values. [Puyol-Antón et al., 2020] automatically segmented the left ventricular blood pool, myocardium, and right ventricular blood pool from native T1 mapping images. They also analyzed the corresponding global and regional myocardial T1 values and used them to characterize the myocardial tissues of different cardiomyopathies. Even though the study used a very large cohort, it did not include the post-contrast T1 mapping images, which are necessary to compute the ECV values.

Both CNN-based and transformer-based segmentation networks have achieved state-of-the-art results in medical image segmentation and in some cases achieving results surpassing expert-level segmentation performances [Bernard et al., 2018, Tang et al., 2022]. Recently, the interest in deploying automatic segmentation frameworks into the clinical routine has increased. However, these methods can generate incorrect segmentation results, which can lead to wrong clinical decisions in the downstream tasks. To avoid this, experts examine the quality of the segmentation results, but this is a time-consuming and very laborious task. As a solution, automatic quality control-based

methods have been proposed. Some studies [Zhang et al., 2006, Kohlberger et al., 2012] have tried to predict segmentation quality using hand-crafted features of the images and segmentation maps. [Valindria et al., 2017] proposed Reverse Classification Accuracy (RCA). It takes predicted segmentation from a new image to train a reverse classifier. This is then evaluated on a set of reference images with ground truth to determine the quality of the segmentation. The method achieved good Dice regression results. However, it was very slow as the time taken to process a single segmentation result was 11 minutes. [Robinson et al., 2018] proposed a CNN based method to regress the Dice score from the image and segmentation pair on a large cohort. Even though it was computationally less expensive than RCA, it did not exploit uncertainty information which can be useful in determining the segmentation quality [Devries et al., 2018b].

Other works have leveraged uncertainty information to estimate the quality of segmentation results. [Devries et al., 2018b] and [Chen et al., 2020c] proposed a quality control method that utilizes a CNN-based QC to determine the quality of the segmentation output from dermoscopic and CT images, respectively. As an input to the classifier, they used the image with its corresponding segmentation and uncertainty maps. However, directly using the image, segmentation, and uncertainty map may not correlate well with the segmentation quality as it is shown in this study. [Puyol-Antón et al., 2020] proposed a two-stage uncertainty-based QC method. To train their QC method, they manually labeled the outputs of the segmentation model as correct or incorrect segmentation. In the first step, they utilized evidence lower bound (ELBO) based thresholding to reject wrong segmentations, and in the second step, a deep learning image classifier was used to classify the segmentation results as correct or incorrect. Using manually labeled segmentation outputs to train the QC method can increase the sensitivity of the QC method; however, obtaining the manual annotations is time-consuming and expensive.

In this chapter, we develop a fully automated and quality-controlled quantification of myocardial tissue characteristics using native myocardial T1 and ECV. To analyze and quantify the myocardial tissues, we first automatically segment the blood pool and the myocardium of the left ventricle and the blood pool of the right ventricle from native and post-contrast T1 images using a Bayesian Swin transformer-based segmentation network. To detect and reject inaccurate segmentation results before further analysis, we introduce uncertainty-based quality control (QC). The QC method uses image-level uncertainty features as input to a random forest-based classifier/regressor to estimate the segmentation result's quality. Using the proposed method, we automatically compute the mean myocardial native T1 and ECV values of healthy and pathological subjects. We also analyze the ability of these values in differentiating a healthy group from cardiac pathological groups.

5.1.2/ CONTRIBUTIONS

In this chapter, we propose a novel uncertainty-based quality control for segmentation to reduce the inclusion of failed segmentations in subsequent analysis such as native myocardial T1 and ECV analysis. The main contributions of our work are:

- We introduced MC-Dropout-based uncertainty estimation to Swin-transformer-based U-Net and systematically studied MC-Dropout in terms of architecture choice (dropout position and amount). We found that the dropout at the multi-layer perceptron (MLP) module of the transformer block gives a good segmentation accuracy and calibration while providing uncertainty estimates and regularization compared to other dropout variants.
- To decrease the effect of inaccurate segmentation on downstream tasks, we proposed uncertainty-based quality control that leverages simple image-level uncertainty metrics to determine the segmentation quality using a random forest classifier/regressor. In this research, we showed that training a classifier using simple inputs that are derived from uncertainty metrics can determine segmentation quality better than the ones that directly use the image, segmentation, and uncertainty map.
- To the best of our knowledge, this is the first study that automatically computes ECV values from native and post-contrast T1 mapping images. From the automatic quality-controlled analysis of T1 mapping and ECV values, we showed that these values can be used to characterize the myocardial tissues of various cardiac diseases.

5.2/ MATERIAL AND METHODS

5.2.1/ MATERIAL

The dataset used for this research is comprised of native T1 mapping and post-contrast T1 mapping CMR images of 295 subjects, of which 31 of them are normal (healthy), and 264 are pathological. The cardiac pathologies include amyloidosis $(n = 6)$, dilated cardiomyopathy (DCM) $(n = 71)$, hypertrophic cardiomyopathy (HCM) $(n = 30)$, myocarditis $(n = 48)$, Tako-Tsubo syndrome $(n = 5)$, and myocardial infarction $(n = 70)$. In addition, there are subjects that have undetermined or complex pathologies $(n = 34)$. The images were collected from different clinical centers in France. Each image was acquired using a Siemens 1.5T MRI scanner. Modified Look-Locker inversion recovery (MOLLI) was utilized to capture the native and post-contrast T1 mapping images. Each patient has three

short-axis slices for each modality (native T1 mapping and post-contrast T1 mapping) which are apical, mid, and basal slices.The slices of native and post-contrast T1 mapping images are realigned according to the center of gravity of the area defined by the manually drawn epicardial contour of the left ventricle. The manual annotation of each case's left ventricular blood pool, myocardium, and right ventricular blood pool are available.

5.2.2/ METHODS

The proposed pipeline consists of three parts, as shown in Figure 5.1. The first one involves segmentation of left ventricular and right ventricular blood pools, and left ventricular myocardium from native and post-contrast T1 mapping images using Bayesian Swin transformer-based U-Net. The Bayesian segmentation model outputs not only the segmentation result but also uncertainty maps. In the second part, to detect the poorly segmented images from the model, we propose an automated quality control (QC) method that utilizes image-level uncertainty metrics generated by the Bayesian model to estimate the quality of the segmentation result. The final part is focused on the automatic analysis of native myocardial T1 mapping and ECV values of the images that were categorized as good quality images by the proposed QC.

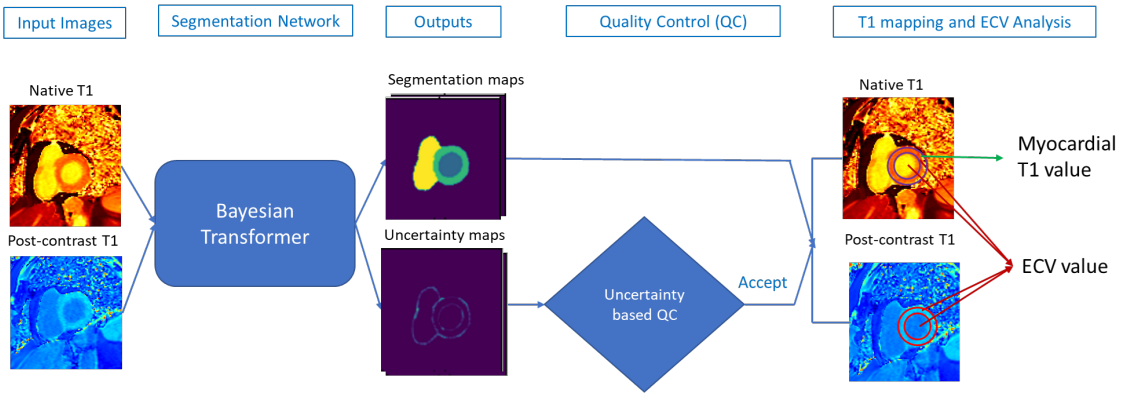

Figure 5.1: Proposed pipeline for automatic quality controlled T1 mapping and ECV analysis from native and post-contrast T1 mapping images. First, the cardiac structures are segmented from native and post-contrast T1 mapping images using the Bayesian segmentation model. Then the quality of the segmentation output is assessed using the uncertainty-based QC. Myocardial T1 and ECV values of the good-quality images are analyzed.

5.2.2.1/ SEGMENTATION

The segmentation network used is a Swin transformer-based U-Net [Liu et al., 2021, Cao et al., 2021]. It is a pure transformer-based model that applies the Swin transformer block to both the encoder and decoder parts of the network.

In the segmentation architecture (Figure 5.2), the input image $(H \times W)$ is divided into non-overlapping patches of size $p \times p$, where we use 4×4 patches. Each patch is linearly embedded into C dimensions. The encoder consists of three stages, each with two consecutive Swin transformer blocks and a patch merging layer. The first stage applies shifted window-based multi-head self-attention (W-MSA) and shifted window MSA to extract feature representations, while the patch merging layer reduces the resolution by concatenating 2×2 neighboring patches. This process is repeated in subsequent stages, resulting in output resolutions of $H/8 \times W/8 \times 2C$, $H/16 \times W/16 \times 4C$, and $H/32 \times W/32 \times 8C$. The bottleneck employs two consecutive Swin transformer blocks for further transformation.

The decoder also has three stages, each comprising a patch-splitting layer and two consecutive Swin transformer blocks. The patch-splitting layer upsamples the bottleneck's feature representations to $H/16 \times W/16 \times 4C$. The upsampled features are then transformed using the transformer blocks, and skip connections concatenate encoder features with the upsampled decoder features. The final layer uses a linear layer to project the feature representations and obtain the segmentation predictions [Liu et al., 2021, Cao et al., 2021].

Dropout at test time enables us to approximate the posterior distribution of the weights by sampling from the Bernoulli distribution across the network's weights [Gal et al., 2016, Kendall et al., 2017a]. The Bayesian Swin transformer-based U-Net has dropout layers at different positions of the network. In particular, the dropout layer is inserted at three different parts of the transformer block, as can be seen from Figure 5.3. In the self-attention module, the dropout is positioned in two places. The first one is called attention dropout, and it is located right after the softmax layer in the self-attention block. The dropout is applied to the attention weights to mask some of the attention weights randomly. The second dropout is placed at the output of the multi-head self-attention module; after concatenating the feature representations of the multi-head attention and then projecting them using a linear layer (Figure 5.3 (a)). It is called projection dropout. It can be helpful in improving the generalization of multi-head attention. The third dropout is located in between the linear layers in the MLP module and at the end of the MLP module (Figure 5.3 (c)). It is named MLP dropout. These dropout layers are placed in all transformer blocks of the encoder and decoder.

A Swin transformer-based U-Net with dropout is trained to segment the heart structures from native and post-contrast T1 mapping CMR images. The model is sampled N times during testing to obtain N Monte-Carlo segmentation samples. The uncertainty metrics are derived from these Monte-Carlo samples. The mean of MC samples is used as the final segmentation. The uncertainty metrics utilized are described in detail in Section 5.2.2.2.

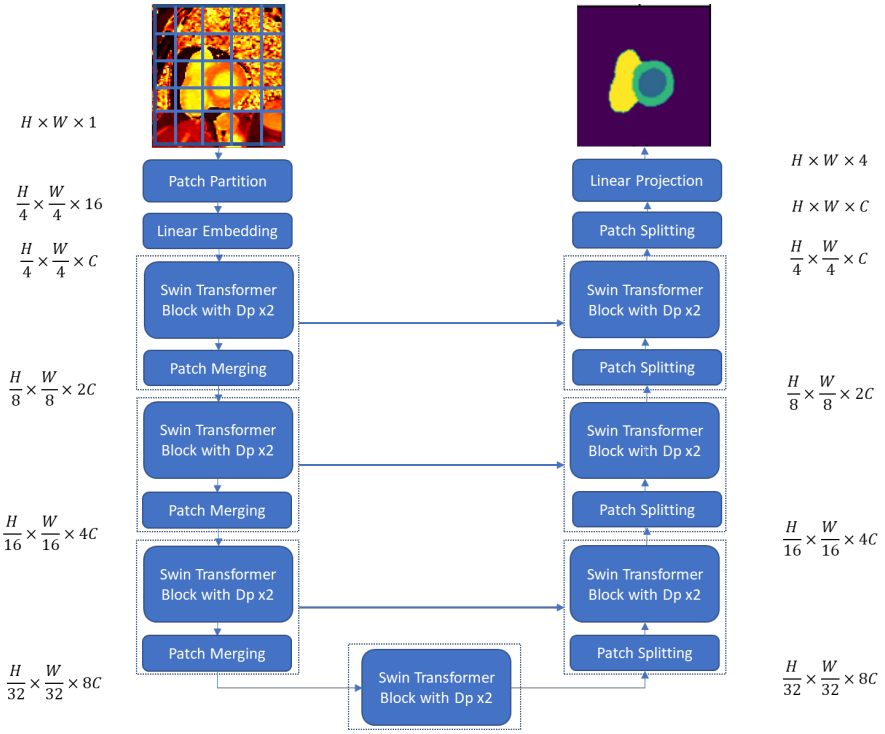

Figure 5.2: Segmentation network architecture: Swin-based U-Net with dropouts (Dp) activated at the MLP part of the transformer block.

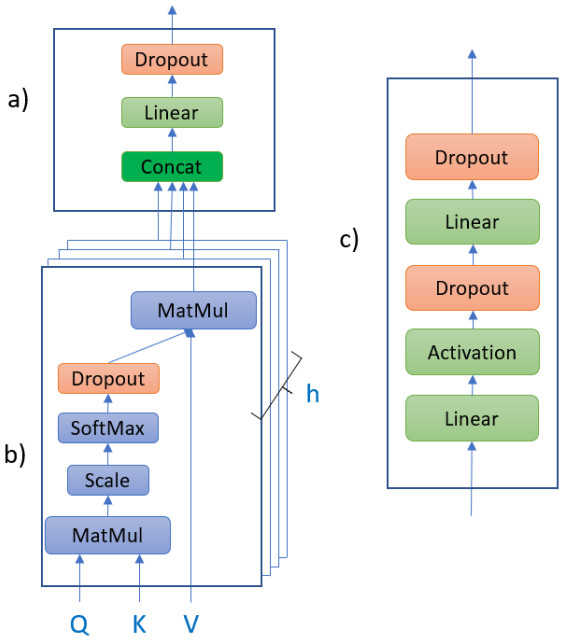

Figure 5.3: Dropout introduced in multi-head attention module (a) Scaled dot-product Attention module (b) and in Multi-Layer Perceptron (MLP) module (c), where h represents the number of heads.

5.2.2.2/ Uncertainty-based Quality Control

From the Bayesian Swin transformer-based U-Net, we get N Monte-Carlo segmentation samples $(S_1, ..., S_N)$ when the model is sampled N times. From these segmentation samples, we compute the mean of the segmentation samples (Eq. 5.1) to get the final prediction. For the uncertainty, two different pixel-wise uncertainties are calculated. The sample variance is computed as the variance of the N Monte-Carlo prediction samples of a pixel (Eq. 5.2) [Nair et al., 2020, Arega et al., 2021a]. It captures the model uncertainty. The second one is predictive entropy, which is estimated by calculating the entropy of the mean probability vector across the class dimension (Eq. 5.3) of each pixel, and it captures predictive uncertainty. To leverage the pixel-wise uncertainty measures to estimate the quality of the segmentation result, we computed the mean of each pixel-wise uncertainty (mean of sample variance and mean of predictive entropy). This way, we have a single uncertainty score for each segmentation.

$$\mu_i = \frac{1}{N} \sum_n (y_{i,n}) \tag{5.1}$$

$$\sigma_i^2 = \frac{1}{N} \sum_n (y_{i,n} - \mu_i)^2 \tag{5.2}$$

$$entropy = -\sum_{c=1}^{C} (p_c \log p_c) \tag{5.3}$$

In addition to the mean pixel-wise uncertainties, we defined two other uncertainty measures: Dice agreement within MC samples (*DiceWithinSamples*) [Roy et al., 2018, Ng et al., 2020] and HD agreement within MC samples *HDWithinsamples*. *DiceWithinSamples* is the average Dice score of the mean predicted segmentation (S_{mean}) and the individual N MC prediction samples as shown in Eq.5.4. When the *DiceWithinsamples* is very high, it shows that the model's MC samples have high agreement among themselves and the model's uncertainty is very low and vice-versa. *HDWithinSamples* is the average HD score of the mean predicted segmentation (S_{mean}) and the individual N MC prediction samples (Eq. 5.5).

$$Dice_{\text{WithinSamples}} = \frac{1}{N} \sum_n Dice(S_{mean}, S_n) \tag{5.4}$$

$$HD_{\text{WithinSamples}} = \frac{1}{N} \sum_n HD(S_{mean}, S_n) \tag{5.5}$$

In this research, we propose a simple uncertainty-based quality control that leverages image level uncertainty metrics such as *DiceWithinsamples*, *HDWithinsamples*, mean

sample variance, and mean predictive entropy to predict the quality of the segmentation. From the uncertainty metrics, the mean of sample variance and mean of predictive entropy are directly computed from the pixel-wise uncertainties whereas *DiceWithinsamples* and *HDWithinsamples* are calculated from the MC segmentation samples and the mean segmentation map, as can be seen from Figure 5.4. These image-level uncertainty features are fed to a random forest (RF) classifier/regressor to train the model to classify the quality of the segmentation result or to directly regress the Dice score. In our experiment, we used RF with 100 trees and gini criterion, and we utilized Scikit-learn's RF implementation [Pedregosa et al., 2011]. The RF classifier/regressor is trained and validated using the image-level uncertainty metrics computed from the images that were used to train and validate the Bayesian segmentation model.

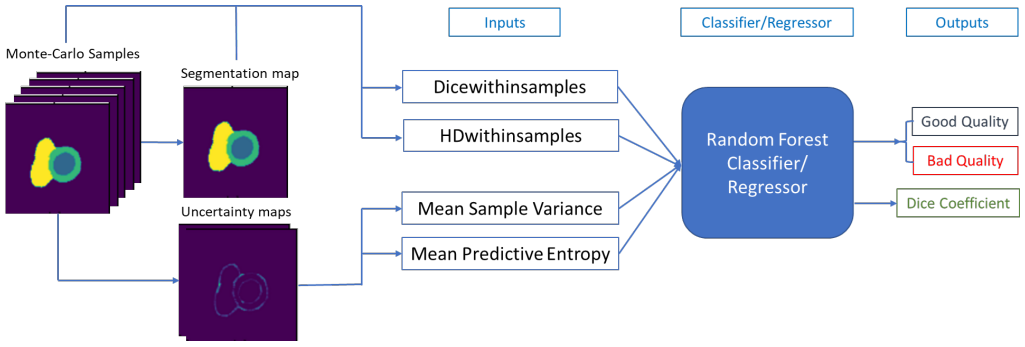

Figure 5.4: The proposed uncertainty-based QC method. The RF classifier/regressor uses four image-level uncertainty features as an input to determine the segmentation quality.

The labels for the quality of the segmentation maps are based on the mean Dice score of the cardiac structures with the ground truth segmentations. If the average Dice score of the cardiac structures (left ventricular blood pool, myocardium, and right ventricular blood pool) is greater than 0.9, the quality of the predicted segmentation map is labeled as good; otherwise, it is labeled poor quality. The threshold is selected after reviewing the inter-observer agreements on cardiac segmentation from different literature [Bai et al., 2018, Bernard et al., 2018, Kushibar et al., 2022].

5.2.2.3/ T1 Mapping and ECV Computation

For T1 mapping analysis, we computed the mean native T1 value of the myocardium to characterize the myocardial tissues. For ECV, it is calculated by measuring myocardial and blood pool T1 before and after administration of contrast agent as well as the patient's hematocrit value (Eq. 2.1) [Haaf et al., 2016].

To exclude papillary muscles while calculating the mean T1 value of the blood pool, we eroded the segmented blood pool until $1/3$ of the area remains [Puyol-Antón et al., 2020].

Similarly, to avoid the partial volume effect while calculating the mean T1 value of the myocardium, we eroded the segmented myocardium until $2/3$ of the area was left.

ECV refers to the space or volume of a tissue that is not occupied by cells and includes the intracapillary plasma volume [Moon et al., 2013]. To compute ECV, the patient's hematocrit value is needed which can be cumbersome to do during cardiac MRI acquisition. When the patient's hematocrit is not measured during the clinical routine, a synthetic hematocrit can be derived from the relaxation rate of blood pool (native T1 values of blood pool) as shown by [Treibel et al., 2016]. According to [Treibel et al., 2016] and [Fent et al., 2017], there is a linear relationship between hematocrit and blood relaxivity (1/blood pool T1 time). The synthetic hematocrit is calculated using the following formula [Treibel et al., 2016]:

$$SyntheticHematocrit = (866.0 \times \frac{1}{T1_{blood\_native}}) - 0.1232 \tag{5.6}$$

Where $T1_{blood\_native}$ is the mean T1 value of the blood pool computed from native T1 mapping images.

Since all the patients in our dataset did not take blood sampling (hematocrit) during the MRI acquisition, ECV was calculated using a synthetic hematocrit value that was derived from the native T1 value of the blood pool as shown in Eq. 5.6.

5.2.2.4/ IMPLEMENTATION

For the segmentation networks, all models are trained for 2000 epochs using Stochastic gradient descent (SGD) with Nesterov momentum ($\mu = 0.9$) with an initial learning rate of 0.01 with a poly learning rate decay strategy. During training, the batch size is set to 24. The loss function employed is a weighted sum of cross-entropy and Dice losses. The weights of the segmentation models are initialized using the pre-trained weights of the Swin transformer (on ImageNet) [Liu et al., 2021]. To improve the model's generalization, a simple data augmentation that includes random rotation and flipping is used. The models were implemented using the Pytorch deep learning framework and trained on NVIDIA Tesla V100 GPUs with 32GB of memory.

5.3/ RESULTS

For segmentation, the performance was mainly evaluated using Dice coefficient and Hausdorff distance (HD) metrics. To measure the calibration quality of the models, we used the Brier score. Brier score measures how close the predicted segmentation proba-

bilities are to their corresponding ground truth probabilities by computing the mean square error of the two probabilities [Arega et al., 2021a]. For the quality control, the performance of the models regarding Dice regression was assessed using the mean absolute error (MAE) metric. The Pearson correlation coefficient (r) of the predicted Dice and the ground truth Dice was also computed. For the binary classification task of the QC method, the classifier's performance was evaluated using the Receiver Operating Characteristics (ROC) curve and the Area under the ROC Curve (AUC).

The dataset was shuffled and randomly split into 60% training, 20% validation and 20% testing datasets. For all the reported results in this chapter, we utilized the test set unless mentioned otherwise. The statistical analysis was performed using SciPy (python library) [Virtanen et al., 2020]. To assess the statistical difference between samples, we utilized the Mann-Whitney U test [Mann et al., 1947], a non-parametric statistical significance test.

5.3.1/ SEGMENTATION

Table 5.1 shows the performance of different dropout variants in the validation set. Most of the dropout variants have similar segmentation accuracy, including the one with no dropout. From all the variants, MLP dropout has relatively better segmentation performance and model calibration. It has the best Dice score on all the heart structures (left ventricular blood pool, myocardium, and right ventricular blood pool). These improvements were mostly statistically significant $(p < 0.05)$ except for the left ventricular blood pool. For the HD, it also achieved the best result in all but the left ventricular myocardium. In terms of calibration quality, MLP dropout yielded the lowest Brier score (the lower the better) $(p < 0.05)$, illustrating that dropout in MLP layer of the transformer block produces well-calibrated uncertainty estimates. Transformer dropout uses dropout at both attention weights and at the end of multi-head attentions. We noticed that the segmentation performance and calibration quality of the Transformer dropout were relatively worse than the other dropout variants.

In terms of dropout rates and the number of Monte-Carlo samples (the N value in Eq. 5.1 and 5.2), we experimented with different dropout rates (0.1, 0.3, 0.5, 0.7, and 0.9) and various numbers of samples (5, 10, 15, 25 and 50). The dropout rate of 0.1 gave us the best result. Performances were improved while increasing the number of samples. However, since the segmentation accuracy difference between sample number 5 and sample number 50 was not significant, for the rest of the experiment we utilized sample number 5 as it was computationally less expensive.

For the Bayesian CNN-based U-Net, similar to [Kendall et al., 2017a], we compared different dropout variants, and the variant that uses Dropout at the central Encoder-Decoder

Table 5.1: Quantitative comparison of Swin-based U-Net at different Dropout positions in terms of Dice score, HD (in mm) and Brier score (BS) in the validation set. The bold results are better. LV: left ventricular blood pool, RV: right ventricular blood pool, MYO: left ventricular myocardium. Statistically significant differences ($p < 0.05$) compared to MLP Dropout are indicated by '*'.

Dropout variant	**Dice LV**	**Dice MYO**	**Dice RV**	**HD LV**	**HD MYO**	**HD RV**	**BS** (10^{-3})
No Dropout	0.976	0.926 *	0.941 *	2.12 *	2.38	4.26 *	3.61 *
MLP Dropout	**0.979**	**0.934**	**0.945**	**1.61**	1.99	**3.60**	**3.31**
Attention Dropout	0.977	0.928 *	0.940 *	2.08 *	2.45	5.15 *	3.56 *
Projection Dropout	0.977	0.924 *	0.937 *	2.11 *	2.57	4.94 *	3.69 *
Transformer Dropout	0.976	0.923 *	0.936 *	2.55 *	2.57	4.36 *	3.73 *
MLP-Projection Dropout	0.977	0.929 *	0.941	1.64	**1.95**	3.90 *	3.56 *
MLP-Attention Dropout	0.977	0.931	0.939 *	1.72 *	2.07	3.91 *	3.46 *
All Dropout	0.978	0.929 *	0.939 *	1.71 *	2.14	5.22 *	3.54 *

position achieved the best result in terms of segmentation accuracy and uncertainty calibration in the validation set. We utilized this variant of the Bayesian CNN-based U-Net in the subsequent stages.

It should be noted that all hyperparameter tunings, including the selection of the best dropout variant for Bayesian Swin-based U-Net (MLP-Dropout) and Bayesian CNN-based U-Net (Central Enc-Dec Dropout), were done in the validation set. These best models were then taken forward in the subsequent stages.

5.3.2/ Uncertainty-based Quality Control

The proposed QC method is compared to different state-of-the-art QC methods, which are based on the various inputs including the image, segmentation map, and uncertainty map. The first QC method, Seg [Chen et al., 2020c], uses only a segmentation map as an input to the classification/regression method to determine the quality of the segmentation result. Seg-Uncert QC [Williams et al., 2021] method utilizes both segmentation map and uncertainty map as an input. Image-Seg QC [Robinson et al., 2018, Huang et al., 2016] employs the image-segmentation pair as an input. Whereas the Image-Seg-Uncertainty QC method [Devries et al., 2018b, Chen et al., 2020c] uses all of the three inputs together (the image, segmentation map, and uncertainty map). All these methods utilize CNN-based network architectures, which are ResNet-18 and ResNet-34 classification/regression networks. These networks are selected because their performance was better than other similar classification/regression networks for the specific task. Similarly, for the proposed method, RF classifier/regressor is chosen due to its superior performance on the task compared to other machine learning and deep learning (multi-layer perceptron (MLP)) based classifiers/regressors.

Table 5.2: Segmentation performance of Bayesian Swin-based U-Net (S U-Net) and Bayesian CNN-based U-Net (C U-Net) on the native T1 mapping and post-contrast T1 mapping dataset in terms of Dice score and HD (in mm). The bold results represent the best. LV: left ventricular blood pool, RV: right ventricular blood pool, MYO: left ventricular myocardium. Statistically significant differences $(p < 0.05)$ compared to Swin-based U-Net are indicated by '*'.

Dataset	Method	Dice LV	Dice MYO	Dice RV	HD LV	HD MYO	HD RV
Native	S U-Net	**0.972**	**0.916**	**0.922**	**1.90**	**2.26**	**4.43**
	C U-Net	0.968	0.907 *	0.905 *	2.10 *	2.79 *	5.93 *
Post-Contrast	S U-Net	**0.954**	**0.887**	**0.893**	**2.81**	**3.22**	**5.43**
	C U-Net	0.941 *	0.857 *	0.867 *	3.31 *	4.01 *	6.70 *

The QC methods are evaluated on segmentation results of CNN-based U-Net and Swin transformer-based U-Net. This can tell us how the QC methods perform in identifying inaccurate segmentation results from a CNN-based U-Net and transformer-based U-Net on native T1 mapping and post-contrast T1 mapping datasets. Regarding the datasets, the native T1 mapping dataset generally has better image quality, and the contrast among the three heart structures is good. The post-contrast T1 mapping dataset is relatively more challenging for the segmentation models as the contrast among the heart structures is lower and some of the images have artifacts and noise. Looking at the quantitative results of CNN-based and Swin-based segmentation networks in Table 5.2, the latter method has better segmentation performance achieving higher Dice scores and lower HD on all the three heart structures. This performance enhancement can be due to the Swin transformer's strong feature representation capability as it leverages the advantages of both the vanilla transformer and CNN.

Table 5.3 summarizes the Dice regression results of the QC methods in terms of mean absolute error (MAE) and Pearson correlation coefficient (Pearson CC) between the predicted Dice and the ground truth Dice. From the table, one can observe that the proposed method achieved the best result in terms of MAE and Pearson CC. From the CNN-based U-Net's segmentation results on native T1 mapping images, the proposed method obtained a mean absolute error of 0.01636, significantly outperforming the Seg (0.01994), Seg-Uncert (0.01997), Image-Seg (0.02072) and Image-Seg-Uncertainty (0.02085) QC methods. In terms of the correlation coefficient, the predicted Dice of the proposed method has the highest Pearson CC with the ground truth Dice $(r = 0.88)$ compared to the other QC methods. For the post-contrast T1 mapping images, the proposed method achieved the lowest MAE (0.02244) with a statistical significance of $p < 0.01$ among the QC methods. It also reached a Pearson CC of 0.82 ahead of all other QC methods by a margin of $9 - 15\%$, showing the robustness of our proposed method on relatively difficult images. After the proposed method, Seg-Uncert QC has the second-best Dice regression performance among the QC methods. Similarly, looking at the Dice regression result from

Table 5.3: Dice score regression results of different quality control (QC) methods: Seg [Chen et al., 2020c], Seg-Uncert [Williams et al., 2021], Image-Seg [Robinson et al., 2018, Huang et al., 2016], Image-Seg-Uncertainty [Devries et al., 2018b, Chen et al., 2020c] and the Proposed QC on various segmentation result types in terms of MAE and Pearson CC between the predicted Dice and the ground truth Dice. Bold results are the best. Asterisks indicate a statistically significant improvement in MAE comparing the proposed QC with the other QC methods.

Model	Dataset	QC Method	MAE	Pearson CC
CNN U-Net	Native	Seg	0.01994 (0.03636) **	0.74
		Seg-Uncert	0.01997 (0.03545) **	0.76
		Image-Seg	0.02072 (0.03757) *	0.76
		Image-Seg-Uncert	0.02085 (0.03617) ***	0.72
		Proposed	**0.01636 (0.02715)**	**0.88**
	Post-contrast	Seg	0.03189 (0.03229) ***	0.71
		Seg-Uncert	0.02937 (0.03822) **	0.70
		Image-Seg	0.02876 (0.03865) **	0.71
		Image-Seg-Uncert	0.02916 (0.03808) **	0.69
		Proposed	**0.02244 (0.03095)**	**0.82**
Swin U-Net	Native	Seg	0.02006 (0.02684) **	0.70
		Seg-Uncert	0.01959 (0.02815) *	0.67
		Image-Seg	0.01980 (0.02726) *	0.72
		Image-Seg-Uncertainty	0.01953(0.02421) *	0.74
		Proposed	**0.01731 (0.02277)**	**0.82**
	Post-contrast	Seg	0.02854 (0.04890) *	0.67
		Seg-Uncert	0.02790 (0.04729)	0.69
		Image-Seg	0.03001 (0.04708) *	0.67
		Image-Seg-Uncert	0.02940 (0.05158) *	0.64
		Proposed	**0.02634 (0.03698)**	**0.82**

* stat. significant with $p < 0.05$ ** stat. significant with $p < 0.01$ *** stat. significant with $p < 0.001$

Swin-based U-Net's segmentation, the proposed QC method, which utilizes very simple uncertainty-based features, achieved the best result in both native T1 and post-contrast images. Comparing the Dice regression results of post-contrast and native T1 mapping images for both models, the MAE of all QC methods on native T1 mapping images was lower than on the corresponding post-contrast images.

The quality control methods were assessed for their ability to classify the segmentation results as poor quality (mean ground truth Dice < 0.9) or good quality (mean ground truth Dice ≥ 0.9). Akin to the Dice regression, their performance was evaluated on the segmentation results of Swin-based U-Net and CNN-based U-Net on both native and post-contrast T1 mapping images. As can be seen from Figure 5.5 (b), on segmentation results of CNN-based U-Net of native T1 mapping images, all the QC methods achieved superior results comparing their classification performance on other segmentation results (Figure 5.5 (a), (c) and (d)). Img-Seg, Seg-Uncert, and Seg QC methods reached an AUC of 0.922, 0.918, and 0.901, respectively. Among the QC methods, the classification

performance of Img-Seg-Uncertainty was the worst $(AUC = 0.867)$. However, when the proposed method is compared to the other methods, it obtained the best result with an AUC of 0.958. In classifying the quality of segmentation results of the more challenging images (post-contrast), the performance of the proposed method was also robust.

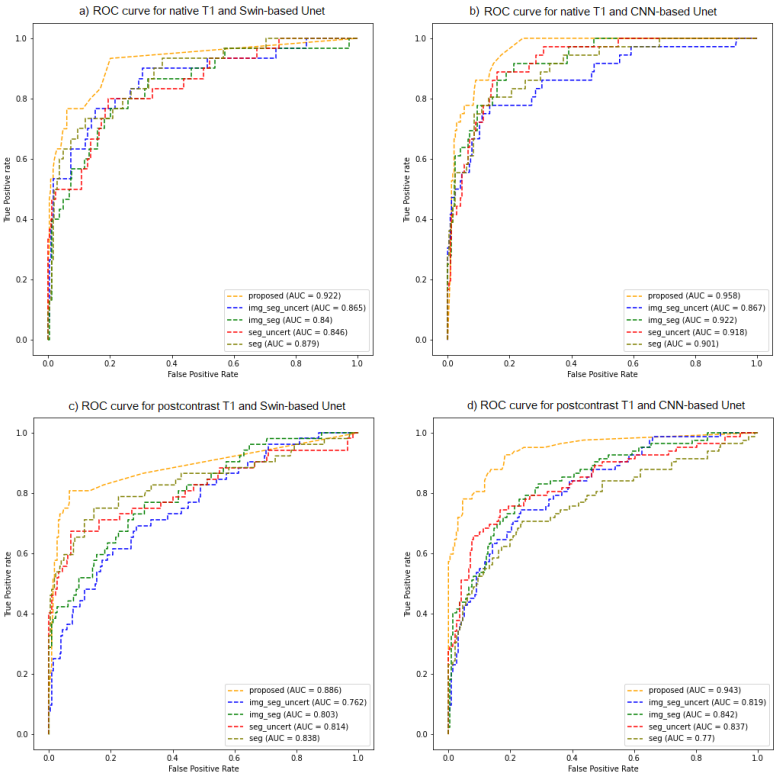

Figure 5.5: ROC curve and AUC comparison of different QC classifiers on four types of segmentation results. In the top row, on segmentation results of Swin-based U-Net (a) and CNN-based U-Net (b) of native T1 images. In the bottom row, on segmentation results of Swin-based U-Net (c) and CNN-based U-Net (d) of post-contrast T1 images.

The CNN-based U-Net on post-contrast images has the lowest mean Dice compared to the other segmentation results (Table 5.2). As can be seen from Figure 5.5 (d), the proposed method's classification performance on the segmentation results of CNN-based U-Net on post-contrast images was superior, reaching an AUC of 0.943 ahead of all other QC methods by a large margin $(10\% - 17.3\%)$. This shows how robust our method is in detecting inaccurate segmentation of a poor-performing segmentation model. Interestingly, the proposed method's performance in detecting failures from a good-performing segmentation model is also strong. However, the performance gap between the proposed method and the other QC methods is smaller when the segmentation model has good segmentation accuracy. From the ROC curves in Figure 5.5, one can also observe that after the proposed method, on average Seg-Uncert QC method performed better than the other QC methods on all of the segmentation results. The Img-Seg-Uncertainty QC method, which uses all the three inputs to the classification network, performed the worst.

After determining the quality of the segmentation results using the QC methods, the bad quality segmentation results are rejected. Then we computed the mean Dice and mean HD (in mm) of the good quality segmentation results (the retained images), as can be seen from the box plots in Figure 5.6. For this analysis, we focused only on the performance of the QC methods on the segmentation results of Swin-based U-Net on post-contrast T1 images. Of the 279 test images (slices), only the segmentation results of 52 images are classified as bad quality because their mean ground truth Dice is less than 0.9. We refer to this as ground truth (GT) for the QC methods (Figure 5.6). This is the ideal QC that rejects all the bad quality segmentations. In the comparison, we also included the performance of No-QC, where all the segmentation results (279) are retained. Looking at the number of rejected segmentations of each QC method, the proposed method rejected 44. Whereas Seg, Seg-Uncert, Image-Seg, Image-Seg-Uncertainty QC methods classified 17, 23, 22, and 17 segmentations as bad quality, respectively.

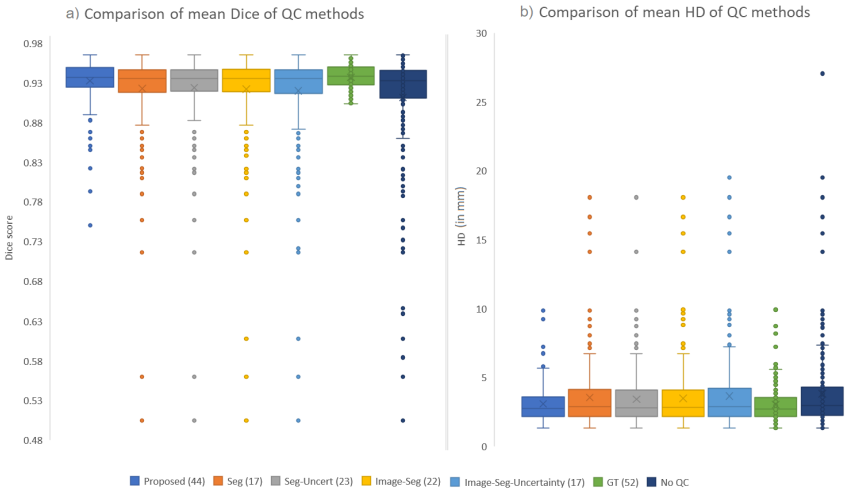

Figure 5.6: Box plots comparing mean Dice (a) and mean HD (in mm) (b) of different QC methods after each QC rejected their poor quality images. Note that the numbers inside the parenthesis of the QC method names represent the number of images rejected by each QC.

From the box plot (Figure 5.6), one can observe that the proposed method has rejected most of the bad quality images and significantly enhanced the mean Dice from 0.912 to 0.932 $(p < 0.001)$ and the mean HD from 3.82 mm to 3.08 mm $(p < 0.01)$ compared to No-QC. The difference in performance between GT QC and the proposed method is not significant $(p > 0.05)$ in terms of mean Dice 0.0045 and mean HD 0.058 mm. This shows us that the proposed method has successfully removed the images with implausible segmentation (very high HD). From Figure 5.6, we also noticed that out of all the QC methods, the proposed QC has the lowest number of outliers particularly in the mean HD. Considering the other QC methods, Seg-Uncert QC has the second-best performance after the proposed QC. It achieved 0.924 mean Dice and 3.41 mm mean HD even though

the difference in performance compared to the GT QC is significant concerning mean Dice ($p < 0.01$). Following Seg-Uncert QC, Image-Seg QC and Seg QC produced a mean Dice of 0.922, 0.923 and mean HD of 3.48 mm, 3.53 mm respectively. We observed that Image-Seg-Uncertainty QC, which has the worst performance in the regression and classification tasks, achieved the lowest mean Dice and the highest mean HD, and the highest number of outliers (Figure 5.6).

Analyzing the images rejected by the QC method according to their position, 73.1% of the images were from the apical slices, and 17.3% were from the mid slices, and 9.61% were from the basal slices. Examining the failed images further in terms of their pathology, we noticed that 35% of the myocardial infarction, 33% of the amyloidosis, 30% of the HCM ,and 17% of the DCM slices were rejected. In contrast, normal and pathological slices such as Tako-Tsubo syndrome and myocarditis have less than 10% failed images.

Figure 5.7 depicts some of the segmentation results which are rejected by the proposed QC method. Figure 5.7 $(a - c)$ shows failed cases from native T1 mapping images and $(d - f)$ are from post-contrast T1 mapping images. One can observe that these images have some artifacts that confuse the segmentation models. It led to higher uncertainty on the location where an artifact was positioned. In the last two columns of Figure 5.7, we can see that there is higher uncertainty whenever there is a segmentation error. In this figure, Uncertainty-I is a sample variance, and Uncertainty-II is a predictive entropy.

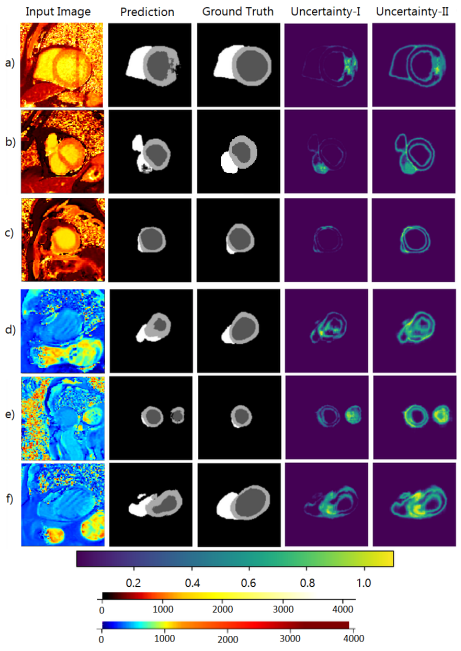

Figure 5.7: Examples of segmentation results rejected by the Proposed QC method. Rows (a-c) show rejected images from native T1 mapping dataset and rows (d-f) show rejected images from post-contrast T1 mapping dataset. Uncertainty-I and Uncertainty-II represent sample variance and predictive entropy uncertainty maps respectively.

To interpret the decision of the random forest classifier, we analyzed the impor-

tance of each input feature on the random forest's decision using Scikit-learn library [Pedregosa et al., 2011]. As depicted in Figure 5.8, the most important feature is $Dicewithinsamples$ which has an importance score of 0.7. The second most important feature is $HDwithinsamples$ with a score of 0.16. The contribution of image-level sample variance ($IL\_mean\_variance$) and predictive entropy ($IL\_mean\_entropy$) is minimal compared to the first two features in altering the decision of the RF classifier. They have a feature importance score of 0.064 and 0.076, respectively.

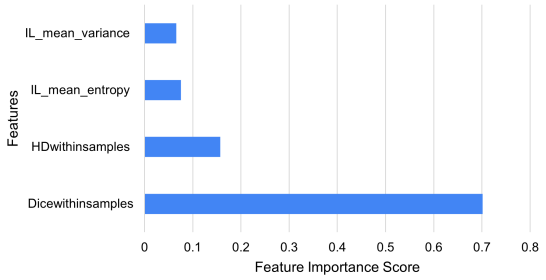

Figure 5.8: Feature importance score for RF classifier. Note that $IL\_mean\_variance$ is the mean of sample variance and $IL\_mean\_entropy$ is the mean of predictive entropy.

For myocardial T1 analysis, we only need the segmentation of the myocardium, whereas, for ECV analysis, the segmentations of all the structures (myocardium and blood pools) are used. Due to this, we compared the performance of a general QC and a separate QC. General QC is the proposed QC method that uses the mean Dice of the three structures (LV, MYO, and RV) to determine the ground truth quality of the segmentation, i.e., poor quality (mean ground truth Dice < 0.9) or good quality (mean ground truth Dice ≥ 0.9). Separate QC is also the proposed QC method, but it uses the Dice of each structure separately to determine the ground truth quality of the segmentation. We compared these two QCs in terms of the mean Dice of the myocardium or all structures and mean absolute error (MAE) between the automatically and manually calculated myocardial T1 or ECV values of the retained good quality segmentation results on the test set.

For myocardial native T1 analysis, a general QC is compared with a separate QC that focuses only on the myocardium (considers a segmentation result as good quality if the ground truth Dice of the myocardium is greater than 0.9). The separate QC enhanced the mean Dice of the myocardium from 0.920 to 0.923 and the MAE of myocardial T1 values from 11.00 ms to 10.25 ms compared to the general QC. This performance enhancement was achieved with the same training and inference time as the general QC. For ECV analysis, even though the separate QC achieved better performance by increasing the mean Dice of all structures from 0.943 to 0.946 and by lowering the MAE of ECV from 0.96% to 0.86% compared to the general QC, its training and inference time was 3 times higher than the general QC. This is because, for separate QC, we need to train six models (3

for native T1 and 3 for post-contrast T1 images) and use these six trained models during inference to decide the segmentation's quality, whereas, for general QC, two models are needed. Furthermore, the training and inference times will linearly increase as the number of structures to be segmented increases. By looking at the trade-off between performance improvement and computational overhead, we decided to use a separate QC for myocardial T1 analysis and a general QC for ECV analysis.

To demonstrate the generalizability of the proposed QC on public datasets, experiments were conducted on public cardiac MRI datasets - ACDC [Bernard et al., 2018] and the Extreme Cardiac MRI Analysis Challenge under Respiratory Motion (CMRxMotion) [Wang et al., 2022] datasets. Across both datasets, the proposed QC demonstrated superior detection performance compared to the other uncertainty-based QC methods, with higher AUC and F1 scores. These results using public benchmark datasets further highlight the efficacy of the proposed approach for quality control in cardiac MR image segmentation. A more detailed comparison of the QC methods on these two public datasets can be found in Appendix B.

5.3.3/ T1 Mapping and ECV Analysis

For the analysis of T1 mapping and ECV, we utilized both the validation and test dataset $(401$ slices) to increase the number of cases per pathology. From 401 images (slices), the proposed QC method rejected 73 images for myocardial T1 analysis and 78 images for ECV analysis due to their inaccurate segmentation. The mean absolute error between the automatic and manual myocardial T1 values (ms) and ECV $(\%)$ of the images retained by the proposed method are 10.41 ms and 0.91%, respectively. Compared to No-QC, which keeps all images, the MAE is lower by 3.7 ms $(p < 0.05)$ for myocardial T1 values and 1% $(p < 0.05)$ for ECV. The proposed method also reached a Pearson correlation coefficient of 0.990 for myocardial T1 values and 0.975 for ECV. The automatic ECV is calculated using Eq. 2.1 from the automatically segmented right-ventricular and left-ventricular blood pools and left-ventricular myocardium, whereas the manual ECV is computed from the manual segmentation of these areas. Similar to ECV, the automatic and manual myocardial T1 values are calculated from the automatically segmented and manually segmented myocardium, respectively. The T1 mapping and ECV analysis are done on the good-quality images that are retained by the proposed QC method. In this analysis, we also did not include some patients who have undetermined or complex pathologies (51).

In Figure 5.9, we computed the mean of each patient group's native myocardial T1 value and ECV to see if there is a clear difference between healthy and pathological groups. From the native myocardial T1 values of each group (Figure 5.9 (a)), we can see that pathology groups with diffuse fibrosis like DCM and HCM have a native myocardial T1

value of 1128 ms and 1057 ms respectively. Other pathologies such as Tako-Tsubo syndrome, amyloidosis and myocardial infarction obtained mean T1 values of 1142 ms, 1115 ms, and 1079 ms, respectively. These native T1 values are much higher than the healthy group's native T1 value which is 1035 ms. Among the pathologies, only the myocarditis group has a lower T1 value (1032 ms) than the healthy group. Comparing the myocardial pathologies with the healthy group, Tako-Tsubo syndrome ($p < 0.001$), amyloidosis ($p < 0.001$), DCM ($p < 0.001$), myocardial infarction ($p < 0.001$) and HCM ($p < 0.05$) have significantly higher mean native T1 values. The difference in native T1 value between myocarditis and the healthy group was not significant ($p > 0.5$).

From Figure 5.9 (b), one can observe that amyloidosis and myocardial infarction patient groups obtained a mean ECV(%) of 44.04 and 35.38 respectively. These are the two highest mean ECVs from all patient groups. Myocardial pathological diseases with diffuse fibrosis such as DCM and HCM have a mean ECV(%) of 30.17 and 30.37 respectively which are slightly higher than the healthy group (27.86). Myocardial diseases such as myocarditis got a mean ECV(%) of 27.39 which is quite similar to the healthy group. Tako-Tsubo syndrome pathology has a mean ECV(%) of 31.2. From the ECV result, we observed that there is a statistically significant difference in ECV between the healthy group and the following myocardial pathological groups: myocardial infarction ($p < 0.001$), amyloidosis ($p < 0.001$), HCM ($p < 0.05$), and DCM ($p < 0.05$). Whereas the difference in ECV between the healthy group and pathological groups such as myocarditis ($p > 0.5$) and Tako-Tsubo syndrome ($p > 0.05$) is not significant.

5.4/ DISCUSSION

In this chapter, we proposed a fully automatic quality-controlled framework for myocardial tissue quantification from native myocardial T1 and ECV values. The framework has three main parts. The first part involves segmenting heart structures using Bayesian Swin transformer-based U-Net from T1 mapping images. Then in the second part, the quality of the segmentation results is evaluated by a novel uncertainty-based quality control method that automatically detects and rejects the failed segmentation results. The last part involves T1 mapping and ECV analysis to distinguish myocardial pathology groups from healthy groups of the retained good quality segmentation results.

To generate uncertainty from Swin-based U-Net using MC-Dropout, we placed dropout at different locations of the segmentation model. Analyzing the dropout variants, using dropout in the self-attention part of the transformer block (attention dropout, projection dropout, etc), which is responsible for computing the representation of a patch by relating to different patches in the same window, results in lower segmentation performance and calibration. This can tell us that features extracted by self-attention layers are better mod-

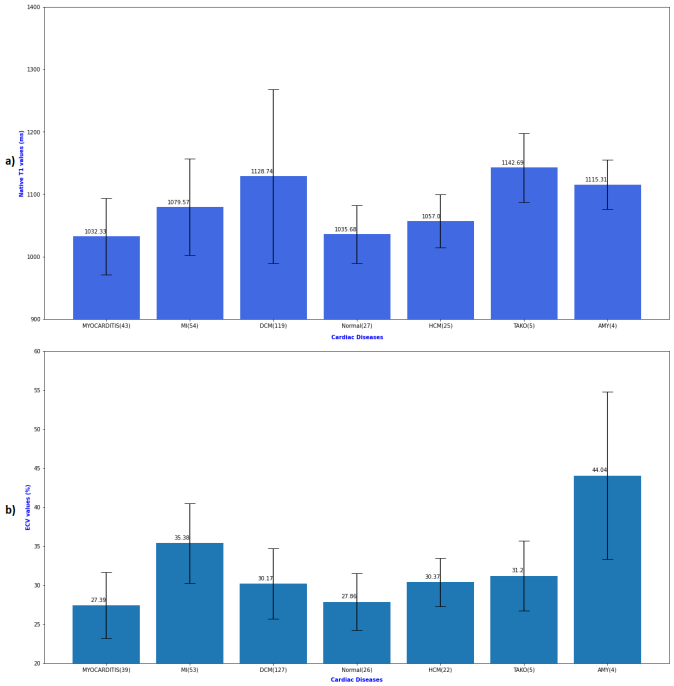

Figure 5.9: Comparison of native myocardial T1 (a) and ECV (b) values of healthy and various cardiac diseases. MI: myocardial infarction, HCM: hypertrophic cardiomyopathy, DCM: dilated cardiomyopathy, TAKO: Tako-Tsubo syndrome, AMY: amyloidosis

eled with deterministic weights. In contrast, adding dropouts to the MLP layers yielded better results in terms of both segmentation accuracy and calibration. This shows us that the features extracted from MLP layers (the later part of the transformer block) can be more useful with Bayesian weights.

From the results in Section 5.3.2, we can conclude that the proposed QC method that leverages four important image-level uncertainty features to determine the quality of segmentation results can accurately detect most of the failed segmentation results of both good performing models (Swin-based U-Net) and relatively poor performing models (CNN-based U-Net). It also performed relatively better on the more challenging post-contrast T1 mapping dataset than on the native T1 mapping dataset. Compared to the other QC methods, its performance was superior in classifying/regressing segmentation results of challenging datasets or poor-performing models. This shows how robust our method is in detecting failed segmentations. However, the performance gap between the proposed method and the other QC methods becomes lower when the segmentation model has good segmentation accuracy or when the dataset has good quality.

Of the other QC methods, the Seg-Uncert QC method has on average the second-best performance in terms of regression/classification after the proposed method. Its performance tells us that when using image-segmentation-uncertainty-based QC methods using a combination of segmentation map and uncertainty map gives a better estimation

of the quality of the segmentation. However, in general, due to the weak correlation between the inputs (image, segmentation map, and uncertainty map) and the segmentation quality, the other QC methods failed to detect most of the poor segmentation results.

Studying the effect of the four image-level uncertainty features on the RF classifier, the results showed us that the image-level uncertainty metrics $Dicewithinsamples$ and $HDwithinsamples$ hold vital information related to the segmentation quality. This can be because $Dicewithinsamples$ and $HDwithinsamples$ are highly correlated with the ground truth Dice coefficient and HD, respectively, as mentioned by [Roy et al., 2018] and [Ng et al., 2020]. Even though the feature importance score of the image level sample variance and predictive entropy features were very small, their contribution was not negligible as the RF classifier performance decreased when these two features were omitted.

Examining the images rejected by the proposed QC based on their position, we noticed that more than two-thirds of them are from the apical slices of the left ventricle. This can be expected as the most difficult and challenging images are mainly the apical slices of the left ventricle even for human experts [Petitjean et al., 2011, Bernard et al., 2018]. Investigating further the rejected images based on their pathological group, we found out that pathological groups such as myocardial infarction, amyloidosis, HCM and DCM were rejected in higher percentages than other pathology groups such as healthy, myocarditis and Tako-Tsubo syndrome. This can be due to the fact that the first group of pathologies changes the contrast and anatomy of the heart structures more than the second group of pathologies. For example, myocardial infarction has lower contrast between the left ventricular myocardium and left ventricular blood pool due to the presence of scar tissues [Rajiah et al., 2013, Cui et al., 2018]. In the case of pathologies like HCM and DCM, the anatomy of the myocardium is changed by thickening or thinning it [Amano et al., 2018, Francone, 2014]. The change in contrast and anatomy can lead to worse segmentation performance. In contrast, images that come from healthy, Tako-Tsubo syndrome, and myocarditis pathology groups affect less the anatomy of the heart structures.

Using the proposed method for downstream tasks such as T1 mapping and ECV analysis, we showed that our approach achieved very low MAE between the automatic and manual myocardial native T1 and ECV values. For both myocardial T1 value and ECV value, even though the proposed QC method removed images with bad segmentation results and decreased the MAE significantly, the improvements were not very high. This can be due to the fact that myocardial T1 value and ECV are calculated by taking the average of the myocardium and blood pool T1 values which can make them tolerant to some mis-segmentations.

From the T1 mapping analysis, similar to the works of [Puyol-Antón et al., 2020] and [Thongsongsang et al., 2021], we showed that myocardial native T1 values can be utilized to distinguish healthy groups from myocardial diseases such as Tako-Tsubo syndrome,

amyloidosis, and myocardial infarction. We also noticed that myocardial pathologies with diffuse fibrosis such as DCM and HCM have significantly higher mean myocardial native T1 values than the healthy group. However, the native T1 values were not helpful in separating myocarditis pathology from the healthy group as their difference was not significant. From the ECV analysis, we demonstrated that the automatically computed synthetic ECV can significantly differentiate a healthy group of patients from other cardiac pathology groups such as myocardial infarction, amyloidosis, HCM and DCM. However, it was difficult to use ECV to separate the healthy group from the myocarditis and Tako-Tsubo syndrome pathology groups because they have very similar ECV values.

The main limitation of the QC method is the metric used to classify the segmentation quality. In this research, similar to other quality control studies [Valindria et al., 2017, Chen et al., 2020c, Hann et al., 2021] we utilized the Dice score to categorize the quality of the segmentation result as good or poor. However, the Dice score does not always reflect the true quality of the segmentation, for example, when the segmented structure is small or when there are outliers, as mentioned in [Reinke et al., 2021]. Furthermore, the selection of the Dice score threshold can also affect the performance of the QC method. In future work, other segmentation metrics will be assessed to determine the quality of segmentation more accurately.

In the T1 mapping and ECV analysis, it is difficult to directly compare the T1 values, and ECVs computed from our cohort with other cohorts [Reiter et al., 2018, Puyol-Antón et al., 2020, Thongsongsang et al., 2021]. This is because the T1 values and ECVs vary as the scanner field strength, vendor type and acquisition techniques are changed [Scully et al., 2018]. The other limitation is in our cohort some pathology groups such as amyloidosis and Tako-Tsubo syndrome have a small number of cases because these diseases are less common. Due to this, it can be difficult to draw conclusions from their T1 and ECV values. Furthermore, it will be essential to validate the generalizability of the proposed framework in a large-scale cohort for the framework to be used clinically, even though there are limited datasets that have both native and post-contrast T1 mapping images for the ECV analysis.

5.5/ CONCLUSION

In this chapter, we proposed a novel uncertainty-based quality control that utilizes image-level uncertainty features to detect failed segmentation results from native and post-contrast T1 mapping images. The proposed framework is applied to automatically characterize myocardial tissues of various cardiac diseases using T1 mapping and ECV analysis. For the segmentation network, we utilized a Bayesian Swin transformer-based U-Net to generate segmentation maps of the heart structures and uncertainty maps. A random

forest classifier/regressor that uses the image-level uncertainty features is proposed to determine the quality of the segmentation results. Compared to other state-of-the-art uncertainty-based QC methods, our method achieved the highest area under the ROC curve (AUC) for classification and the lowest MAE Dice for regression with less computational complexity and training time. From the results, we noticed that these improvements were notably higher when the dataset is more challenging or when the performance of the segmentation model is poor. This shows how robust our method is in detecting inaccurate segmentations. From T1 mapping and ECV analysis, we automatically characterize the myocardial tissues of cases with different cardiac pathologies from their mean native myocardial T1 values and mean ECV. From the results, we observed that the automatically computed T1 mapping and ECV values can significantly differentiate a healthy group from myocardial diseases like myocardial infarction, amyloidosis, HCM and DCM.

6

Uncertainty-based and Feature space-based Out-of-distribution detection in cardiac MRI segmentation

In the previous chapter (Chapter 5), we proposed an uncertainty-based quality control framework to detect inaccurate segmentations and exclude them before proceeding with further analysis. However, the model can still fail on unusual out-of-distribution (OOD) inputs which are very different from the training images. This motivates developing techniques to explicitly detect OOD inputs before segmentation, as explored in this chapter. By identifying OOD images upfront, we can reject them to avoid unreliable downstream analysis. While uncertainty helps indicate errors on nearby distributions, additional OOD detection better handles distant deviations. Our proposed framework leverages the segmentation model's rich feature representations and similarity metrics to recognize outlier inputs. Together with the quality control approach in Chapter 5, the OOD detection technique aims to further improve the reliability of cardiac MR analysis on real-world data. Detecting outliers complements quality control to make segmentation systems more robust and reliable.

6.1/ Introduction

Deep learning-based models have achieved remarkable performance in medical image segmentation tasks, including cardiac MRI segmentation, and the models are trained under the assumption that the test data will be drawn from the same distribution as the training data. However, when these models are deployed in a clinical setting, test samples may deviate from the in-distribution (ID) and fall under the out-of-distribution (OOD)

samples category. OOD samples in medical image segmentation can arise due to various reasons, such as changes in scanners, acquisition protocols, clinical centers, or pathology, which can lead to domain shifts, referred to as near-OOD (OOD images whose distribution are near ID) [Yang et al., 2021]. We could also have some cases where the samples come from a slightly different cardiac MR imaging modality, which we call mild OOD. Or, the samples could come from a completely different domain or task, such as CT scans of the abdomen or natural images. These would be far-OOD cases, where the distribution of the OOD images is very far from that of the in-distribution (ID) images. As a result, it is essential to develop models that can handle OOD samples in medical image segmentation to ensure safe and effective clinical practice.

Several studies have proposed methods for detecting OOD data in image classification. Hendrycks et al. [Hendrycks et al., 2016] proposed a baseline method that uses a classifier's Maximum Softmax Probability (MSP) to detect out-of-distribution examples without requiring additional network training. The method works based on the observation that a well-trained neural network typically assigns higher softmax scores to in-distribution examples compared to out-of-distribution. Guo et al. [Guo et al., 2017] have enhanced the MSP method by incorporating temperature scaling in the softmax function to increase the softmax score difference between in-and out-of-distribution examples. Liang et al. [Liang et al., 2017] employed temperature scaling and input preprocessing to further enhance the gap between in- and out-of-distribution samples.

Some works used distance metrics in the feature space to identify OOD images [Yang et al., 2021]. Lee et al. [Lee et al., 2018] suggested fitting a Gaussian discriminant model on the last hidden layer of a pre-trained model and used the minimum Mahalanobis distance to the class centroids to detect the OOD images. Other studies utilized the cosine [Techapanurak et al., 2020] and Euclidean distances [Huang et al., 2020] between the class centroids and the input's embedding to detect OOD samples [Yang et al., 2021].

Other OOD detection techniques involve modifying the network architecture or training process. Devries and Taylor, [Devries et al., 2018a] appended an OOD scoring branch onto a classification network. The trained model outputs confidence estimates for each input, which is used to differentiate between in and out-of-distribution examples. Lee et al. [Lee et al., 2017] proposed an OOD detection method that involves jointly training a classifier and a generator to detect and generate out-of-distribution samples by minimizing their losses alternatively. Other methods [Hendrycks et al., 2018, Vyas et al., 2018] also trained an OOD detector by exposing the model to OOD samples during training.

Another approach for detecting OOD samples is based on reconstruction models such as auto-encoders (AEs) [Guo et al., 2018, Denouden et al., 2018], variational auto-encoders (VAEs) [An et al., 2015], and GANs [Perera et al., 2019]. In this approach, the model is trained with a reconstruction loss using in-distribution (ID) data. Once the model is

trained, it is assumed that the reconstruction of unseen OOD samples will fail since they deviate from the ID distribution [Berger et al., 2021]. This makes it possible to detect OOD samples based on the reconstruction error, which is usually higher for OOD samples compared to ID samples. The problem with reconstruction models based OOD detection is that it requires training an auxiliary network (reconstruction model) using the ID images in addition to the main task. This can be time-consuming and computationally expensive, especially for large datasets.

Some works have used prediction uncertainty to detect OOD samples. The most commonly used uncertainty estimation methods are deep ensemble [Lakshminarayanan et al., 2016] and MC-dropout [Gal et al., 2015]. Sample variance among the predictions [Lambert et al., 2022a] or the entropy of the predicted class probability [Mehrtash et al., 2019] is mostly used as an uncertainty metric to detect the OOD samples. However, other image-level uncertainty metrics have not been explored and used by these methods.

A few works have been proposed for OOD detection in medical image segmentation. Mehrtash et al. [Mehrtash et al., 2019] utilized prediction uncertainty to determine the segmentation quality and detect OOD inputs. As shown by [Chen et al., 2020c, Arega et al., 2023a], information extracted from predictive uncertainty can be useful in detecting errors of segmentation models when the models perform poorly in the in-distribution test set images [Arega et al., 2023a]. However, predictive uncertainty information may not be helpful when the samples are far from the training data distribution as the models output empty segmentation outputs with high confidence, as shown in this research. Lambert et al. [Lambert et al., 2022a] studied that the predictive uncertainty of binary segmentation models, which focuses only on anatomical segmentation, fails to detect OOD inputs. However, they found that incorporating anatomical label segmentation alongside lesion segmentation (multi-class segmentation) could improve OOD detection in Multiple Sclerosis lesions segmentation. Other works like Gonzalez et al. [González et al., 2021] used features extracted from the latent space of the segmentation network and Mahalanobis distance between the latent space features of the test image and the training images to detect OOD inputs, whereas Karimi et al. [Karimi et al., 2023] utilized the features from the penultimate layer of the segmentation model and Euclidean distance to detect the OOD inputs. However, these works only focused on features extracted from only two specific layers of the network and did not take advantage of the features extracted from the other parts of the network.

In this chapter, we proposed a post-hoc OOD detection method that leverages the features extracted from the encoder layers of a pre-trained segmentation model to differentiate in-distribution images from out-of-distribution images in cardiac MR segmentation. A 2D U-Net segmentation model was pre-trained on a publicly available short-axis

cine CMR dataset. To measure the distance between encoder features, we studied and compared the two commonly used distance metrics: Euclidean and Mahalanobis distances. We also studied uncertainty-based OOD detection methods and leveraged a Dice coefficient-based image-level uncertainty metric to detect OOD images in segmentation tasks efficiently. Furthermore, we investigated the correlation between the Mahalanobis distance and the segmentation quality, as well as the correlation between uncertainty scores and the segmentation accuracy.

6.2/ DATASETS

6.2.1/ IN-DISTRIBUTION DATASET

As an in-distribution (ID) dataset, we used a publicly available dataset called Automatic Cardiac Diagnosis Challenge (ACDC) [1] dataset [Bernard et al., 2018]. The dataset consists of short-axis cine cardiac MR images of 100 patients acquired at the University Hospital of Dijon using two MRI scanners of different magnetic strengths (1.5 T (Siemens Area, Siemens Medical Solutions, Germany) and 3.0 T (Siemens Trio Tim, Siemens Medical Solutions, Germany)). The cine MR images were acquired in breath hold with a retrospective or prospective gating and with an SSFP sequence in short axis orientation. The CMR images have a spatial resolution ranging from 1.37 to 1.68 $mm^2/pixel$ and a slice thickness of 5-10 mm. The cardiac structures (left-ventricular cavity, left-ventricular myocardium, and right-ventricular cavity) were segmented manually by clinical experts at the end-diastolic (ED) and end-systolic (ES) phases [Bernard et al., 2018].

As a pre-processing step, all ID images were pre-processed by resizing the images to have a spatial size of 400 by 400 and by normalizing the intensity of every image to have zero mean and unit variance. The dataset was shuffled and randomly split into three subsets, 60% for the training dataset (1163 slices), 15% for the validation dataset (291 slices), and 25% for the testing dataset (448 slices).

6.2.2/ OUT-OF-DISTRIBUTION DATASETS

The out-of-distribution images include artificially transformed in-distribution (ACDC) test images, cine cardiac MRI images acquired using different MR scanners or imaging protocols, cardiac images from different MR modalities, and non-cardiac images such as abdominal and lung CT scans and ADE20K scene-centric natural images. For all OOD images, we applied the same pre-processing steps as ID images.

[1] https://www.creatis.insa-lyon.fr/Challenge/acdc/

Table 6.1: Summary of the in-distribution and OOD datasets used in this research.

Type	Dataset	Modality	No. of Images
In-distribution (ID)	ACDC	cine CMR	448
Near OOD	RandomBiasField	cine CMR with random bias field	448
	RandomMotion	cine CMR with random motion artifact	448
	RandomNoise	cine CMR with Gaussian noise	448
	RandomGamma	cine CMR with contrast enhancement	448
	Adversarial	cine CMR with adversarial examples	191
	M&Ms	cine CMR from different centers,vendors	391
Mild OOD	Native_T1	Cardiac native T1 mapping MRI	279
	PostContrast_T1	Cardiac post-contrast T1 mapping MRI	279
	Emidec_LGE	Cardiac LGE MRI	358
	Camus_US	Cardiac Ultrasound (echocardiography)	285
Far OOD	Abdominal_CT	CT scans of abdomen	589
	Lung_CT	CT scans of lung	918
	ADE_RGB	Natural images	64

The OOD datasets are grouped into three sub-categories: near OOD, mild OOD, and far OOD, depending on their similarity to the ID images. The near OOD datasets include cine cardiac MR images from different centers or vendors like the M&Ms dataset and artificially transformed in-distribution (ACDC) test images such as random bias field, random motion artifact, contrast enhancement, Gaussian noise as well as adversarial images. In the mild OOD datasets, we include cardiac images but from different modalities such as native T1 mapping, post-contrast T1 mapping, late gadolinium enhancement cardiac MR, and cardiac ultrasound or echocardiogram. In the far OOD datasets, we include datasets far from cardiac MR, such as abdominal CT scans, lung tumor CT scans, and natural RGB images. A summary of the datasets is shown in Table 6.1.

6.2.2.1/ Artificially Transformed ACDC

For the near-OOD dataset, we artificially transformed the in-distribution (ACDC) test images to emulate some imaging artifacts such as random bias field, random motion artifacts, contrast enhancement, and Gaussian noise [Pérez-García et al., 2020]. More specifically, we used the TorchIO library [Pérez-García et al., 2020] to apply the following parameters: for the random bias field, the bias field was modeled as a linear combination of polynomial basis functions with polynomial coefficients of (0.5, 1.5) and an order of the basis polynomial functions of (3, 5) for MRI magnetic fields. For random motion artifacts, we used simulated movements within the range of (10,20), with a rotation range of 60 degrees and a translation range of 60 mm. For contrast enhancement, we used gamma values within the range of (-3.5, 3.5), and for Gaussian noise, we used a mean of 0 and a standard deviation of (0.05, 0.25). Each of these transformations was applied to every ID

test image, resulting in 448 slices for each transformation.

6.2.2.2/ Multi-Centre, Multi-Vendor & Multi-Disease Cine Cardiac MRI - M&Ms

The M&Ms[2] challenge dataset consists of short-axis cine cardiac MR images scanned in five clinical centers in Spain and Germany using three different MR vendors [Campello et al., 2021, Martín-Isla et al., 2023]. This can be an ideal near-OOD dataset as its distribution deviates from the training data distribution due to changes in acquisition protocol and MR scanner. We used cine cardiac MR images of 36 patients (391 slices) as an OOD.

6.2.2.3/ Adversarial ACDC

As an OOD, we also used adversarial examples of the in-distribution ACDC test dataset. These are intentionally created examples designed to deceive a model into producing inaccurate predictions. These adversarial examples were generated using the Dense Adversary Generation (DAG) algorithm [Xie et al., 2017a]. The algorithm works by utilizing an incorrect segmentation mask and computing a minimum perturbation that will change the prediction of a set of non-background pixels from the correct classification to an incorrect classification [Xie et al., 2017a, Paschali et al., 2018]. As OOD images, we utilized 191 slices.

6.2.2.4/ Cardiac Native and Post-contrast T1 mapping MRI

In-house native T1 mapping and post-contrast T1 mapping CMR images consisting of 93 subjects that have different cardiac pathologies. The images were collected from different clinical centers in France. Each image was acquired using Siemens 1.5T MRI scanner. Modified Look-Locker inversion recovery (MOLLI) was utilized to capture the native and post-contrast T1 mapping images. Each patient has three short-axis slices for each modality (native T1 mapping and post-contrast T1 mapping) [Arega et al., 2023a].

6.2.2.5/ Cardiac LGE MRI - EMIDEC

The Automatic Evaluation of Myocardial Infarction from Delayed-Enhancement Cardiac MRI challenge (EMIDEC)[3] is a MICCAI 2020 challenge that focuses on the segmentation of cardiac structures and myocardial infarction from Late gadolinium enhancement (LGE)

[2]https://www.ub.edu/mnms/
[3]http://emidec.com/

cardiac MR images. The acquisitions were obtained using Siemens MRI scanners (Area (1.5 T) and Skyra (3T)) during conventional cardiovascular exams with no specific protocol [Lalande et al., 2020]. As an OOD, we used LGE CMR images of 50 patients (358 slices).

6.2.2.6/ CARDIAC ULTRASOUND - CAMUS

The Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) dataset [4] contains 2D echocardiographic sequences of patients with two and four-chamber views. It was acquired from GE Vivid E95 ultrasound scanners (GE Vingmed Ultrasound, Horten Norway), with a GE M5S probe (GE Healthcare, US) at the University Hospital of St Etienne (France) [Leclerc et al., 2019]. As an OOD, we used cardiac ultrasound images (two and four-chamber views) of 10 patients (285 slices).

6.2.2.7/ ABDOMINAL CT SCAN - BCVA

The Multi Atlas Labeling Beyond the Cranial Vault - Abdomen challenge (BCVA)[5] is a 2015 MICCAI abdominal segmentation challenge [Landman et al., 2015]. The dataset consists of 30 abdominal CT scans. The CT scans were acquired during the portal venous contrast enhancement phase at Vanderbilt University Medical Center. As an OOD, we used CT scans of 8 patients (589 slices).

6.2.2.8/ LUNG CT SCAN - TCIA

The lung CT scan dataset is comprised of patients with non-small cell lung cancer from Stanford University's publicly available through The Cancer Imaging Archive (TCIA) [6]. As an OOD, we used CT scans of 5 patients (918 slices).

6.2.2.9/ ADE20K RGB

The ADE20K [7] dataset is a large-scale scene-centric dataset that focuses on recognizing and segmenting objects and "stuff" (non-object regions like water or sky) from natural images (RGB images) [Zhou et al., 2016, Zhou et al., 2017]. In our OOD experiment, we utilized 64 images from the ADE20K dataset.

[4] https://www.creatis.insa-lyon.fr/Challenge/camus/index.html
[5] https://www.synapse.org/Synapse:syn3193805/wiki/217752
[6] cancerimagingarchive.net
[7] https://groups.csail.mit.edu/vision/datasets/ADE20K/

6.3/ METHODS

6.3.1/ SEGMENTATION NETWORK

To segment the cardiac structures from the CMR images, a 2D U-Net segmentation network was utilized [Ronneberger et al., 2015]. The network consists of an encoder and a decoder part connected via skip connections. The encoder has four convolutional blocks or stages, in which each stage is made up of 2 consecutive convolutional layers. Each convolution is followed by batch normalization and rectified linear unit (ReLU) activation function. At the end of each stage, a max pooling layer is used to reduce the spatial resolution of the feature maps by a factor of 2 while retaining the most important features. After the encoder stages, there is a bottleneck block that has two convolutional layers, each followed by batch normalization and ReLU activation. Each encoder bock (including the bottleneck) doubles the number of channels of the feature maps.

The decoder consists of four transpose convolutional layers that upsample the feature maps by a factor of 2. Each decoder layer is connected to the corresponding encoder layer via a skip connection that concatenates the feature maps from the encoder and the decoder layers. The concatenated feature maps are then fed into a decoder block that performs two 3x3 convolutions followed by batch normalization and ReLU activation. The final convolutional layer in the network is a 1x1 convolution that maps the feature maps to the desired number of output channels. The decoder layer then upsamples the image and concatenates it with the corresponding feature map from the encoder layer, as shown in Fig 6.1.

6.3.2/ UNCERTAINTY-BASED OOD DETECTION

Uncertainty-based OOD detection methods can be promising, particularly in identifying near-OOD images. Most of the methods use uncertainty metrics based on pixel-wise uncertainty, such as predictive entropy [Karimi et al., 2023] and sample variance [Lambert et al., 2022a]. The average of predictive entropy and the average of sample variance are used as image-level uncertainties to decide whether the input sample is an outlier. However, these image-level uncertainty metrics may not work well as they are insensitive to class imbalance. In segmentation, the class distribution is often imbalanced, with the background class being much more prevalent than the objects of interest. Thus taking the average of the pixel-wise uncertainty values can sometimes be misleading, as it may appear low even if the model is uncertain on the minority classes.

In this research, we proposed to use a Dice coefficient-based image-level uncertainty metric called Dice within samples (Dice-ws) to detect OOD samples. It is the average Dice

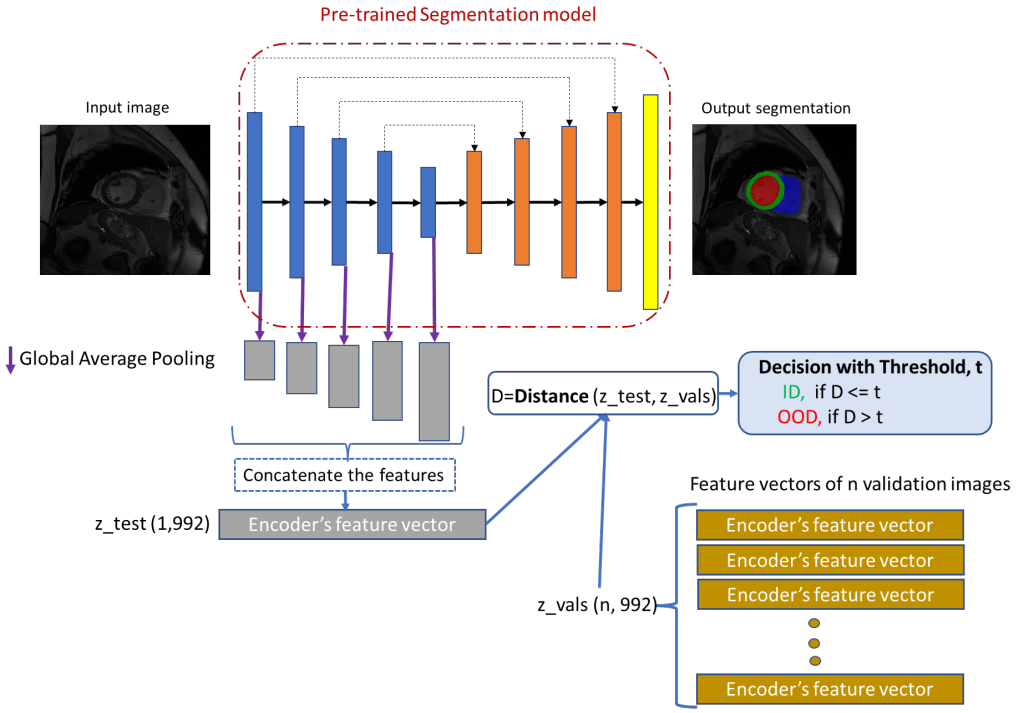

Figure 6.1: Proposed OOD detection method, which leverages the features extracted from the encoder blocks of a pre-trained segmentation model. To reduce the dimensionality of the features, global average pooling is used for each of the extracted feature maps before concatenating them. To measure the similarity between the input image and the validation in-distribution images, a Mahalanobis distance is used. To determine a threshold for this distance, we can either use a distance threshold that achieves a 95% true positive rate on the validation in-distribution dataset [Liang et al., 2017, González et al., 2021] or compute it from the mean and standard deviation of the distances [Karimi et al., 2023] calculated from the validation ID datasets.

coefficient of the mean predicted segmentation and the individual M prediction samples (Eq. 5.4). This metric is more sensitive to class imbalance as it takes into account the overlap between the segmentations.

To generate segmentation uncertainty, we employed two techniques: MC Dropout and Deep Ensemble. For MC Dropout, a dropout with a rate of 0.1 was applied at the middle layers (central Encoder-Decoder position) of the network. For Deep Ensemble, multiple models were trained from scratch with random weight initialization.

6.3.3/ Feature space-based OOD Detection

The encoder of a segmentation network is used to extract high-level features from an input image and create a condensed representation of that image that can be used by the decoder to generate a segmentation map. This is accomplished by using a series of convolutional and pooling layers that increase the number of learned features while

reducing the spatial dimensionality of the input image.

Features extracted by the encoder layers of a segmentation network can be important for OOD detection because they capture the key visual characteristics of the input image. The encoder layers of a segmentation network are designed to capture multi-scale representations of the input image at multiple levels of abstraction, starting with low-level features such as edges and corners and moving up to higher-level features such as object parts, textures, and shapes. This multi-scale representation provides a rich and diverse set of features that can be used to distinguish between in-distribution and out-of-distribution images.

One problem with using all the encoder features is the dimensionality of the encoder features, as it can be very large. To mitigate this problem, global average pooling was applied to each of the extracted encoder feature maps before merging them to create a feature vector representing the input image in the feature space, as seen in Fig 6.1.

For an input image x, each encoder stage of the pre-trained segmentation model produces a feature map $f_l \in R^{n_l,h_l,w_l}$, where n_l is the number of channels, and (h_l, w_l) indicate the spatial-size of the feature map of an encoder stage (block) l. To reduce the dimensionality of each feature map, a global average pooling is applied, resulting in feature map $f_l \in R^{n_l}$. Then all the resized feature maps of the encoder stages are merged to create a single feature vector $(z \in R^k, k = \sum_l n_l)$ that represents the input image, where k is the sum of the number of feature maps of all the encoder stages. In our case, the encoder of the pre-trained segmentation network has five stages (including the bottleneck or latent space). When an input image of size $(400, 400)$ is given, feature maps of size $(32, 200, 200)$, $(64, 100, 100)$, $(128, 50, 50)$, $(256, 25, 25)$, and $(512, 25, 25)$ are generated by each of the encoder stages. Afterward, global average pooling is applied to each feature map to obtain feature maps of size $(32, 1)$, $(64, 1)$, $(128, 1)$, $(256, 1)$, and $(512, 1)$, respectively. Finally, all the resized feature maps are concatenated to form the final feature vector of size $(992, 1)$.

Different distance metrics can be used to measure the similarity between the input image's feature vector and the training images' feature vectors to detect OOD images using the encoder features of a segmentation network. In this research, two commonly used distance metrics in OOD detection are explored: Mahalanobis and Euclidean distances.

The Mahalanobis distance is a multivariate distance metric that takes into account the covariance between features. It measures the distance between a point and the distribution of data. From the encoder features of the validation images, we computed the mean $(\mu \in R^k)$ and covariance $(\Sigma \in R^{k,k})$ of the features to estimate the multivariate Gaussian distribution. During inference, when a new input image x_t is given, its encoder feature $z_t \in R^k$ is extracted and the Mahalanobis distance to the estimated Gaussian distribution is computed as follows (Eq. 6.1):

$$D_{\text{Mah}}(z_t) = \sqrt{(z_t - \mu)^T \Sigma^{-1} (z_t - \mu)} \tag{6.1}$$

Euclidean distance is a simple distance metric that measures the straight-line distance between two vectors in a high-dimensional space. It is sensitive to the scale of the features, so the feature vectors were normalized prior to using Euclidean distance. During inference, when a new input image x_t is given, its encoder feature $z_t \in R^k$ is extracted, and the Euclidean distance between the input image's feature vector z_t and its nearest neighbor from features of the validation set z_v is calculated as shown in Eq. 6.2.

$$D_{\text{Eucl}}(z_t, z_v) = \sqrt{(z_{t_1} - z_{v_1})^2 + (z_{t_2} - z_{v_2})^2 + ... + (z_{t_k} - z_{v_k})^2} \tag{6.2}$$

6.3.4/ Implementation

The 2D segmentation model was trained for 200 epochs, utilizing the ADAM optimizer with a learning rate of 0.001 and a batch size of four. For the segmentation loss, a hybrid loss function was used, which consisted of both cross-entropy and dice loss with equal weights (1.0). In order to generate the segmentation uncertainty, we used 5 Monte Carlo samples or Deep Ensemble samples. The models were implemented using the Pytorch deep learning framework and were trained on NVIDIA Tesla V100 GPUs with 32GB of memory.

6.4/ Results

To evaluate the performance of the OOD detection methods, the following two metrics are employed: false positive rate (FPR) at 95% true positive rate (TPR) [González et al., 2021, Liang et al., 2017] and Area Under the Receiver Operating Characteristic curve (AUC). FPR at 95% TPR represents the probability that a negative (OOD) example is classified wrongly as positive (ID) when the true positive rate (TPR) is as high as 95%. The AUC is a useful metric for evaluating the overall performance of the OOD detection method, as it summarizes the method's ability to distinguish between positive (ID) and negative (OOD) cases over a range of thresholds.

6.4.1/ Ablation Study

To evaluate the effect of the different distance metrics, the performance of Mahalanobis (Mah.) and Euclidean (Eucl.) distances were compared in differentiating ID and OOD images using the encoder features of the images as shown in Table 6.2. The comparison

Table 6.2: Ablation results comparing the performance of different distance metrics and the features extracted from different parts of the segmentation network in near, mild, and far OOD datasets in terms of AUC and FPR at 95% TPR. The better results are highlighted in bold.

Ablation	Method	Near OOD		Mild OOD		Far OOD		Average	
		AUC	FPR	AUC	FPR	AUC	FPR	AUC	FPR
Distance	Eucl.	0.810	0.504	0.989	0.065	1	0	0.91	0.252
	Mah.	**0.842**	**0.401**	**1**	**0**	**1**	**0**	**0.927**	**0.185**
Position	Latent	0.787	0.599	0.998	0.01	1	0	0.901	0.28
	Encoder	**0.842**	**0.401**	**1**	**0**	**1**	**0**	**0.927**	**0.185**
	Decoder	0.652	0.699	0.950	0.241	0.977	0.176	0.819	0.437
	All	0.830	0.429	1	0	1	0	0.922	0.198

was done on the grouped OOD images (near, mild, and far OOD) using AUC and FPR at 95% TPR. The Mahalanobis distance performed the best, achieving an average AUC of 0.933 and FPR at 95% TPR of 0.173. The gain in performance was mostly in the near and mild OOD datasets, whereas in the far OOD datasets, the performance of the two distance metrics was the same.

In our experiments, the performance of the OOD detection methods in terms of the features extracted from different parts of the segmentation network was also studied. These features include latent-space features (latent), encoder features (encoder), decoder features (decoder), and features from all encoder and decoder blocks (all). The size of the latent space feature vector is 512, while encoder features, decoder features, and all features have feature vector sizes of 992, 992, and 1472, respectively. It should be noted that the latent-space feature is also included as part of the encoder and decoder features. To compare the performance of the different features, the Mahalanobis distance metric was utilized. As can be seen from Table 6.2, the encoder and all features performed perfectly with an AUC of 1 and FPR of 0 in both mild and far OOD, whereas in the near OOD datasets, the encoder features achieved slightly better detection results than all features. Looking at the performance of the latent space features, even though it uses a comparatively small-sized feature vector to represent an image, its detection performance was weaker than the encoder and all features. As for the decoder features, its performance was much lower than the other three features, particularly in the near OOD datasets.

6.4.2/ Segmentation Performance

Table 6.3 presents a comparison of the segmentation model's performance on both in-distribution (ID) and out-of-distribution (OOD) datasets. Additionally, the table includes information about uncertainty (Dice-ws) and Mahalanobis distance (Mah. Distance) for each dataset. Note that the encoder features were utilized for computing the Mahalanobis

distance. The table also shows the Pearson Correlation Coefficient (PCC) between the Dice score and Mahalanobis distance (PCC Mah. Dist) as well as the Pearson correlation coefficient between the Dice score and Dice-ws uncertainty (PCC Dice-ws), which are displayed in columns 5 and 6, respectively. The Dice score is obtained by calculating the mean Dice of the three cardiac structures: left ventricular blood pool (LV), myocardium (MYO), and right ventricular blood pool (RV).

The model achieved an average Dice score of 0.808 on the ACDC test dataset images, which are the in-distribution images. However, its performance decreased significantly on the near OOD datasets. Despite this, the model demonstrated relatively good performance on images with random bias fields, random gamma (contrast enhancement), and M&Ms datasets, achieving average Dice scores of 0.687, 0.688, and 0.680, respectively. On the other hand, the model struggled to segment well the artificially transformed ACDC images with random motion artifacts, yielding an average Dice score of 0.560. For the artificially transformed ACDC images with random noise and adversarial artifacts, the model failed to segment them correctly. Furthermore, the standard deviations (std) of the Dice scores of most of the near OOD datasets were very high, indicating that the model performed well on some images that were similar to the ID dataset but very poorly on others that were more dissimilar. Similarly, the model performed poorly in the mild OOD datasets showing the generalization problem of deep learning models when tested on images with different modalities.

Looking at the certainty of the model on the different datasets, generally, the model's certainty decreases when the OOD images are farther away from the ID images. As can be seen from Table 6.3, the certainty in terms of Dice-within-samples (Dice-ws) of the model decreased from 0.922 for ID dataset to 0.880, 0.875, 0.874, 0.866, 0.793, and 0.627 for random noise, random bias field, M&Ms dataset, random gamma, random motion, and adversarial artifacts, respectively. This indicates the strong correlation between the segmentation performance (Dice score) and the certainty of the model for most of the near OOD images (with the exception of the random noise dataset), as shown in column 6 of the table. This is also true for some mild OOD datasets like native T1 and LGE CMR images. However, the certainty starts increasing in the far OOD dataset. For example, the model is more certain of its prediction of the cardiac ultrasound images (0.963) than of the ID images (0.922), even though for the ultrasound cardiac images, it outputs empty segmentation maps. In the far OOD datasets, the images are very different from the ID datasets in which the images are CT scans of different organs or natural images instead of cardiac MR images, and the segmentation model outputs empty segmentation maps with high certainty.

For the Mahalanobis distance, we utilized the encoder features and discussed them in detail in Section 6.4.4.

Table 6.3: Quantitative comparison of the segmentation model's performance in the ID and OOD datasets in terms of Dice score as well as the Mahalanobis distance and certainty scores of the OOD detection methods. PCC (Mah. Dist) denotes the Pearson correlation coefficient between Dice Score and Mahalanobis Distance, while PCC (Dice-ws) represents the Pearson correlation coefficient between Dice Score and Dice-ws certainty. The values displayed are the mean values, while those inside the parentheses represent the standard deviations.

Datasets	**Dice Score**	**Mah. Distance**	**Uncertainty** (Dice-ws)	**PCC** (Mah. Dist)	**PCC** (Dice-ws)
ACDC (ID)	0.808 (0.24)	4730 (1160)	0.922 (0.14)	-0.149	0.744
RandomBiasField	0.687 (0.33)	25,170 (61,805)	0.875 (0.17)	-0.481	0.644
RandomMotion	0.560 (0.32)	9194 (2623)	0.793 (0.20)	-0.260	0.632
RandomNoise	0.424 (0.41)	167,417 (167,397)	0.880 (0.21)	-0.722	0.134
RandomGamma	0.688 (0.33)	34,930 (78,284)	0.866 (0.19)	-0.607	0.755
Adversarial	0.237 (0.25)	5023 (1327)	0.627 (0.20)	0.046	0.396
M&Ms	0.680 (0.29)	9085 (4828)	0.874 (0.15)	-0.112	0.632
Native_T1	0.219 (0.21)	15,460 (3815)	0.609 (0.19)	-0.324	0.454
PostContrast_T1	0 (0)	13,520 (3164)	0.779 (0.21)	-0.05	0.0
Emidec_LGE	0 (0)	45,107 (20,613)	0.590 (0.19)	-0.245	0.582
Camus_US	- (-)	25,608 (6179)	0.963 (0.10)	-	-
Abdominal_CT	- (-)	892,542 (87,555)	0.917 (0.14)	-	-
Lung_CT	- (-)	33,345 (16,276)	0.827 (0.19)	-	-
ADE_RGB	- (-)	72,399 (50,608)	0.829 (0.18)	-	-

6.4.3/ Uncertainty-based OOD Detection

In Table 6.4, we compared the OOD detection performance of the commonly used image-level uncertainty metrics, which are the average of sample variance (IL_var) and predictive entropy (IL_Ent), with our proposed Dice score-based image-level uncertainty metric that is Dice-within-samples (Dice-ws) in the near, mild and far OOD datasets. This comparison was done on both deep ensemble-based and MC-dropout-based uncertainty estimation methods.

For both deep ensemble-based and MC-dropout-based uncertainty, the proposed Dice-within-samples metric outperforms the other two metrics, the average of sample variance and the average of predictive entropy, in almost all of the datasets in terms of AUC and FPR at 95% TPR. The only exception is for random bias field and native T1 datasets where the average of sample variance from deep ensemble outperforms Dice-within-samples in terms of FPR.

Comparing the deep ensemble-based and MC-dropout-based Dice-within-samples method, the former has on average better detection performance than the latter in terms of AUC and FPR. However, the MC-dropout-based Dice-within-samples outperforms the deep ensemble based in terms of FPR, particularly in most of the near OOD datasets and far OOD datasets. The Dice-within-samples-based OOD detection method per-

formed well on adversarial, native T1, and postcontrast T1 datasets, achieving AUCs of 0.895, 0.911, and 0.913, respectively. While its performance was very low in the far OOD datasets and in some of the near OOD datasets. Comparing the performance of the average of sample variance (IL_var) and average predictive entropy (IL_Ent), the former consistently outperforms the latter in all the datasets in terms of both AUC and FPR.

6.4.4/ Comparison with state-of-the-art

The proposed feature space-based OOD detection method is compared to different state-of-the-art OOD detection methods in terms of AUC and FPR at 95% TPR in Table 6.5. The proposed feature space-based method uses the encoder features of the images and utilizes Mahalonobis distance to detect OOD images. The state-of-the-art methods consist of uncertainty-based methods such as Dice-within-samples (Dice-ws), the average sample variance (IL_Var) [Lambert et al., 2022a], maximum softmax probability (MSP) [Hendrycks et al., 2016], and softmax with temperature scaling (Temp_Scale) [Guo et al., 2017] as well as feature space-based OOD detection methods like spectral features (Spectral) [Karimi et al., 2023], and latent space features (Latent_Space) [González et al., 2021]. For the uncertainty-based methods, we used deep ensemble [Lakshminarayanan et al., 2016] as our uncertainty estimation method, as it yielded the best OOD detection result, as indicated in Table 6.4. Additionally, we assessed temperature scaling [Guo et al., 2017] with three different temperatures (10, 100, and 1000) and found that a temperature of 100 yielded the best results, which we reported in the table. The OOD detection methods are evaluated in the 13 OOD datasets, as outlined in Section 6.2, to assess their ability to detect OOD images.

As can be seen from Table 6.5, the proposed method achieved the best result in terms of AUC and FPR, outperforming the other state-of-the-art OOD detection methods. Compared to the uncertainty-based OOD detection methods, the performance improvement was in almost all the datasets, except for the adversarial dataset, where Dice-ws performed better. Compared to the feature space-based OOD detection methods, the performance enhancement was mostly in the near and mild OOD datasets. In the far OOD datasets, the spectral and latent space feature-based methods perfectly detected the OOD images similar to the proposed method with almost an AUC of 1 and FPR of 0.

The proposed method exhibited excellent performance in the mild and far OOD datasets, achieving an AUC of 1 and an FPR of 0. However, in the near OOD dataset, although our method still outperformed the other OOD detection methods, its performance was lower, especially in terms of FPR. For instance, in the M&Ms, random bias field, random gamma, and adversarial datasets, the FPR values were 0.355, 0.355, 0.46, and 0.948, respectively. Additionally, all the other feature space-based OOD detection methods failed

Table 6.4: OOD detection performance comparison of deep ensemble-based and MC-dropout-based image-level uncertainty metrics in terms of AUC and FPR at 95% TPR. Dice-ws represents the Dice within samples, IL_Var denotes the average of sample variance, and IL_Ent for the average predictive entropy. The bold results are better.

Datasets	**Deep Ensemble**						**MC-Dropout**					
	Dice-ws		**IL_Var**		**IL_Ent**		**Dice-ws**		**IL_Var**		**IL_Ent**	
	AUC	**FPR**	**AUC**	**FPR**	**AUC**	**FPR**	**AUC**	**FPR**	**AUC**	**FPR**	**AUC**	**FPR**
RandomBiasField	**0.619**	0.908	0.557	**0.897**	0.541	0.902	0.598	**0.897**	0.527	0.906	0.501	0.915
RandomMotion	0.759	0.801	0.68	0.79	0.615	0.819	**0.772**	**0.734**	0.66	0.801	0.575	0.83
RandomNoise	0.432	0.875	0.368	0.897	0.327	0.924	**0.453**	**0.864**	0.383	0.931	0.34	0.944
RandomGamma	**0.599**	**0.886**	0.554	0.924	0.522	0.935	0.572	0.895	0.499	0.924	0.472	0.942
Adversarial	**0.895**	**0.555**	0.603	0.88	0.408	0.921	0.719	0.822	0.363	0.958	0.262	0.942
M&Ms	**0.626**	0.916	0.524	0.852	0.505	0.854	0.6	**0.849**	0.524	0.87	0.492	0.862
Native_T1	**0.911**	0.52	0.811	**0.47**	0.744	0.509	0.894	0.545	0.803	0.509	0.739	0.502
PostContrast_T1	**0.694**	0.778	0.245	0.957	0.149	0.989	0.544	**0.746**	0.234	0.971	0.124	0.996
Emidec_LGE	**0.913**	**0.556**	0.489	0.958	0.232	0.986	0.714	0.751	0.209	0.978	0.131	0.989
Camus_US	**0.223**	0.979	0.025	1	0.012	1	0.153	**0.972**	0.025	1	0.027	1
Abdominal_CT	**0.443**	0.946	0.102	0.988	0.044	0.995	0.352	**0.907**	0.095	0.986	0.035	0.997
Lung_CT	**0.664**	0.873	0.124	0.995	0.057	1	0.543	**0.855**	0.124	0.996	0.087	0.997
ADE_RGB	**0.551**	**0.922**	0.242	0.953	0.125	1	0.291	0.938	0.066	1	0.053	1
Average	**0.641**	**0.809**	0.41	0.889	0.329	0.91	0.554	0.829	0.347	0.91	0.295	0.917

to properly detect the adversarial images, whereas the uncertainty-based Dice-within-samples method achieved the best result with an AUC of 0.895 and an FPR of 0.555. Among all the methods, the MSP and temperature scaling, which are based on the softmax scores, have the worst OOD detection result. They achieved the lowest AUC and the highest FPR in most of the OOD datasets.

In Figure 6.2, we computed and analyzed the Mahalanobis distance (of the proposed feature space-based OOD detection method) of the ID and OOD datasets from the ID validation set using box plots. The ID images, which are the ACDC test images, obtained a mean and standard deviation Mahalanobis distance of $4730(1160)$, with very few outliers as shown in Figure 6.2 and Table 6.3 (second column). Among the near OOD datasets, the adversarial dataset has the nearest distance to the ID images with a mean and standard deviation Mahalanobis distance of $5023(1327)$. For the M&Ms dataset, even though the mean Mahalanobis distance is near to the ID images, there are some outliers that are very far. Random bias field, random gamma, and random noise have the farthest distance from the ID images and have many outliers, and particularly random noise has the highest standard deviation of the Mahalanobis distance. For the mild OOD datasets, they have a mean Mahalanobis distance ranging from 13,520 to 45,107 and their standard deviation is relatively low, with the exception of the LGE dataset. Mahalanobis distances of the far OOD datasets are characterized by their very high mean and high standard deviation with many outliers. Among all OOD datasets, the abdominal CT scans have the highest mean Mahalanobis distance (892,542), making it the farthest OOD dataset, even greater than the natural RGB images.

Furthermore, we assessed the qualitative result of the segmentation model and its pixel-wise sample variance uncertainty (deep ensemble based) and its image-level uncertainties like the average of sample variance and Dice-within-samples as well as the Mahalanobis distance of the proposed feature-based OOD detection in Figure 6.3 and Figure 6.4. In Figure 6.3, which contains the qualitative results of the ID and near OOD datasets, the segmentation model either segmented the image very well with low uncertainty or segmented the image inaccurately but with high uncertainty. In this case, both the image level uncertainty and the Mahalonobis distance provide a good indication of the quality of the segmentation result with the exception of the adversarial dataset where the feature space-based OOD detection method has confused the adversarial images with the ID images because they look very similar. Looking at the Mahalonobis distances, the ACDC ID image has a Mahalonobis distance of 5392, whereas the adversarial version of the same image has a lower Mahalonobis distance which is 4827. However, the image level uncertainty (Dice-within-samples) of the ACDC ID image is much lower than its adversarial version by more than 30%. The prediction of the segmentation model is also empty. This shows the advantage of the uncertainty-based OOD detection method in detecting adversarial images, which look very similar to the original images.

Figure 6.4 contains the qualitative results of the mild and far OOD datasets. For the images of these datasets, the segmentation model predicted an empty segmentation map except for the native T1 mapping image. The model's uncertainty is low for most of the empty predictions; the model is confident with its prediction. This makes the uncertainty very difficult for detecting mild and far OOD images. However, the Mahalanobis distance provided a better estimation of how similar the images are with respect to the ID images, as the distance becomes larger as the images are more dissimilar to the ID images. For example, the native T1 mapping image, which is the most similar image among the mild and far OOD datasets, has a Mahalanobis distance of 13,138, whereas the RGB natural image, which is considered the most dissimilar image, has a Mahalanobis of 56,290.

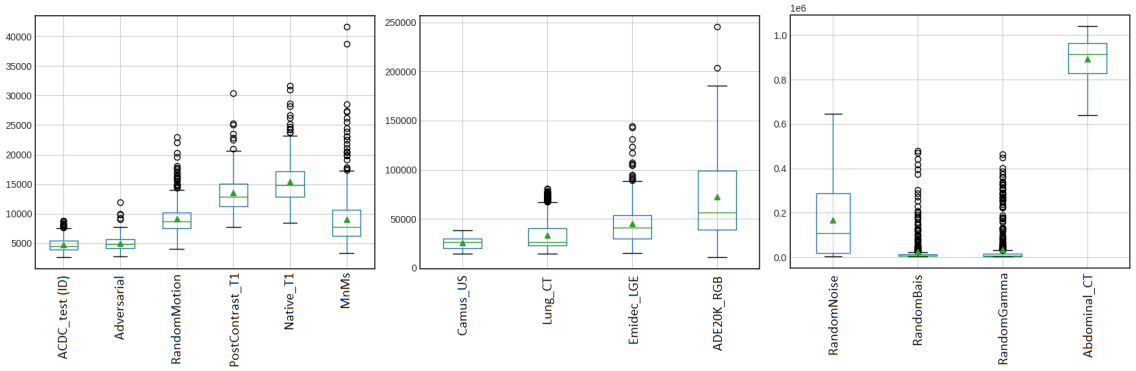

Figure 6.2: Box plots comparing the Mahalanobis distances (proposed method) of the ID (ACDC test set) and the 13 different OOD datasets

6.5/ DISCUSSION

In this chapter, we proposed a simple post-hoc OOD detection method that can be used with any pre-trained segmentation model. Our method uses the features extracted from the encoder blocks of the segmentation model and Mahalanobis distance to distinguish the different OOD datasets from the in-distribution cardiac cine MR dataset. We compared its detection performance with state-of-the-art uncertainty-based and feature space-based OOD detection methods. In addition, we studied the correlation between the Mahalanobis distance and the segmentation quality, as well as the correlation between uncertainty scores and the segmentation quality.

Regarding the distance metrics, the Mahalanobis distance metric outperformed Euclidean distance in distinguishing ID and OOD images. This indicates that considering the covariance between the features, as captured by the Mahalanobis distance, is crucial for effective OOD detection. However, it is worth noting that the performance gains were more prominent for near and mild OOD datasets, suggesting that the covariance information is more discriminative as the OOD datasets are less dissimilar.

Table 6.5: Quantitative comparison of different uncertainty and feature space-based OOD detection methods: Dice-within-samples (Dice-ws), the average sample variance (IL_Var) [Lambert et al., 2022a], maximum softmax probability (MSP) [Hendrycks et al., 2016], and softmax with temperature scaling (Temp_Scale) [Guo et al., 2017], spectral features (Spectral) [Karimi et al., 2023], latent space features (Latent_Space) [González et al., 2021] and the Proposed method in 13 different OOD datasets in terms of AUC and FPR at 95% TPR. The bold results are better. RandomBiasField: Random bias field, RandomMotion: random motion artifact, RandomNoise: Gaussian noise, RandomGamma: contrast enhancement, Adversarial: adversarial images, Native_T1: native T1 mapping, PostContrast_T1: post-contrast T1 mapping, Emidec_LGE: late gadolinium enhancement cardiac MR, Camus_US: cardiac ultrasound or echocardiogram, Abdominal_CT: abdominal CT scans, Lung_CT: lung tumor CT scans, and ADE_RGB: natural RGB images.

Datasets	**Dice-ws**		**IL_Var**		**MSP**		**Temp_Scale**		**Spectral**		**Latent_Space**		**Proposed**	
	AUC	**FPR**	**AUC**	**FPR**	**AUC**	**FPR**	**AUC**	**FPR**	**AUC**	**FPR**	**AUC**	**FPR**	**AUC**	**FPR**
RandomBiasField	0.619	0.908	0.557	0.897	0.529	0.906	0.752	0.772	0.801	0.518	0.849	0.467	**0.887**	**0.355**
RandomMotion	0.759	0.801	0.68	0.79	0.613	0.806	0.759	0.828	0.879	0.475	0.847	0.571	**0.971**	**0.129**
RandomNoise	0.432	0.875	0.368	0.897	0.326	0.924	0.231	0.982	0.882	0.257	0.884	0.257	**0.926**	**0.163**
RandomGamma	0.599	0.886	0.554	0.924	0.509	0.933	0.677	0.77	0.769	0.54	0.722	0.621	**0.803**	**0.46**
Adversarial	**0.895**	**0.555**	0.603	0.88	0.405	0.916	0.411	0.974	0.671	0.921	0.659	0.942	0.572	0.948
M&Ms	0.626	0.916	0.524	0.852	0.502	0.854	0.641	0.793	0.714	0.775	0.765	0.739	**0.898**	**0.355**
Native_T1	0.911	0.52	0.811	0.47	0.721	0.53	0.743	0.703	0.941	0.272	0.997	0.018	**1**	**0**
PostContrast_T1	0.694	0.778	0.245	0.957	0.145	0.978	0.211	1	0.993	0.036	0.996	0.022	**1**	**0**
Emidec_LGE	0.913	0.556	0.489	0.958	0.182	0.986	0.468	0.978	**1**	**0**	**1**	**0**	**1**	**0**
Camus_US	0.223	0.979	0.025	1	0.011	1	0.008	1	**1**	**0**	**1**	**0**	**1**	**0**
Abdominal_CT	0.443	0.946	0.102	0.988	0.048	0.993	0	1	**1**	**0**	**1**	**0**	**1**	**0**
Lung_CT	0.664	0.873	0.124	0.995	0.051	1	0.036	1	0.994	0.022	**1**	**0**	**1**	**0**
ADE_RGB	0.551	0.922	0.242	0.953	0.097	1	0.097	1	0.999	0.016	**1**	**0**	**1**	**0**
Average	0.641	0.809	0.41	0.889	0.318	0.91	0.387	0.908	0.896	0.295	0.901	0.28	**0.927**	**0.185**

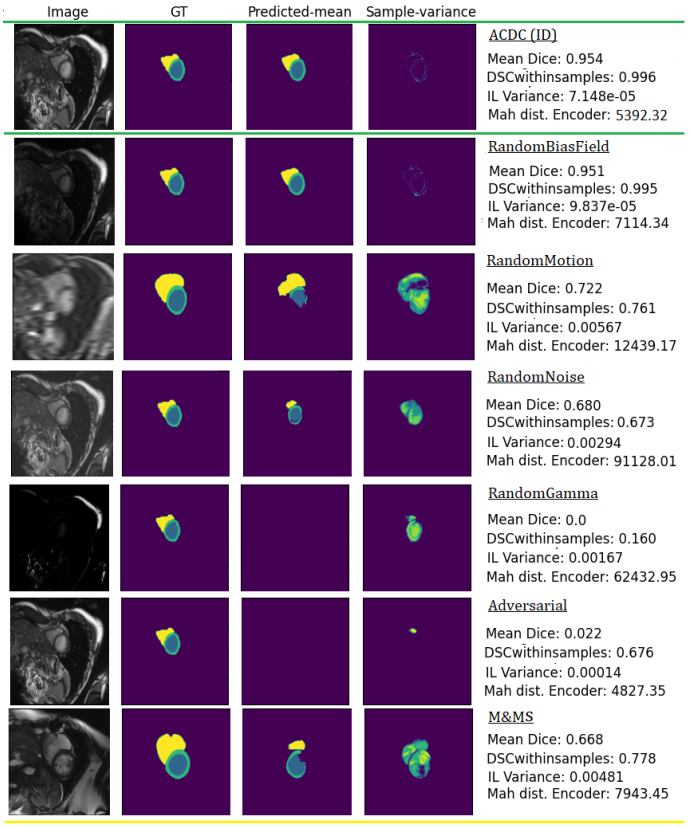

Figure 6.3: Qualitative results of ID and near OOD datasets with their corresponding mean Dice score, Dice within samples uncertainty, image level sample variance (IL_Variance) uncertainty, and Mahalanobis distance for the encoder features (proposed method). Image: the input image, GT: ground truth, Predicted-mean: the final prediction, Sample-variance: the pixel-wise uncertainty

The encoder features exhibited superior performance in OOD detection compared to other features, particularly in the near OOD images. This shows that the shifts in data are well captured by the encoder features more than the other features. This can be due to the fact that the encoder layers capture multi-scale representations of the input image at multiple levels of abstraction. All features, which combine encoder and decoder features, achieved similar performance as the encoder features. However, its feature vector size is 48% larger than encoder features. On the other hand, the weaker performance of decoder features implies that this region does not capture sufficient discriminative information for OOD detection.

The segmentation model's performance varied across different OOD datasets. It achieved relatively good segmentation results on images with random bias fields, random gamma, and the M&Ms dataset, indicating robustness to certain types of artifacts and variations. The model's poor performance on near OOD datasets, such as those with random noise and adversarial artifacts, demonstrates the difficulty of the model in handling images with Gaussian noises and adversarial perturbations. The high standard deviation of the Dice

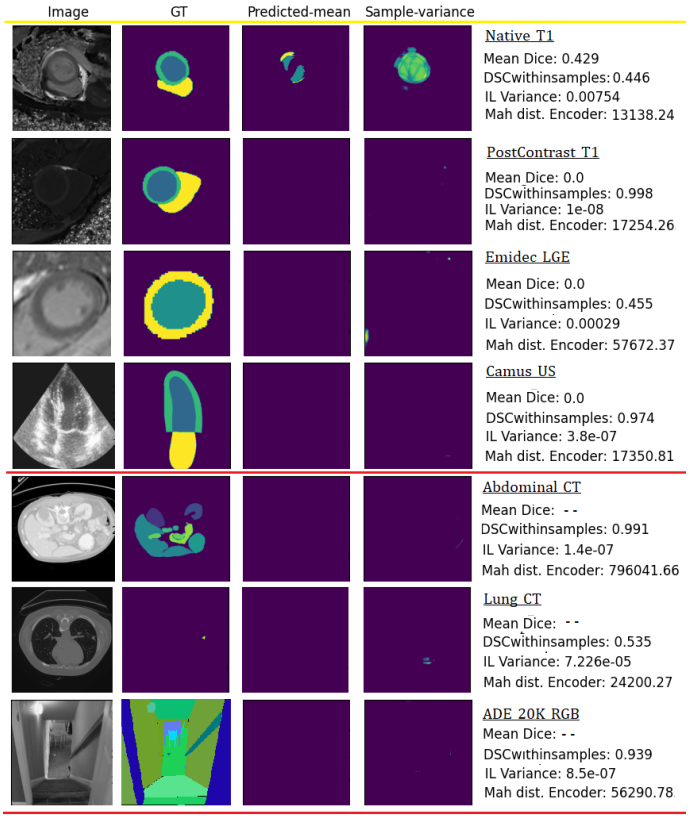

Figure 6.4: Qualitative results of Mild and Far OOD datasets with their corresponding mean Dice score, Dice within samples uncertainty, image level sample variance (IL_Variance) uncertainty, and Mahalanobis distance for the encoder features (proposed method). Image: the input image, GT: ground truth, Predicted-mean: the final prediction, Sample-variance: the pixel-wise uncertainty

scores in the near OOD datasets indicates that the model was very sensitive to the level of artifacts added to the images. This is because when the amount of artifacts is light, the model's segmentation was good, but its performance decreased as the amount of artifacts increased.

In relation to the association between the Mahalanobis distance, uncertainty (Dice-ws), and segmentation quality (Dice score), we found that the Mahalanobis distance has a weak correlation with the segmentation quality. This signifies its inability to distinguish poorly segmented images in both the in-distribution (ID) and most of the near OOD datasets. On the other hand, in agreement with the findings of Arega et al. [Arega et al., 2023a], the uncertainty-based approach (Dice-ws) demonstrates a strong correlation with the segmentation quality and effectively identifies suboptimal segmentation results, notably in the ID dataset and several near OOD datasets, such as random gamma, random bias field, random motion, and the M&Ms datasets. This outcome is anticipated because the feature space-based method lacks sufficient information pertaining to the segmentation result, given that the encoder features primarily reflect information

related to the input image, whereas the estimated uncertainties reflect the confidence of the deep ensemble or MC-Dropout segmentation models in their segmentation results.

The results from uncertainty-based OOD detection show that the Dice-based image level uncertainty metric is more robust and correlates better to the segmentation result than the ones which are based on pixel-wise uncertainty metrics (sample variance and predictive entropy) in both deep ensemble-based and MC-dropout-based uncertainty estimation methods. The weaker performance of the pixel-wise uncertainty metrics can be due to the fact that taking the average of these metrics to estimate the uncertainty at the image level sometimes does not accurately convey the level of uncertainty, especially when there is a class imbalance. The certainty of the model's predictions, as indicated by the Dice-within-samples metric, generally decreased as the OOD datasets deviated further from the ID images. This correlation suggests that the model's segmentation performance is closely linked to its certainty in near OOD scenarios. However, interestingly, in far OOD datasets such as cardiac ultrasound images, and abdominal CT scans, the model exhibited high certainty while producing empty segmentation maps. This indicates that uncertainty-based methods are not good enough for OOD detection because the model's uncertainty is not reliable for some mild and far OOD images.

The proposed feature space-based OOD detection method, which utilizes Mahalanobis distance, achieves the best results in terms of AUC and FPR when compared to other state-of-the-art methods in most of the OOD datasets. Generally, the feature space-based OOD detection methods outperform the uncertainty-based methods, except for adversarial datasets. This tells us that the feature space-based methods capture the data shifts better when the images are more dissimilar than the ID images. However, when the images are confusingly similar to the ID images, like adversarial images, the uncertainty-based method (Dice-within-samples) detects them better than the feature space-based methods. The adversarial perturbation added to the original ID image confuses the segmentation network even though it is visually undetectable. This also makes the adversarial images confusing for the feature space-based OOD detection methods. Since the MC-dropout or deep ensemble methods produce wrong segmentation results with high uncertainty for the adversarial images, it is easier for the uncertainty-based methods to detect them.

Compared to feature space-based methods, the proposed method shows superior detection performance, mainly in the near OOD datasets. This indicates that instead of using only the spectral features or the latent space features, it is better to utilize the combined features which are extracted from the encoder blocks to detect OOD images because the encoder layers extract multi-scale and semantically rich representations of the input image. These representations capture the underlying patterns and main characteristics of the image that can be helpful in distinguishing in-distribution (ID) and OOD images.

Analyzing the results of the proposed method in the near OOD datasets, despite achieving the best results in terms of AUC and FPR, we noticed that the FPR results of M&Ms, random bias field, random gamma, and random noise datasets were relatively poorer than the other OOD datasets. Interestingly, these poor performances have some correlation with the high standard deviation of their Dice scores. This can tell us that the relatively high FPR of these datasets could be due to the presence of images that are very similar to ID images in which the model segments them well.

The qualitative analysis of segmentation results, uncertainties, and Mahalanobis distance supports the quantitative findings. The uncertainty metrics and Mahalanobis distance provide some insights into the quality of segmentation and the similarity of images to the ID images.

Finally, the proposed OOD detection method offers a straightforward implementation that can be easily integrated with any pre-trained segmentation model as a post-hoc, requiring no modifications to the model or its pre-trained weights. To deploy the OOD detection system, it is necessary to determine a threshold for the OOD score or distance. One approach is to select a distance threshold that achieves a 95% true positive rate (TPR) on the validation in-distribution (ID) dataset [Liang et al., 2017, González et al., 2021]. Alternatively, the threshold can be based on the mean and standard deviation of the distances or OOD scores calculated from the validation ID datasets [Karimi et al., 2023]. These thresholds serve as effective criteria for distinguishing between ID and OOD samples, facilitating the practical implementation of the OOD detection system.

6.6/ CONCLUSION

In this chapter, we propose a post-hoc method for detecting out-of-distribution images, which can be easily integrated with any pre-trained segmentation model. Our method utilizes the features extracted from the encoder blocks of the segmentation model and employs Mahalanobis distance as a metric to measure the distance between the encoder features of the input image and the in-distribution images to determine whether the input image is OOD or not. To evaluate the performance of our approach, we conducted experiments using a pre-trained segmentation model trained on a publicly available cardiac short-axis cine MRI dataset. We assessed the detection performance on 13 different out-of-distribution datasets, categorized as near, mild, and far based on their similarity to the in-distribution dataset. The results demonstrate that our proposed method outperforms state-of-the-art feature space-based and uncertainty-based out-of-distribution detection methods. It achieves the best detection results across all datasets, except for the adversarial dataset, in terms of both AUC and FPR at a TPR of 95%. Our method successfully detects out-of-distribution datasets, particularly the mild and far out-of-distribution

datasets, with an AUC of 1.0 and an FPR of 0.0. This highlights the advantage of utilizing the multi-scale and semantically rich representations of the encoder for out-of-distribution detection, as opposed to solely relying on the latent space features. Although the uncertainty-based method, specifically the Dice-within-samples approach, exhibits better detection performance for the adversarial dataset and shows a strong correlation with the segmentation quality in the near out-of-distribution datasets, it fails to detect mild and far out-of-distribution images, showing the weakness of these methods when the images are more dissimilar. Future work will focus on exploring the potential of combining both Mahalanobis distance and uncertainty scores to enhance the identification and detection of challenging out-of-distribution images that are difficult to segment.

III

CONCLUSION

7

Conclusion

This chapter provides a summary of the work presented in this thesis and explores the limitations of our work as well as the potential future directions for improvement and further development.

7.1/ Thesis Summary

In this thesis, we first focused on improving the segmentation of the more challenging scar segmentation from LGE MRI with the help of uncertainty information. Then to detect failed segmentation results before analyzing the segmentations in the downstream tasks, we proposed uncertainty-based quality control to determine the quality of the segmentation results and to reduce incorrect analysis in the downstream tasks. Finally, to identify and reject outlier images during inference, we proposed feature-space and uncertainty-based out-of-distribution detection for cardiac MR segmentation.

Deep learning-based segmentation of the heart structures from cardiac MR images has achieved state-of-the-art results, sometimes even matching the segmentation performance of experts. However, the segmentation of scar tissues, such as myocardial scar, remains challenging for deep learning methods due to their small size and lack of contrast with surrounding structures. Previous studies have utilized Bayesian neural networks to generate uncertainty estimates, with higher uncertainty indicating challenging image regions. These uncertain areas can provide insights into potential segmentation errors. To leverage this uncertainty information, we propose a segmentation model that integrates uncertainty into the learning process. To improve the segmentation of the challenging regions such as scars in cardiac MRI, we propose a segmentation model that integrates uncertainty information into the learning process. More specifically, Monte-Carlo dropout is employed to estimate uncertainty during training, and the uncertainty (mean of the pixel-wise sample variance) is then incorporated into the loss function to improve segmentation accuracy and probability calibration. The proposed method is evaluated on

two publicly available datasets: EMIDEC MICCAI 2020 and LAScarQS MICCAI 2022, which specifically target the segmentation of infarcted myocardium and left atrial scars from LGE MRI. The experimental results demonstrate that our method achieves state-of-the-art performance, surpassing the top-ranked approaches from both challenges. From the ablation study, we observe that the benefits of incorporating uncertainty information are most pronounced in apical slices and scar segmentation, which are visually difficult cases with higher epistemic uncertainty. This confirms that uncertainty provides useful guidance, especially for challenging examples.

Despite achieving high accuracy, deep learning models lack reliability for clinical adoption. Even top segmentation models generate anatomically implausible cardiac MRI segmentations, unlike human experts. Flawed segmentations can lead to erroneous clinical decisions in subsequent tasks. To address this issue, we propose an uncertainty-based quality control (QC) method to identify failed segmentations before further analysis. The proposed QC framework for T1 mapping and ECV analysis consists of three key components. Firstly, we employ a Bayesian Swin transformer-based U-Net to segment the left ventricular and right ventricular blood pools, as well as the left ventricular myocardium, from native and post-contrast T1 mapping images. Secondly, we introduce an automated QC method to detect poorly segmented images generated by the model. This QC method utilizes image-level uncertainty metrics derived from the Bayesian model, including metrics such as Dice agreement within Monte Carlo samples, Hausdorff distance agreement within MC samples, and the mean of pixel-wise uncertainty metrics like sample variance and predictive entropy. These image-level uncertainty features are fed into a random forest (RF) classifier, which is trained to classify the quality of the segmentation results. Experimental results using private and public datasets demonstrate that our proposed QC method significantly outperforms other state-of-the-art uncertainty-based QC methods, as evidenced by the mean area under the ROC curve. Notably, the improvements are particularly pronounced when dealing with challenging datasets or when the segmentation model's performance is suboptimal, showcasing the robustness of our method in detecting inaccurate segmentations. After rejecting the inaccurate segmentation results identified by the QC method, T1 mapping and ECV values are automatically computed, enabling the characterization of myocardial tissues in both healthy and pathological cardiac cases. The computed myocardial T1 and ECV values show excellent agreement with manual segmentations, as indicated by high Pearson correlation coefficients. These automatically computed values effectively capture the characteristics of myocardial tissues. Overall, our proposed fully automatic uncertainty-based QC framework for T1 mapping and ECV analysis has great potential to enhance the accuracy and reliability of cardiac MR segmentation, thereby improving the clinical decision-making process. The method's robustness in detecting failed segmentations and the strong agreement between automatic and manual segmentations underscores its value as a valuable tool for character-

izing myocardial tissues in healthy and pathological cardiac cases.

Real-world segmentation models encounter out-of-distribution (OOD) inputs deviating from training data. Such differences can arise from changes in scanners, protocols, or even modalities. Unseen OOD images cause unpredictable model behavior, threatening clinical safety. Predictive uncertainty information can be valuable in detecting segmentation errors when models perform poorly on in-distribution test set images. However, the uncertainty information may not be effective when samples deviate significantly from the training data distribution. To address this challenge and enhance the trustworthiness of the models by detecting and rejecting OOD images that differ greatly from the in-distribution images, we propose a post-hoc out-of-distribution (OOD) detection method. Our method can be applied to any pre-trained segmentation model without requiring modifications to the model or its pre-trained weights. It utilizes the features extracted from the encoder blocks of the segmentation model and employs the Mahalanobis distance as a metric to measure the distance between the encoder features of the input image and the in-distribution images. This distance measurement helps determine whether the input image is OOD or not. To evaluate the performance of our approach, we conducted experiments using a pre-trained segmentation model that was trained on a publicly available cardiac short-axis cine MRI dataset. We assessed the detection performance on 13 different OOD datasets, categorized as near, mild, and far based on their similarity to the in-distribution dataset. The results demonstrate that our method outperforms state-of-the-art feature space-based and uncertainty-based OOD detection methods across the various OOD datasets. Our method successfully detects near, mild, and far OOD images with high detection accuracy, showcasing the advantage of leveraging the multi-scale and semantically rich representations of the encoder. While the uncertainty-based method, specifically the Dice-within-samples approach, exhibits better detection performance for the adversarial dataset and shows a strong correlation with segmentation quality in the near OOD datasets, it fails to detect mild and far OOD images. This highlights the limitations of these methods when dealing with images that are more dissimilar from the training distribution.

7.2/ PERSPECTIVES

In this section, we discuss some limitations of the work presented in this thesis and provide comprehensive discussions and suggestions for future research directions that build upon the main contributions.

Improving segmentation performance through refined image-level uncertainty features: In Chapter 4, we explored the use of uncertainty information to enhance cardiac MR segmentation performance. Our approach utilized an image-level uncertainty mea-

sure obtained by averaging pixel-wise sample variances. However, this metric may not be optimal when dealing with imbalanced class distributions, where the background class significantly outweighs the objects of interest. Averaging pixel-wise uncertainty may not accurately represent underlying segmentation uncertainty.

To address this limitation, a potential future direction is to employ image-level uncertainty metrics less affected by class imbalance. One such metric is the Dice coefficient-based uncertainty metric [Roy et al., 2018, Ng et al., 2020, Arega et al., 2021a], Dice within samples. By incorporating the Dice within samples uncertainty-based loss alongside segmentation loss, the approach can achieve more refined segmentation performance in cases where the background is significantly larger than the objects of interest. Dice within samples focuses solely on foreground classes, disregarding the background (true negative) class, making it a more effective measure in these scenarios.

Uncovering the underlying causes of segmentation failure: In the pursuit of improving the reliability of the segmentation model, a quality control method was proposed in Chapter 5 to identify and reject incorrect segmentation results. However, in addition to identifying these failures, it could be valuable to understand the underlying causes or sources behind them. By gaining insights into why certain segmentation results are rejected, we can further enhance the model's reliability.

Explaining the cause of segmentation failure can involve investigating whether the incorrect results are due to the low quality of the input image or due to the poor generalization capability of the segmentation model. For the latter case, techniques can be devised to improve the model's generalization and reduce the occurrence of incorrect segmentations. This may involve exploring model performance-enhancing methods [Chen et al., 2020a, Garcea et al., 2022] tailored to address the model generalization problem.

Regarding low-quality input images, it could be worth considering image quality enhancement techniques [Tsai et al., 2013, Zhou et al., 2019, Bing et al., 2019] to improve the overall quality of the images. These techniques could include noise or artifact reduction, contrast or brightness enhancement, or deep learning-based super-resolution approaches [Chen et al., 2020b]. By enhancing the image quality, the number of flawed segmentation results can be reduced, thereby improving the overall reliability and accuracy of the segmentation model.

Correcting segmentation errors before or after the QC method: The proposed quality control (QC) method in Chapter 5 primarily focuses on detecting erroneous segmentation results and excluding them from downstream tasks. However, it could be interesting also to correct some of the bad segmentation results with minor errors using different deep learning-based post-processing methods. For instance, techniques such as those presented in [Painchaud et al., 2019] and [Larrazabal et al., 2020] could be employed to rec-

tify and refine the detected segmentation errors. As proposed in [Painchaud et al., 2019] and [Larrazabal et al., 2020], one approach can involve leveraging constrained Variational Autoencoders (cVAE) or Denoising Autoencoders (DAE) to ensure anatomical validity in the segmentation results. These methods employ a warping step that guides the segmentation predictions toward the closest anatomically valid cardiac shape.

Such correction procedures could be incorporated either before the QC method to reduce the number of anatomical errors in the segmentation results or after the QC to correct some of the rejected segmentation results that have minor errors. This expanded approach can allow for the inclusion of more patient cases (which were previously rejected due to minor errors) in the downstream tasks, providing additional valuable information about various pathologies and ultimately enhancing the overall clinical decision-making process.

Considerations for T1 and ECV values comparison: In Chapter 5, we examined the T1 and ECV values associated with various myocardial pathologies. These values were obtained from T1 mapping images acquired using one T1-mapping imaging technique (MOLLI). However, it is crucial to consider that differences in MR field strength, MRI scanners, imaging techniques, or even versions of T1-mapping sequences can impact T1 estimations [Puyol-Antón et al., 2020]. Therefore, it is essential to exercise caution when directly translating the T1 and ECV values from this study to images acquired using different vendor types or T1-mapping acquisition techniques.

Multi-Modal myocardial pathology analysis through combined LGE and T1 Mapping: Another interesting study regarding myocardial pathology characterization could be to leverage complementary modalities like LGE imaging alongside T1 mapping [Puyol-Antón et al., 2020]. LGE visualizes fibrotic scar while T1 mapping quantifies diffuse fibrosis. Combining modalities may offer deeper tissue insights and provide a more robust pathology assessment. Future studies exploring the correlation and complementarity between LGE and T1 mapping measurements can contribute valuable insights to the field of cardiac imaging and myocardial pathology assessment. However, it is worth noting that registering or aligning these modalities could be challenging.

Combining uncertainty-based and feature space-based information for better OOD image detection: For the OOD detection method presented in Chapter 6, as a future work, it could be interesting to see the combination of feature space-based and uncertainty-based OOD detection methods to distinguish ID from OOD images better. The results from our experiments in Chapter 6 also suggest this approach, as the uncertainty-based method performs well in detecting adversarial OOD images, while feature space-based methods struggle with them. In addition, the uncertainty-based method is good at identifying poorly segmented images (quality control) and provides useful uncertainty estimates for the near ODD images, which can be utilized alongside the distance

information from the feature space-based methods to provide complementary information for the OOD detection system.

Model interpretability and explainability: While the methods proposed in this thesis focus primarily on improving model performance and reliability, an equally important consideration is model interpretability and explainability [Singh et al., 2020]. Ensuring that clinical users can understand and trust model outputs is critical for real-world adoption. Therefore, a valuable research direction lies in enhancing the interpretability and explainability of our approaches.

By incorporating state-of-the-art explainable AI techniques, such as those discussed in [Arrieta et al., 2020, Singh et al., 2020], into our methods, we can enhance the transparency and trustworthiness of the models. These techniques aim to uncover the internal mechanisms of deep learning models and provide insights into the features and patterns that contribute to their decisions. This integration of explainable AI techniques could enhance the interpretability of the models and facilitate their integration into clinical workflows, making them more usable in real-world scenarios.

BIBLIOGRAPHY

[Abbas et al., 2015] Abbas, A., Matthews, G. H., Brown, I. W., Shambrook, J., Peebles, C., et Harden, S. (2015). **Cardiac MR assessment of microvascular obstruction.** *The British journal of radiology*, 88 1047:20140470.

[Abdelhamed et al., 2023] Abdelhamed, M. K., et Meriaudeau, F. (2023). **NesT UNet: pure transformer segmentation network with an application for automatic cardiac myocardial infarction evaluation**. In *Medical Imaging 2023: Computer-Aided Diagnosis*, volume 12465, pages 608–619. SPIE.

[AD Elster, 2023] AD Elster, E. L. (2023). **True fisp**. https://mriquestions.com/true-fispfiesta.html.

[Ali et al., 2021] Ali, N., Behairy, N., Kharabish, A., Elmozy, W., Hegab, A., et Saraya, S. (2021). **Cardiac MRI T1 mapping and extracellular volume application in hypertrophic cardiomyopathy**. *Egyptian Journal of Radiology and Nuclear Medicine*, 52:1–9.

[Alzubaidi et al., 2021a] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., et Farhan, L. (2021a). **Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions**. *Journal of big Data*, 8:1–74.

[Alzubaidi et al., 2021b] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A. Q., Duan, Y., Al-Shamma, O., Santamaría, J. I., Fadhel, M. A., Al-Amidie, M., et Farhan, L. (2021b). **Review of deep learning: concepts, CNN architectures, challenges, applications, future directions**. *Journal of Big Data*, 8.

[Amado et al., 2004] Amado, L. C., Gerber, B. L., Gupta, S. N., Rettmann, D. W., Szarf, G., Schock, R. B., Nasir, K., Kraitchman, D. L., et Lima, J. A. C. (2004). **Accurate and objective infarct sizing by contrast-enhanced magnetic resonance imaging in a canine myocardial infarction model.** *Journal of the American College of Cardiology*, 44 12:2383–9.

[Amano et al., 2018] Amano, Y., Kitamura, M., Takano, H., Yanagisawa, F., Tachi, M., Suzuki, Y., Kumita, S., et Takayama, M. (2018). **Cardiac MR imaging of hypertrophic cardiomyopathy: Techniques, findings, and clinical relevance**. *Magnetic Resonance in Medical Sciences*, 17:120 – 131.

[An et al., 2015] An, J., et Cho, S. (2015). **Variational autoencoder based anomaly detection using reconstruction probability**.

[Araújo et al., 2019] Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., Carneiro, Â., Mendonça, A. M., et Campilho, A. J. C. (2019). **DR—GRADUATE: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images**. *Medical image analysis*, 63:101715.

[Arega et al., 2020] Arega, T. W., et Bricq, S. (2020). **Automatic Myocardial Scar Segmentation from Multi-sequence Cardiac MRI Using Fully Convolutional Densenet with Inception and Squeeze-Excitation Module**. In *Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge*, pages 102–117. Springer.

[Arega et al., 2023a] Arega, T. W., Bricq, S., Legrand, F., Jacquier, A., Lalande, A., et Mériaudeau, F. (2023a). **Automatic uncertainty-based quality controlled T1 mapping and ECV analysis from native and post-contrast cardiac T1 mapping images using bayesian vision transformer.** *Medical image analysis*, 86:102773.

[Arega et al., 2023b] Arega, T. W., Bricq, S., Legrand, F., Jacquier, A., Lalande, A., et Meriaudeau, F. (2023b). **Uncertainty-based quality controlled T1 mapping and ECV analysis using bayesian vision transformer**. In *Medical Imaging with Deep Learning, short paper track*.

[Arega et al., 2021a] Arega, T. W., Bricq, S., et Mériaudeau, F. (2021a). **Leveraging uncertainty estimates to improve segmentation performance in cardiac mr**. In *UNSURE/PIPPI@MICCAI*.

[Arega et al., 2022a] Arega, T. W., Bricq, S., et Mériaudeau, F. (2022a). **Automatic quality assessment of cardiac MR images with motion artefacts using multi-task learning and k-space motion artefact augmentation**. In *STACOM@MICCAI*.

[Arega et al., 2022b] Arega, T. W., Bricq, S., et Mériaudeau, F. (2022b). **Using polynomial loss and uncertainty information for robust left atrial and scar quantification and segmentation**. In *LAScarQS@MICCAI*.

[Arega et al., 2021b] Arega, T. W., Grand, F. L., Bricq, S., et Mériaudeau, F. (2021b). **Using MRI-specific data augmentation to enhance the segmentation of right ventricle in multi-disease, multi-center and multi-view cardiac MRI**. In *STACOM@MICCAI*.

[Arrieta et al., 2020] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et others (2020). **Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI**. *Information fusion*, 58:82–115.

[Ayhan et al., 2019] Ayhan, M. S., Kuehlewein, L., Aliyeva, G., Inhoffen, W., Ziemssen, F., et Berens, P. (2019). **Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection**. *Medical image analysis*, 64:101724.

[Bai et al., 2017] Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A. M., Aung, N. L., Lukaschuk, E., Sanghvi, M. M., Zemrak, F., Fung, K., Paiva, J. M., Carapella, V., Kim, Y. J., Suzuki, H., Kainz, B., Matthews, P. M., Petersen, S. E., Piechnik, S. K., Neubauer, S., Glocker, B., et Rueckert, D. (2017). **Automated cardiovascular magnetic resonance image analysis with fully convolutional networks**. *Journal of Cardiovascular Magnetic Resonance*, 20.

[Bai et al., 2018] Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A. M., Aung, N. L., Lukaschuk, E., Sanghvi, M. M., Zemrak, F., Fung, K., Paiva, J. M., Carapella, V., Kim, Y. J., Suzuki, H., Kainz, B., Matthews, P. M., Petersen, S. E., Piechnik, S. K., Neubauer, S., Glocker, B., et Rueckert, D. (2018). **Automated cardiovascular magnetic resonance image analysis with fully convolutional networks**. *Journal of Cardiovascular Magnetic Resonance*, 20.

[Baron et al., 2013] Baron, N., Kachenoura, N., Cluzel, P., Frouin, F., Herment, A., Grenier, P. A., Montalescot, G., et Beygui, F. (2013). **Comparison of various methods for quantitative evaluation of myocardial infarct volume from magnetic resonance delayed enhancement data.** *International journal of cardiology*, 167 3:739–44.

[Baumgartner et al., 2017] Baumgartner, C. F., Koch, L. M., Pollefeys, M., et Konukoglu, E. (2017). **An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation**. In *STACOM@MICCAI*.

[Baumgartner et al., 2019] Baumgartner, C. F., Tezcan, K. C., Chaitanya, K., Hötker, A. M., Muehlematter, U. J., Schawkat, K., Becker, A. S., Donati, O. F., et Konukoglu, E. (2019). **Phiseg: Capturing uncertainty in medical image segmentation**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

[Berger et al., 2021] Berger, C., Paschali, M., Glocker, B., et Kamnitsas, K. (2021). **Confidence-based out-of-distribution detection: A comparative study and analysis**. In *UNSURE/PIPPI@MICCAI*.

[Bernard et al., 2018] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M. A. G., Sanromá, G., Napel, S., Petersen, S. E., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V. A., Krishnamurthi, G., Rohé, M.-M., Pennec, X., Sermesant, M., Isensee, F., Jäger, P. F., Maier-Hein, K., Full, P. M., Wolf, I., Engelhardt, S., Baumgartner, C. F., Koch, L. M., Wolterink, J. M., Ivsgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., et Jodoin, P.-M. (2018).

Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37:2514–2525.

[Bing et al., 2019] Bing, X., Zhang, W., Zheng, L., et Zhang, Y. (2019). **Medical image super resolution using improved generative adversarial networks**. *IEEE Access*, 7:145030–145038.

[Blundell et al., 2015] Blundell, C., Cornebise, J., Kavukcuoglu, K., et Wierstra, D. (2015). **Weight uncertainty in neural networks**. *ArXiv*, abs/1505.05424.

[Boser et al., 1992] Boser, B. E., Guyon, I. M., et Vapnik, V. N. (1992). **A training algorithm for optimal margin classifiers**. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

[Brahim et al., 2022] Brahim, K., Arega, T. W., Boucher, A., Bricq, S., Sakly, A., et Mériaudeau, F. (2022). **An improved 3D deep learning-based segmentation of left ventricular myocardial diseases from delayed-enhancement MRI with inclusion and classification prior information u-net (ICPIU-Net)**. *Sensors (Basel, Switzerland)*, 22.

[Brahim et al., 2020] Brahim, K., Qayyum, A., Lalande, A., Boucher, A., Sakly, A., et Mériaudeau, F. (2020). **Efficient 3D deep learning for myocardial diseases segmentation**. In *M&Ms and EMIDEC/STACOM@MICCAI*.

[Bshyamanth, 2023] Bshyamanth, jaintarun, h. (2023). **Decision tree**. https://www.geeksforgeeks.org/decision-tree/.

[Bushaev, 2017] Bushaev, V. (2017). **How do we 'train' neural networks ?** https://towardsdatascience.com/how-do-we-train-neural-networks-edd985562b73.

[Caiazzo et al., 2020] Caiazzo, G., Musci, R. L., Frediani, L., Umińska, J. M., Wańha, W., Filipiak, K. J., Kubica, J., et Navarese, E. P. (2020). **State of the art: No-reflow phenomenon.** *Cardiology clinics*, 38 4:563–573.

[Callie Tayrien, 2023] Callie Tayrien, Stacey Wojcik, S. K. (2023). **Anatomy and function of the heart valves**. https://www.urmc.rochester.edu/encyclopedia/content.aspx?ContentTypeID=90&ContentID=P03059.

[Campello et al., 2021] Campello, V. M., Gkontra, P., Izquierdo, C., Martín-Isla, C., Sojoudi, A., Full, P. M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., Parreño, M., Albiol, A., Kong, F., Shadden, S. C., Acero, J. C., Sundaresan, V., Saber, M., Elattar, M., Li, H., Menze, B., Khader, F., Haarburger, C., Scannell, C. M., Veta, M., Carscadden, A., Punithakumar, K., Liu, X., Tsaftaris, S. A., Huang, X., Yang, X., Li, L., Zhuang, X.,

Viladés, D., Descalzo, M. L., Guala, A., Mura, L. L., Friedrich, M. G., Garg, R., Lebel, J., Henriques, F., Karakas, M., Çavuş, E., Petersen, S. E., Escalera, S., Seguí, S., Rodríguez-Palomares, J. F., et Lekadir, K. (2021). **Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge**. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554.

[Cao et al., 2021] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et Wang, M. (2021). **Swin-unet: Unet-like pure transformer for medical image segmentation**. *ArXiv*, abs/2105.05537.

[Cao et al., 2020] Cao, X., Chen, H., Li, Y., Peng, Y., Wang, S., et Cheng, L.-J. (2020). **Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation**. *IEEE Transactions on Medical Imaging*, 40:431–443.

[Chen et al., 2020a] Chen, C., Bai, W., Davies, R. H., Bhuva, A. N., Manisty, C. H., Augusto, J. B., Moon, J. C., Aung, N., Lee, A. M., Sanghvi, M. M., et others (2020a). **Improving the generalizability of convolutional neural network-based segmentation on CMR images**. *Frontiers in cardiovascular medicine*, 7:105.

[Chen et al., 2019a] Chen, C., Biffi, C., Tarroni, G., Petersen, S. E., Bai, W., et Rueckert, D. (2019a). **Learning shape priors for robust cardiac MR segmentation from multi-view images**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

[Chen et al., 2019b] Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., et Rueckert, D. (2019b). **Deep learning for cardiac image segmentation: A review**. *Frontiers in Cardiovascular Medicine*, 7.

[Chen et al., 2021a] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., et Zhou, Y. (2021a). **Transunet: Transformers make strong encoders for medical image segmentation**. *ArXiv*, abs/2102.04306.

[Chen et al., 2021b] Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., et Miao, Y. (2021b). **Review of image classification algorithms based on convolutional neural networks**. *Remote. Sens.*, 13:4712.

[Chen et al., 2020b] Chen, L., Yang, X., Jeon, G., Anisetti, M., et Liu, K. (2020b). **A trusted medical image super-resolution method based on feedback adaptive weighted dense network**. *Artificial Intelligence in Medicine*, 106:101857.

[Chen et al., 2019c] Chen, M., Fang, L., et Liu, H. (2019c). **Fr-net: Focal loss constrained deep residual networks for segmentation of cardiac MRI**. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 764–767.

[Chen et al., 2020c] Chen, X., Men, K., Chen, B., Tang, Y., Zhang, T., Wang, S., Li, Y., et Dai, J. (2020c). **CNN -based quality assurance for automatic segmentation of breast cancer in radiotherapy**. *Frontiers in Oncology*, 10.

[Cleve et al., 2018] Cleve, J., et McCulloch, M. L. (2018). **Conducting a cardiac ultrasound examination**.

[Clinic, 2021] Clinic, C. (2021). **Heart conduction system (cardiac conduction)**. https://my.clevelandclinic.org/health/body/21648-heart-conduction-system.

[Clough et al., 2019] Clough, J. R., Öksüz, I., Byrne, N., Schnabel, J. A., et King, A. P. (2019). **Explicit topological priors for deep-learning based image segmentation using persistent homology**. In *Information Processing in Medical Imaging*.

[Cortes et al., 1995] Cortes, C., et Vapnik, V. (1995). **Support-vector networks**. *Machine learning*, 20:273–297.

[Cui et al., 2018] Cui, C., Wang, S., Lu, M., Duan, X., Wang, H., Jia, L., Tang, Y., Sirajuddin, A., Prasad, S. K., Kellman, P., Arai, A. E., et Zhao, S. (2018). **Detection of recent myocardial infarction using native T1 mapping in a swine model: A validation study**. *Scientific Reports*, 8.

[de la Rosa et al., 2019] de la Rosa, E., Sidibé, D., Decourselle, T., Leclercq, T., Cochet, A., et Lalande, A. (2019). **Myocardial infarction quantification from late gadolinium enhancement MRI using top-hat transforms and neural networks**. *ArXiv*, abs/1901.02911.

[Denouden et al., 2018] Denouden, T., Salay, R., Czarnecki, K., Abdelzad, V., Phan, B., et Vernekar, S. (2018). **Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance**. *ArXiv*, abs/1812.02765.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., et Toutanova, K. (2019). **Bert: Pre-training of deep bidirectional transformers for language understanding**. *ArXiv*, abs/1810.04805.

[Devries et al., 2018a] Devries, T., et Taylor, G. W. (2018a). **Learning confidence for out-of-distribution detection in neural networks**. *ArXiv*, abs/1802.04865.

[Devries et al., 2018b] Devries, T., et Taylor, G. W. (2018b). **Leveraging uncertainty estimates for predicting segmentation quality**. *ArXiv*, abs/1807.00502.

[Dhua, 2020] Dhua, D. (2020). **Cardiomyopathy**. https://www.drdebarghadhua.com/wp-content/themes/Doctor/documents/Cardiomyopathy.html.

[Dice, 1945] Dice, L. R. (1945). **Measures of the amount of ecologic association between species**. *Ecology*, 26(3):297–302.

[Dimitriadis et al., 2018] Dimitriadis, S. I., Liparas, D., Initiative, A. D. N., et others (2018). **How random is the random forest? random forest algorithm on the service of structural imaging biomarkers for alzheimer's disease: from alzheimer's disease neuroimaging initiative (adni) database**. *Neural regeneration research*, 13(6):962.

[Dolezal et al., 2022] Dolezal, J. M., Srisuwananukorn, A., Karpeyev, D. A., Ramesh, S., Kochanny, S. E., Cody, B., Mansfield, A. S., Rakshit, S., Bansal, R., Bois, M. C., Bungum, A. O., Schulte, J. J., Vokes, E. E., Garassino, M. C., Husain, A. N., et Pearson, A. T. (2022). **Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology**. *Nature Communications*, 13.

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et others (2020). **An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020**. *arXiv preprint arXiv:2010.11929*.

[Drozdzal et al., 2016] Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., et Pal, C. J. (2016). **The importance of skip connections in biomedical image segmentation**. *ArXiv*, abs/1608.04117.

[Ed Burns, 2023] Ed Burns, N. L. (2023). **artificial intelligence (ai)**. https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence.

[Fahmy et al., 2019a] Fahmy, A. S., El-Rewaidy, H., Nezafat, M., Nakamori, S., et Nezafat, R. V. (2019a). **Automated analysis of cardiovascular magnetic resonance myocardial native T1 mapping images using fully convolutional neural networks**. *Journal of Cardiovascular Magnetic Resonance*, 21.

[Fahmy et al., 2019b] Fahmy, A. S., Neisius, U., Chan, R. H., Rowin, E. J., Manning, W. J., Maron, M. S., et Nezafat, R. V. (2019b). **Three-dimensional deep convolutional neural networks for automated myocardial scar quantification in hypertrophic cardiomyopathy: A multicenter multivendor study.** *Radiology*, page 190737.

[Feng et al., 2022] Feng, M., Xu, K., Wu, N., Huang, W., Bai, Y., Wang, C., et Wang, H. (2022). **Trusted multi-scale classification framework for whole slide image**. *ArXiv*, abs/2207.05290.

[Feng et al., 2020] Feng, X., Kramer, C., Salerno, M., et Meyer, C. H. (2020). **Automatic Scar Segmentation from DE-MRI Using 2D Dilated UNet with Rotation-Based Augmentation**. In *M&Ms and EMIDEC/STACOM@MICCAI*.

[Fent et al., 2017] Fent, G., Garg, P., Foley, J. R. J., Swoboda, P. P., Dobson, L. E., Erhayiem, B., Treibel, T. A., Moon, J. C., Greenwood, J. P., et Plein, S. (2017). **Synthetic myocardial extracellular volume fraction.** *JACC. Cardiovascular imaging*, 10 11:1402–1404.

[Filos et al., 2019] Filos, A., Farquhar, S., Gomez, A. N., Rudner, T. G. J., Kenton, Z., Smith, L., Alizadeh, M., de Kroon, A. A. W. M., et Gal, Y. (2019). **A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks**. *ArXiv*, abs/1912.10481.

[Fortunato et al., 2017] Fortunato, M., Blundell, C., et Vinyals, O. (2017). **Bayesian recurrent neural networks**. *arXiv preprint arXiv:1704.02798*.

[Foundation, 2017] Foundation, H. (2017). **How the heart works**. https://www.heartfoundation.org.nz/your-heart/how-the-heart-works.

[Francone, 2014] Francone, M. (2014). **Role of cardiac magnetic resonance in the evaluation of dilated cardiomyopathy: Diagnostic contribution and prognostic significance**. *ISRN Radiology*, 2014.

[Full et al., 2020] Full, P. M., Isensee, F., Jäger, P. F., et Maier-Hein, K. H. (2020). **Studying robustness of semantic segmentation under domain shift in cardiac MRI**. *ArXiv*, abs/2011.07592.

[Gal, 2016] Gal, Y. (2016). **Uncertainty in deep learning**. https://www.cs.ox.ac.uk/people/yarin.gal/website/thesis/thesis.pdf.

[Gal et al., 2015] Gal, Y., et Ghahramani, Z. (2015). **Dropout as a bayesian approximation: Representing model uncertainty in deep learning**. *ArXiv*, abs/1506.02142.

[Gal et al., 2016] Gal, Y., et Ghahramani, Z. (2016). **Dropout as a Bayesian approximation: Representing model uncertainty in deep learning**. In *international conference on machine learning*, pages 1050–1059. PMLR.

[Galati et al., 2022] Galati, F., Ourselin, S., et Zuluaga, M. A. (2022). **From accuracy to reliability and robustness in cardiac magnetic resonance image segmentation: a review**. *Applied Sciences*, 12(8):3936.

[Gao et al., 2021] Gao, Y., Zhou, M., et Metaxas, D. N. (2021). **Utnet: A hybrid transformer architecture for medical image segmentation**. *ArXiv*, abs/2107.00781.

[Garcea et al., 2022] Garcea, F., Serra, A., Lamberti, F., et Morra, L. (2022). **Data augmentation for medical imaging: A systematic literature review**. *Computers in Biology and Medicine*, page 106391.

[Garg et al., 2018] Garg, S., et Awate, S. P. (2018). **Perfect mcmc sampling in bayesian mrfs for uncertainty estimation in segmentation**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

[Gerig et al., 2001] Gerig, G., Jomier, M., et Chakos, M. (2001). **Valmet: A new validation tool for assessing and improving 3D object segmentation**. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2001: 4th International Conference Utrecht, The Netherlands, October 14–17, 2001 Proceedings 4*, pages 516–523. Springer.

[Ghadri et al., 2018] Ghadri, J. R., Wittstein, I. S., Prasad, A., Sharkey, S. W., Dote, K., Akashi, Y. J., Cammann, V. L., Crea, F., Galiuto, L., Desmet, W., Yoshida, T., Manfredini, R., Eitel, I., Kosuge, M., Nef, H. M., Deshmukh, A. J., Lerman, A., Bossone, E., Citro, R., Ueyama, T., Corrado, D., Kurisu, S., Ruschitzka, F., Winchester, D. E., Lyon, A. R., Omerovic, E., Bax, J. J., Meimoun, P., Tarantini, G., Rihal, C. S., Y-Hassan, S., Migliore, F., Horowitz, J. D., Shimokawa, H., Lüscher, T. F., et Templin, C. (2018). **International expert consensus document on takotsubo syndrome (part i): Clinical characteristics, diagnostic criteria, and pathophysiology**. *European Heart Journal*, 39:2032 – 2046.

[Gilotra, 2023] Gilotra, N. A. (2023). **Myocarditis**. https://www.hopkinsmedicine.org/health/conditions-and-diseases/myocarditis.

[Ginat et al., 2011] Ginat, D. T., Fong, M. W., Tuttle, D. J., Hobbs, S. K., et Vyas, R. C. (2011). **Cardiac imaging: Part 1, MR pulse sequences, imaging planes, and basic anatomy.** *AJR. American journal of roentgenology*, 197 4:808–15.

[Girum et al., 2020] Girum, K. B., Skandarani, Y., Hussain, R., Grayeli, A. B., Créhange, G., et Lalande, A. (2020). **Automatic Myocardial Infarction Evaluation from Delayed-Enhancement Cardiac MRI using Deep Convolutional Networks**. In *M&Ms and EMIDEC/STACOM@MICCAI*.

[González, 2023] González, C. (2023). **Lifelong learning in the clinical open world**. https://api.semanticscholar.org/CorpusID:259149923.

[González et al., 2021] González, C., Gotkowski, K., Bucher, A. M., Fischbach, R., Kaltenborn, I., et Mukhopadhyay, A. (2021). **Detecting when pre-trained nnU-Net models fail silently for covid-19 lung lesion segmentation**. *ArXiv*, abs/2107.05975.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., et Courville, A. (2016). **Deep learning**. MIT press.

[Graves, 2011] Graves, A. (2011). **Practical variational inference for neural networks**. In *NIPS*.

[Gu et al., 2018] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et others (2018). **Recent advances in convolutional neural networks**. *Pattern recognition*, 77:354–377.

[Guo et al., 2017] Guo, C., Pleiss, G., Sun, Y., et Weinberger, K. Q. (2017). **On calibration of modern neural networks**. In *International Conference on Machine Learning*.

[Guo et al., 2022] Guo, F., Ng, M., Kuling, G., et Wright, G. (2022). **Cardiac MRI segmentation with sparse annotations: Ensembling deep learning uncertainty and shape priors**. *Medical image analysis*, 81:102532.

[Guo et al., 2018] Guo, J., Liu, G., Zuo, Y., et Wu, J. (2018). **An anomaly detection framework based on autoencoder and nearest neighbor**. *2018 15th International Conference on Service Systems and Service Management (ICSSSM)*, pages 1–6.

[Guo, 2017] Guo, L. (2017). **Interventricular septum**. https://www.osmosis.org/answers/interventricular-septum.

[Haaf et al., 2016] Haaf, P., Garg, P., Messroghli, D. R., Broadbent, D. A., Greenwood, J. P., et Plein, S. (2016). **Cardiac T1 mapping and extracellular volume (ECV) in clinical practice: a comprehensive review**. *Journal of Cardiovascular Magnetic Resonance*, 18.

[Han et al., 2020] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., et Tao, D. (2020). **A survey on visual transformer**. *ArXiv*, abs/2012.12556.

[Hann et al., 2021] Hann, E., Popescu, I. A., Zhang, Q., Gonzales, R. A., Barutcu, A., Neubauer, S., Ferreira, V. M., et Piechnik, S. K. (2021). **Deep neural network ensemble for on-the-fly quality control-driven segmentation of cardiac MRI T1 mapping**. *Medical Image Analysis*, 71.

[Harvard, 2022] Harvard, P. H. S. (2022). **Preventing heart disease**. https://www.hsph.harvard.edu/nutritionsource/disease-prevention/cardiovascular-disease/preventing-cvd/.

[He et al., 2022] He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., et Shen, D. (2022). **Transformers in medical image analysis: A review**. *ArXiv*, abs/2202.12165.

[He et al., 2016] He, K., Zhang, X., Ren, S., et Sun, J. (2016). **Deep residual learning for image recognition**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Hendrycks et al., 2016] Hendrycks, D., et Gimpel, K. (2016). **A baseline for detecting misclassified and out-of-distribution examples in neural networks**. *ArXiv*, abs/1610.02136.

[Hendrycks et al., 2018] Hendrycks, D., Mazeika, M., et Dietterich, T. G. (2018). **Deep anomaly detection with outlier exposure**. *ArXiv*, abs/1812.04606.

[Hennemuth et al., 2012] Hennemuth, A., Friman, O., Hüllebrand, M., et Peitgen, H.-O. (2012). **Mixture-model-based segmentation of myocardial delayed enhancement MRI**. In *International Workshop on Statistical Atlases and Computational Models of the Heart*.

[Herzog et al., 2020] Herzog, L., Murina, E., Dürr, O., Wegener, S., et Sick, B. (2020). **Integrating uncertainty in deep neural networks for MRI based stroke analysis**. *Medical image analysis*, 65:101790.

[Hoffmann et al., 2021] Hoffmann, L., et Elster, C. (2021). **Deep ensembles from a bayesian perspective**. *ArXiv*, abs/2105.13283.

[Hosny et al., 2018] Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., et Aerts, H. J. (2018). **Artificial intelligence in radiology**. *Nature Reviews Cancer*, 18:500 – 510.

[Hu et al., 2017] Hu, J., Shen, L., Albanie, S., Sun, G., et Wu, E. (2017). **Squeeze-and-excitation networks**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:2011–2023.

[Huang et al., 2016] Huang, C., Wu, Q., et Meng, F. (2016). **Qualitynet: Segmentation quality evaluation with deep convolutional networks**. *2016 Visual Communications and Image Processing (VCIP)*, pages 1–4.

[Huang et al., 2017] Huang, G., Liu, Z., Van Der Maaten, L., et Weinberger, K. Q. (2017). **Densely connected convolutional networks**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

[Huang et al., 2020] Huang, H., Li, Z., Wang, L., Chen, S., Dong, B., et Zhou, X. (2020). **Feature space singularity for out-of-distribution detection**. *ArXiv*, abs/2011.14654.

[Hüllebrand et al., 2020] Hüllebrand, M., Ivantsits, M., Zhang, H., Kohlmann, P., Kuhnigk, J.-M., Kühne, T., Schönberg, S. O., et Hennemuth, A. (2020). **Comparison of a hybrid mixture model and a CNN for the segmentation of myocardial pathologies in delayed enhancement MRI**. In *M&Ms and EMIDEC/STACOM@MICCAI*.

[Huttenlocher et al., 1993] Huttenlocher, D. P., Klanderman, G. A., et Rucklidge, W. J. (1993). **Comparing images using the hausdorff distance**. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863.

[IBM, 2023] IBM (2023). **What is random forest?** https://www.ibm.com/topics/random-forest.

[Institute, 2019] Institute, C. (2019). **Making sense of an echocardiogram report - for gps!** https://cardiologyinstitute.co.nz/gp-info/2019/1/28/making-sense-of-an-echocardiogram-report.

[Isensee et al., 2017] Isensee, F., Jaeger, P. F., Full, P. M., Wolf, I., Engelhardt, S., et Maier-Hein, K. (2017). **Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features**. In *STACOM@MICCAI*.

[Isensee et al., 2021] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., et Maier-Hein, K. H. (2021). **nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation**. *Nature methods*, 18(2):203–211.

[Jain, 2018] Jain, S. (2018). **An overview of regularization techniques in deep learning**. https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/.

[Jakkula, 2006] Jakkula, V. (2006). **Tutorial on support vector machine (svm)**. *School of EECS, Washington State University*, 37(2.5):3.

[Jang et al., 2017] Jang, Y., Hong, Y., Ha, S., Kim, S., et Chang, H. J. (2017). **Automatic segmentation of lv and rv in cardiac MRI**. In *STACOM@MICCAI*.

[Janiesch et al., 2021] Janiesch, C., Zschech, P., et Heinrich, K. (2021). **Machine learning and deep learning**. *Electronic Markets*, 31:685–695.

[JavaTpoint, 2022] JavaTpoint (2022). **Single layer perceptron in tensorflow**. https://www.javatpoint.com/single-layer-perceptron-in-tensorflow.

[JavaTpoint, 2023] JavaTpoint (2023). **Support vector machine algorithm**. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[Jayawardana et al., 2021] Jayawardana, R., et Bandaranayake, T. S. (2021). **Analysis of optimizing neural networks and artificial intelligent models for guidance, control, and navigation systems**. *International Research Journal of Modernization in Engineering, Technology and Science*, 3(3):743–759.

[Jiang, 2019] Jiang, Z. (2019). **A novel crop weed recognition method based on transfer learning from vgg16 implemented by keras**. In *IOP Conference Series: Materials Science and Engineering*, volume 677, page 032073. IOP Publishing.

[Jungo et al., 2019] Jungo, A., et Reyes, M. (2019). **Assessing reliability and challenges of uncertainty estimations for medical image segmentation**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 48–56. Springer.

[Karim et al., 2016] Karim, R., Bhagirath, P., Claus, P., Housden, R. J., Chen, Z., Karimaghaloo, Z., Sohn, H.-M., Rodríguez, L. L., Vera, S., Albà, X., et others (2016). **Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late gadolinium enhancement MR images**. *Medical image analysis*, 30:95–107.

[Karim et al., 2018] Karim, R., Blake, L.-E., Inoue, J., Tao, Q., Jia, S., Housden, R. J., Bhagirath, P., Duval, J.-L., Varela, M., Behar, J. M., et others (2018). **Algorithms for left atrial wall segmentation and thickness–evaluation on an open-source CT and MRI image database**. *Medical image analysis*, 50:36–53.

[Karim et al., 2013] Karim, R., Housden, R. J., Balasubramaniam, M., Chen, Z., Perry, D., Uddin, A., Al-Beyatti, Y., Palkhi, E., Acheampong, P., Obom, S., et others (2013). **Evaluation of current algorithms for segmentation of scar tissue from late gadolinium enhancement cardiovascular magnetic resonance of the left atrium: an open-access grand challenge**. *Journal of Cardiovascular Magnetic Resonance*, 15(1):1–17.

[Karimi et al., 2023] Karimi, D., et Gholipour, A. (2023). **Improving calibration and out-of-distribution detection in deep models for medical image segmentation**. *IEEE Transactions on Artificial Intelligence*, 4:383–397.

[Karpathy, 2016] Karpathy, A. (2016). **Convolutional neural networks (CNN/convnets)**. *CS231n Convolutional Neural Networks for Visual Recognition*.

[Kate Meier et al., 2009] Kate Meier, C., et Oyama, M. A. (2009). **Chapter 41 - Myocardial Infarction**. In Silverstein, D. C., et Hopper, K., editors, *Small Animal Critical Care Medicine*, pages 174–176. W.B. Saunders, Saint Louis.

[Kendall et al., 2015] Kendall, A., Badrinarayanan, V., et Cipolla, R. (2015). **Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding**. *arXiv preprint arXiv:1511.02680*.

[Kendall et al., 2017a] Kendall, A., Badrinarayanan, V., et Cipolla, R. (2017a). **Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding**. *ArXiv*, abs/1511.02680.

[Kendall et al., 2017b] Kendall, A., et Gal, Y. (2017b). **What uncertainties do we need in bayesian deep learning for computer vision?** *Advances in neural information processing systems*, 30.

[Khan et al., 2023] Khan, A., Rauf, Z., Sohail, A., Rehman, A., Asif, H., Asif, A., et Farooq, U. (2023). **A survey of the vision transformers and its CNN-transformer based variants**. *ArXiv*, abs/2305.09880.

[Khened et al., 2018] Khened, M., Varghese, A., et Krishnamurthi, G. (2018). **Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers**. *Medical Image Analysis*, 51:21–45.

[Kim et al., 1999] Kim, R. J., Fieno, D. S., Parrish, T. B., Harris, K. E., Chen, E., Simonetti, O. P., Bundy, J. M., Finn, J. P., Klocke, F. J., et Judd, R. M. (1999). **Relationship of MRI delayed contrast enhancement to irreversible injury, infarct age, and contractile function.** *Circulation*, 100 19:1992–2002.

[Kohl et al., 2018] Kohl, S. A. A., Romera-Paredes, B., Meyer, C., Fauw, J. D., Ledsam, J. R., Maier-Hein, K., Eslami, S. M. A., Rezende, D. J., et Ronneberger, O. (2018). **A probabilistic u-net for segmentation of ambiguous images**. *ArXiv*, abs/1806.05034.

[Kohlberger et al., 2012] Kohlberger, T., Singh, V., Alvino, C., Bahlmann, C., et Grady, L. (2012). **Evaluating segmentation error without ground truth**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 528–536. Springer.

[Kotsiantis et al., 2006] Kotsiantis, S. B., Zaharakis, I. D., et Pintelas, P. E. (2006). **Machine learning: a review of classification and combining techniques**. *Artificial Intelligence Review*, 26:159–190.

[Kramer et al., 2020] Kramer, C. M., Barkhausen, J., Bucciarelli-Ducci, C., Flamm, S. D., Kim, R. J., et Nagel, E. (2020). **Standardized cardiovascular magnetic resonance imaging (cmr) protocols: 2020 update**. *Journal of Cardiovascular Magnetic Resonance*, 22.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., et Hinton, G. E. (2012). **Imagenet classification with deep convolutional neural networks**. *Advances in neural information processing systems*, 25.

[Kushibar et al., 2022] Kushibar, K., Campello, V. M., Moras, L. G., Linardos, A., Radeva, P., et Lekadir, K. (2022). **Layer ensembles: A single-pass uncertainty estimation in deep learning for segmentation**. *ArXiv*, abs/2203.08878.

[Lakshminarayanan et al., 2016] Lakshminarayanan, B., Pritzel, A., et Blundell, C. (2016). **Simple and scalable predictive uncertainty estimation using deep ensembles**. In *NIPS*.

[Lalande et al., 2020] Lalande, A., Chen, Z., Decourselle, T., Qayyum, A., Pommier, T., Lorgis, L., de la Rosa, E., Cochet, A., Cottin, Y., Ginhac, D., et others (2020). **Emidec: A database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac MRI**. *Data*, 5(4):89.

[Lalande et al., 2021] Lalande, A., Chen, Z., Pommier, T., Decourselle, T., Qayyum, A., Salomon, M., Ginhac, D., Skandarani, Y., Boucher, A., Brahim, K., de Bruijne, M., Camarasa, R., Correia, T., Feng, X., Girum, K. B., Hennemuth, A., Huellebrand, M., Hussain, R., Ivantsits, M., Ma, J., Meyer, C. H., Sharma, R., Shi, J., Tsekos, N. V., Varela, M., Wang, X., Yang, S., Zhang, H., Zhang, Y., Zhou, Y., Zhuang, X., Couturier, R., et Mériaudeau, F. (2021). **Deep learning methods for automatic evaluation of delayed enhancement-MRI. the results of the EMIDEC challenge**. *Medical image analysis*, 79:102428.

[Lambert et al., 2022a] Lambert, B., Forbes, F., Doyle, S., Tucholka, A., et Dojat, M. (2022a). **Improving uncertainty-based out-of-distribution detection for medical image segmentation**. *ArXiv*, abs/2211.05421.

[Lambert et al., 2022b] Lambert, B., Forbes, F., Tucholka, A., Doyle, S., Dehaene, H., et Dojat, M. (2022b). **Trustworthy clinical ai solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis**. *ArXiv*, abs/2210.03736.

[Landman et al., 2015] Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., et Klein, A. (2015). **Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge**. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12.

[Larrazabal et al., 2020] Larrazabal, A. J., Mart'inez, C., Glocker, B., et Ferrante, E. (2020). **Post-DAE: Anatomically plausible segmentation via post-processing with denoising autoencoders**. *IEEE Transactions on Medical Imaging*, 39:3813–3820.

[Leclerc et al., 2019] Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E. A. R., Jodoin, P.-M., Grenier, T., Lartizien, C., D'hooge, J., Løvstakken, L., et Bernard, O. (2019). **Deep learning for segmentation using an open large-scale dataset in 2D echocardiography**. *IEEE Transactions on Medical Imaging*, 38:2198–2210.

[LeCun et al., 2015] LeCun, Y., et others (2015). **Lenet-5, convolutional neural networks**. *URL: http://yann. lecun. com/exdb/lenet*, 20(5):14.

[Lee et al., 2011] Lee, J. J. S., Liu, S., Nacif, M. S., Ugander, M., Han, J., Kawel, N., Sibley, C. T., Kellman, P., Arai, A. E., et Bluemke, D. A. (2011). **Myocardial T1 and extracellular volume fraction mapping at 3 tesla**. *Journal of Cardiovascular Magnetic Resonance*, 13:75 – 75.

[Lee et al., 2017] Lee, K., Lee, H., Lee, K., et Shin, J. (2017). **Training confidence-calibrated classifiers for detecting out-of-distribution samples**. *ArXiv*, abs/1711.09325.

[Lee et al., 2018] Lee, K., Lee, K., Lee, H., et Shin, J. (2018). **A simple unified framework for detecting out-of-distribution samples and adversarial attacks**. *Advances in neural information processing systems*, 31.

[Leibig et al., 2016] Leibig, C., Allken, V., Ayhan, M. S., Berens, P., et Wahl, S. (2016). **Leveraging uncertainty information from deep neural networks for disease detection**. *Scientific Reports*, 7.

[Li et al., 2019] Li, C., Tong, Q., Liao, X., Si, W., Chen, S., Wang, Q., et Yuan, Z. (2019). **APCP-NET: Aggregated parallel cross-scale pyramid network for CMR segmentation**. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 784–788.

[Li et al., 2022a] Li, L., Wu, F., Wang, S., Luo, X., Martín-Isla, C., Zhai, S., Zhang, J., Liu7, Y., Zhang, Z., Ankenbrand, M. J., Jiang, H., Zhang, X., Wang, L., Arega, T. W., Altunok, E., Zhao, Z., Li, F., Ma, J., Yang, X., Puybareau, É., Oksuz, I., Bricq, S., Li, W., Punithakumar, K., Tsaftaris, S. A., Schreiber, L. M., Yang, M., Liu, G., quan Xia, Y., Wang, G., Escalera, S., et Zhuang, X. (2022a). **MyoPS: A benchmark of myocardial pathology segmentation combining three-sequence cardiac magnetic resonance images**. *Medical image analysis*, 87:102808.

[Li et al., 2021] Li, L., Zimmer, V. A. M., Schnabel, J. A., et Zhuang, X. (2021). **Atrialgeneral: Domain generalization for left atrial segmentation of multi-center LGE MRIs**. In *MICCAI*.

[Li et al., 2022b] Li, L., Zimmer, V. A. M., Schnabel, J. A., et Zhuang, X. (2022b). **AtriaJSQnet: A new framework for joint segmentation and quantification of left atrium and scars incorporating spatial and shape information**. *Medical image analysis*, 76:102303.

[Li et al., 2022c] Li, L., Zimmer, V. A. M., Schnabel, J. A., et Zhuang, X. (2022c). **Medical image analysis on left atrial LGE MRI for atrial fibrillation studies: A review**. *Medical image analysis*, 77:102360.

[Liang et al., 2017] Liang, S., Li, Y., et Srikant, R. (2017). **Enhancing the reliability of out-of-distribution image detection in neural networks**. *arXiv: Learning*.

[Linmans et al., 2022] Linmans, J., Elfwing, S., van der Laak, J., et Litjens, G. J. S. (2022). **Predictive uncertainty estimation for out-of-distribution detection in digital pathology**. *Medical image analysis*, 83:102655.

[Liu et al., 2022] Liu, T., Hou, S., Zhu, J., Zhao, Z., et Jiang, H. (2022). **UGformer for robust left atrium and scar segmentation across scanners**. In *Challenge on Left Atrial and Scar Quantification and Segmentation*, pages 36–48. Springer.

[Liu et al., 2021] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., et Guo, B. (2021). **Swin transformer: Hierarchical vision transformer using shifted windows**. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002.

[Lopez et al., 2022] Lopez, E. O., Ballard, B. D., et Jan, A. (2022). **Cardiovascular disease**. In *StatPearls [Internet]*. StatPearls Publishing.

[Ma, 2020a] Ma, J. (2020a). **Cascaded Framework for Automatic Evaluation of Myocardial Infarction from Delayed-Enhancement Cardiac MRI**. *arXiv preprint arXiv:2012.14556*.

[Ma, 2020b] Ma, J. (2020b). **Histogram matching augmentation for domain adaptation with application to multi-centre, multi-vendor and multi-disease cardiac image segmentation**. *ArXiv*, abs/2012.13871.

[Ma, 2021] Ma, J. (2021). **Cutting-edge 3D medical image segmentation methods in 2020: Are happy families all alike?** *arXiv preprint arXiv:2101.00232*.

[Ma et al., 2021] Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., et Martel, A. L. (2021). **Loss odyssey in medical image segmentation**. *Medical Image Analysis*, 71:102035.

[Manivannan, 2020] Manivannan, I. (2020). **A comparative study of uncertainty estimation methods in deep learning based classification models**.

[Mann et al., 1947] Mann, H. B., et Whitney, D. R. (1947). **On a test of whether one of two random variables is stochastically larger than the other**. *Annals of Mathematical Statistics*, 18:50–60.

[Martín-Isla et al., 2023] Martín-Isla, C., Campello, V. M., Izquierdo, C., Kushibar, K., Sendra-Balcells, C., Gkontra, P., Sojoudi, A., Fulton, M. J., Arega, T. W., Punithakumar, K., Li, L., Sun, X., Khalil, Y. A., Liu, D., Jabbar, S., Queirós, S., Galati, F., Mazher, M., Gao, Z., Beetz, M., Tautz, L., Galazis, C., Varela, M., Hullebrand, M., Grau, V., Zhuang, X., Puig, D., Zuluaga, M. A., Mohy-ud Din, H., Metaxas, D., Breeuwer, M., Geest, R. J. v. d., Noga, M., Bricq, S., Rentschler, M. E., Guala, A., Petersen, S. E., Escalera, S., Palomares, J. F. R., et Lekadir, K. (2023). **Deep learning segmentation of the right ventricle in cardiac MRI: The M&Ms challenge**. *IEEE Journal of Biomedical and Health Informatics*, pages 1–14.

[Mazher et al., 2022] Mazher, M., Qayyum, A., Abdel-Nasser, M., et Puig, D. (2022). **Automatic semi-supervised left atrial segmentation using deep-supervision 3DResUnet with pseudo labeling approach for LAScarQS 2022 challenge**. In *Challenge on Left Atrial and Scar Quantification and Segmentation*, pages 153–161. Springer.

[McClure et al., 2014] McClure, P., Khalifa, F., Soliman, A., El-Ghar, M. A., Gimel'farb, G. L., Elmagraby, A., et El-Baz, A. S. (2014). **A novel NMF guided level-set for DWI prostate segmentation**. *Journal of Computer Science & Systems Biology*, 7:209–216.

[Mehrtash et al., 2019] Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., et Kapur, T. (2019). **Confidence calibration and predictive uncertainty estimation for deep medical image segmentation**. *IEEE Transactions on Medical Imaging*, 39:3868–3878.

[Mehrtash et al., 2020] Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., et Kapur, T. (2020). **Confidence calibration and predictive uncertainty estimation for deep medical image segmentation**. *IEEE transactions on medical imaging*, 39(12):3868–3878.

[Messroghli et al., 2017] Messroghli, D. R., Moon, J. C., Ferreira, V. M., Grosse-Wortmann, L., He, T., Kellman, P., Mascherbauer, J., Nezafat, R. V., Salerno, M., Schelbert, E. B., Taylor, A. J., Thompson, R. B., Ugander, M., van Heeswijk, R. B., et Friedrich, M. G. (2017). **Clinical recommendations for cardiovascular magnetic resonance mapping of T1, T2, T2* and extracellular volume: A consensus statement by the society for cardiovascular magnetic resonance (SCMR) endorsed by the european association for cardiovascular imaging (EACVI)**. *Journal of Cardiovascular Magnetic Resonance*, 19.

[Milletarì et al., 2016] Milletarì, F., Navab, N., et Ahmadi, S.-A. (2016). **V-net: Fully convolutional neural networks for volumetric medical image segmentation**. *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571.

[Moccia et al., 2018] Moccia, S., Banali, R., Martini, C., Muscogiuri, G., Pontone, G., Pepi, M., et Caiani, E. G. (2018). **Development and testing of a deep learning-based strategy for scar segmentation on CMR-LGE images**. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 32:187–195.

[Molle et al., 2019] Molle, P. V., Verbelen, T., Boom, C. D., Vankeirsbilck, B., Vylder, J. D., Diricx, B., Kimpe, T., Simoens, P., et Dhoedt, B. (2019). **Quantifying uncertainty of deep neural networks in skin lesion classification**. In *UNSURE/CLIP@MICCAI*.

[Montesinos López et al., 2022] Montesinos López, O. A., Montesinos López, A., et Crossa, J. (2022). **Fundamentals of artificial neural networks and deep learning**. In *Multivariate statistical machine learning methods for genomic prediction*, pages 379–425. Springer.

[Moon et al., 2013] Moon, J. C., Messroghli, D. R., Kellman, P., Piechnik, S. K., Robson, M. D., Ugander, M., Gatehouse, P. D., Arai, A. E., Friedrich, M. G., Neubauer, S.,

Schulz-Menger, J., et Schelbert, E. B. (2013). **Myocardial T1 mapping and extracellular volume quantification: a society for cardiovascular magnetic resonance (SCMR) and CMR working group of the european society of cardiology consensus statement**. *Journal of Cardiovascular Magnetic Resonance*, 15:92 – 92.

[Mukhoti et al., 2018] Mukhoti, J., et Gal, Y. (2018). **Evaluating bayesian deep learning methods for semantic segmentation**. *ArXiv*, abs/1811.12709.

[Nair et al., 2020] Nair, T., Precup, D., Arnold, D. L., et Arbel, T. (2020). **Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation**. *Medical image analysis*, 59:101557.

[Nakamori et al., 2018] Nakamori, S., Dohi, K., Ishida, M., Goto, Y., Imanaka-Yoshida, K., Omori, T., Goto, I., Kumagai, N., Fujimoto, N., Ichikawa, Y., Kitagawa, K., Yamada, N., Sakuma, H., et Ito, M. (2018). **Native T1 mapping and extracellular volume mapping for the assessment of diffuse myocardial fibrosis in dilated cardiomyopathy.** *JACC. Cardiovascular imaging*, 11 1:48–59.

[Nath et al., 2020] Nath, V., Yang, D., Landman, B. A., Xu, D., et Roth, H. R. (2020). **Diminishing uncertainty within the training pool: Active learning for medical image segmentation**. *IEEE Transactions on Medical Imaging*, 40:2534–2547.

[National Institutes of Health, 2022] National Institutes of Health, N. (2022). **Magnetic resonance imaging (MRI)**. https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri.

[Neal, 1995] Neal, R. M. (1995). **Bayesian learning for neural networks**.

[Ng et al., 2020] Ng, M., Guo, F., Biswas, L., Petersen, S. E., Piechnik, S. K., Neubauer, S., et Wright, G. A. (2020). **Estimating uncertainty in neural networks for cardiac MRI segmentation: A benchmark study**. *IEEE Transactions on Biomedical Engineering*, 70:1955–1966.

[NHS, 2022] NHS (2022). **Cardiovascular disease**. https://www.nhs.uk/conditions/cardiovascular-disease/.

[Ojha et al., 2022] Ojha, N., et Dhamoon, A. S. (2022). **Myocardial infarction**. In *StatPearls [Internet]*. StatPearls Publishing.

[Oktay et al., 2017] Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M. P., Bai, W., Caballero, J., Cook, S. A., de Marvao, A., Dawes, T. J. W., O'Regan, D. P., Kainz, B., Glocker, B., et Rueckert, D. (2017). **Anatomically constrained neural networks (ACNNs): Application to cardiac image enhancement and segmentation**. *IEEE Transactions on Medical Imaging*, 37:384–395.

[O'Shea et al., 2015] O'Shea, K., et Nash, R. (2015). **An introduction to convolutional neural networks**. *arXiv preprint arXiv:1511.08458*.

[Otto, 2013] Otto, C. M. (2013). **Textbook of clinical echocardiography**. Elsevier Health Sciences.

[Painchaud et al., 2019] Painchaud, N., Skandarani, Y., Judge, T., Bernard, O., Lalande, A., et Jodoin, P.-M. (2019). **Cardiac MRI segmentation with strong anatomical guarantees**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

[Pandit, 2023] Pandit, B. (2023). **A comprehensive guide to the 7 key loss functions in deep learning**. https://dataaspirant.com/loss-functions-in-deep-learning/.

[Paschali et al., 2018] Paschali, M., Conjeti, S., Navarro, F., et Navab, N. (2018). **Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Louppe, G., Prettenhofer, P., Weiss, R., Weiss, R. J., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., et Duchesnay, E. (2011). **Scikit-learn: Machine learning in python**. *J. Mach. Learn. Res.*, 12:2825–2830.

[Perera et al., 2019] Perera, P., Nallapati, R., et Xiang, B. (2019). **Ocgan: One-class novelty detection using gans with constrained latent representations**. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2893–2901.

[Pérez-García et al., 2020] Pérez-García, F., Sparks, R., et Ourselin, S. (2020). **Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning**. *Computer Methods and Programs in Biomedicine*, 208.

[Petitjean et al., 2011] Petitjean, C., et Dacher, J.-N. (2011). **A review of segmentation methods in short axis cardiac MR images**. *Medical image analysis*, 15(2):169–184.

[Petitjean et al., 2015] Petitjean, C., Zuluaga, M. A., Bai, W., Dacher, J.-N., Grosgeorge, D., Caudron, J., Ruan, S., Ayed, I. B., Cardoso, M. J., Chen, H.-C., et others (2015). **Right ventricle segmentation from cardiac MRI: a collation study**. *Medical image analysis*, 19(1):187–202.

[Prince et al., 2023] Prince, E. W., Ghosh, D., Görg, C., et Hankinson, T. C. (2023). **Uncertainty-aware deep learning classification of adamantinomatous craniopharyngioma from preoperative MRI**. *Diagnostics*, 13.

[Puyol-Antón et al., 2020] Puyol-Antón, E., Ruijsink, B., Baumgartner, C. F., Sinclair, M., Konukoglu, E., Razavi, R., et King, A. P. (2020). **Automated quantification of myocardial tissue characteristics from native T1 mapping using neural networks with uncertainty-based quality-control**. *Journal of Cardiovascular Magnetic Resonance*, 22.

[Radau et al., 2009] Radau, P., Lu, Y., Connelly, K., Paul, G., Dick, A. J., et Wright, G. A. (2009). **Evaluation framework for algorithms segmenting short axis cardiac MRI.** *The MIDAS Journal*.

[Rajiah et al., 2013] Rajiah, P., Desai, M. Y., Kwon, D. H., et Flamm, S. D. (2013). **MR imaging of myocardial infarction.** *Radiographics : a review publication of the Radiological Society of North America, Inc*, 33 5:1383–412.

[Reinke et al., 2021] Reinke, A., Eisenmann, M., Tizabi, M. D., Sudre, C. H., Radsch, T., Antonelli, M., Arbel, T., Bakas, S., Cardoso, M. J., Cheplygina, V., Farahani, K., Glocker, B., Heckmann-Notzel, D., Isensee, F., Jannin, P., Kahn, C. E., Kleesiek, J., Kurç, T. M., Kozubek, M., Landman, B. A., Litjens, G. J. S., Maier-Hein, K. H., Menze, B. H., Muller, H., Petersen, J., Reyes, M., Rieke, N., Stieltjes, B., Summers, R. M., Tsaftaris, S. A., van Ginneken, B., Kopp-Schneider, A., Jager, P. F., et Maier-Hein, L. (2021). **Common limitations of image processing metrics: A picture story**. *ArXiv*, abs/2104.05642.

[Reiter et al., 2018] Reiter, U., Reiter, C., Kräuter, C., Fuchsjäger, M., et Reiter, G. (2018). **Cardiac magnetic resonance T1 mapping. Part 2: Diagnostic potential and applications.** *European journal of radiology*, 109:235–247.

[Reynolds, 2019] Reynolds, A. H. (2019). **Convolutional neural networks CNNs**. https://anhreynolds.com/blogs/cnn.html.

[Ripley et al., 2016] Ripley, D. P., Musa, T. A., Dobson, L. E., Plein, S., et Greenwood, J. P. (2016). **Cardiovascular magnetic resonance imaging: what the general cardiologist should know**. *Heart*, 102:1589 – 1603.

[Robinson et al., 2018] Robinson, R., Oktay, O., Bai, W., Valindria, V. V., Sanghvi, M. M., Aung, N. L., Paiva, J. M., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A. M., Carapella, V., Kimm, Y. J., Kainz, B., Piechnik, S. K., Neubauer, S., Petersen, S. E., Page, C., Rueckert, D., et Glocker, B. (2018). **Subject-level prediction of segmentation failure using real-time convolutional neural nets**.

[Rodgers et al., 2019] Rodgers, J. L., Jones, J., Bolleddu, S. I., Vanthenapalli, S., Rodgers, L. E., Shah, K., Karia, K., et Panguluri, S. K. (2019). **Cardiovascular risks associated with gender and aging**. *Journal of Cardiovascular Development and Disease*, 6.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., et Brox, T. (2015). **U-net: Convolutional networks for biomedical image segmentation**. *ArXiv*, abs/1505.04597.

[Roy et al., 2018] Roy, A. G., Conjeti, S., Navab, N., et Wachinger, C. (2018). **Inherent brain segmentation quality control from fully convnet Monte Carlo sampling**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 664–672. Springer.

[Russell et al., 1995] Russell, S. J., et Norvig, P. (1995). **Artificial intelligence: A modern approach**.

[Sander et al., 2019] Sander, J., de Vos, B. D., Wolterink, J. M., et Išgum, I. (2019). **Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI**. In *Medical Imaging 2019: Image Processing*, volume 10949, page 1094919. International Society for Optics and Photonics.

[Schülke, 1968] Schülke, R. (1968). **Anatomy and physiology**. *Zahntechnik; Zeitschrift fur Theorie und Praxis der wissenschaftlichen Zahntechnik*, 9 2:77–9.

[Schultheiss et al., 2019] Schultheiss, H. P., Fairweather, D., Caforio, A. L. P., Escher, F., Hershberger, R., Lipshultz, S. E., Liu, P. P., Matsumori, A., Mazzanti, A., McMurray, J., et Priori, S. G. (2019). **Dilated cardiomyopathy**. *Nature Reviews Disease Primers*, 5:1–19.

[Schulz-Menger et al., 2020] Schulz-Menger, J., Bluemke, D. A., Bremerich, J., Flamm, S. D., Fogel, M. A., Friedrich, M. G., Kim, R. J., von Knobelsdorff-Brenkenhoff, F., Kramer, C. M., Pennell, D. J., Plein, S., et Nagel, E. (2020). **Standardized image interpretation and post-processing in cardiovascular magnetic resonance - 2020 update**. *Journal of Cardiovascular Magnetic Resonance*, 22.

[Scully et al., 2018] Scully, P. R., Bastarrika, G., Moon, J. C., et Treibel, T. A. (2018). **Myocardial extracellular volume quantification by cardiovascular magnetic resonance and computed tomography**. *Current Cardiology Reports*, 20.

[Shaaf et al., 2022] Shaaf, Z. F., Jamil, M. M. A., Ambar, R. B., Alattab, A. A., Yahya, A. A., et Asiri, Y. (2022). **Automatic left ventricle segmentation from short-axis cardiac MRI images based on fully convolutional neural network**. *Diagnostics*, 12.

[Shapiro, 2003] Shapiro, A. (2003). **Monte carlo sampling methods**. *Handbooks in operations research and management science*, 10:353–425.

[Shelhamer et al., 2014] Shelhamer, E., Long, J., et Darrell, T. (2014). **Fully convolutional networks for semantic segmentation**. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440.

[Simonyan et al., 2014] Simonyan, K., et Zisserman, A. (2014). **Very deep convolutional networks for large-scale image recognition**. *arXiv preprint arXiv:1409.1556*.

[Singh et al., 2020] Singh, A., Sengupta, S., et Lakshminarayanan, V. (2020). **Explainable deep learning models in medical image analysis**. *Journal of imaging*, 6(6):52.

[Soberanis-Mukul et al., 2019] Soberanis-Mukul, R. D., Albarqouni, S., et Navab, N. (2019). **An uncertainty-driven GCN refinement strategy for organ segmentation**. *ArXiv*, abs/1906.02191.

[Srivastava et al., 2014] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., et Salakhutdinov, R. (2014). **Dropout: a simple way to prevent neural networks from overfitting**. *J. Mach. Learn. Res.*, 15:1929–1958.

[Stanley Oiseth, 2023] Stanley Oiseth, Lindsay Jones, E. M. (2023). **Takotsubo cardiomyopathy**. https://www.lecturio.com/concepts/takotsubo-cardiomyopathy/.

[Suinesiaputra et al., 2012] Suinesiaputra, A., Cowan, B. R., Finn, J. P., Fonseca, C. G., Kadish, A. H., Lee, D. C., Medrano-Gracia, P., Warfield, S. K., Tao, W., et Young, A. A. (2012). **Left ventricular segmentation challenge from cardiac MRI: a collation study**. In *Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges: Second International Workshop, STACOM 2011, Held in Conjunction with MICCAI 2011, Toronto, ON, Canada, September 22, 2011, Revised Selected Papers 2*, pages 88–97. Springer.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., et Rabinovich, A. (2015). **Going deeper with convolutions**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

[Taha et al., 2015] Taha, A. A., et Hanbury, A. (2015). **Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool**. *BMC Medical Imaging*, 15.

[Tan et al., 2019] Tan, M., et Le, Q. V. (2019). **Efficientnet: Rethinking model scaling for convolutional neural networks**. *ArXiv*, abs/1905.11946.

[Tang et al., 2022] Tang, Y., Yang, D., Li, W., Roth, H. R., Landman, B., Xu, D., Nath, V., et Hatamizadeh, A. (2022). **Self-supervised pre-training of swin transformers for 3D medical image analysis**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740.

[Taylor et al., 2016] Taylor, A. J., Salerno, M., Dharmakumar, R., et Jerosch-Herold, M. (2016). **T1 mapping: Basic techniques and clinical applications.** *JACC. Cardiovascular imaging*, 9 1:67–81.

[Techapanurak et al., 2020] Techapanurak, E., Suganuma, M., et Okatani, T. (2020). **Hyperparameter-free out-of-distribution detection using cosine similarity**. In *Asian Conference on Computer Vision*.

[Thongsongsang et al., 2021] Thongsongsang, R., Songsangjinda, T., Tanapibunpon, P., et Krittayaphong, R. (2021). **Native T1 mapping and extracellular volume fraction for differentiation of myocardial diseases from normal CMR controls in routine clinical practice**. *BMC Cardiovascular Disorders*, 21.

[Tobon-Gomez et al., 2015] Tobon-Gomez, C., Geers, A. J., Peters, J., Weese, J., Pinto, K., Karim, R., Ammar, M., Daoudi, A., Margeta, J., Sandoval, Z., et others (2015). **Benchmark for algorithms segmenting the left atrium from 3D CT and MRI datasets**. *IEEE Transactions on medical imaging*, 34(7):1460–1473.

[Tousignant et al., 2019] Tousignant, A., Lemaître, P., Precup, D., Arnold, D. L., et Arbel, T. (2019). **Prediction of disease progression in multiple sclerosis patients using deep learning analysis of MRI data**. In *International Conference on Medical Imaging with Deep Learning*.

[Tran, 2016] Tran, P. V. (2016). **A fully convolutional neural network for cardiac segmentation in short-axis MRI**. *ArXiv*, abs/1604.00494.

[Treibel et al., 2016] Treibel, T. A., Fontana, M., Maestrini, V., Castelletti, S., Rosmini, S., Simpson, J., Nasis, A., Bhuva, A. N., Bulluck, H., Abdel-Gadir, A., White, S. K., Manisty, C. H., Spottiswoode, B. S., Wong, T. C., Piechnik, S. K., Kellman, P., Robson, M. D., Schelbert, E. B., et Moon, J. C. (2016). **Automatic measurement of the myocardial interstitium: Synthetic extracellular volume quantification without hematocrit sampling.** *JACC. Cardiovascular imaging*, 9 1:54–63.

[Tricia Kinman, 2022] Tricia Kinman, J. R. (2022). **Heart attack symptoms, causes, and treatment**. https://www.healthline.com/health/heart-attack.

[Tsai et al., 2013] Tsai, D.-Y., Matsuyama, E., Chen, H.-M., et others (2013). **Improving image quality in medical images using a combined method of undecimated wavelet transform and wavelet coefficient mapping**. *International Journal of Biomedical Imaging*, 2013.

[Tu et al., 2022] Tu, C., Huang, Z., Deng, Z., Yang, Y., Ma, C., He, J., Ye, J., Wang, H., et Ding, X. (2022). **Self pre-training with single-scale adapter for left atrial segmentation**. In *Challenge on Left Atrial and Scar Quantification and Segmentation*, pages 24–35. Springer.

[Valindria et al., 2017] Valindria, V. V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E. O., Rockall, A. G., Rueckert, D., et Glocker, B. (2017). **Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth**. *IEEE Transactions on Medical Imaging*, 36:1597–1606.

[van Engelen et al., 2019] van Engelen, J. E., et Hoos, H. H. (2019). **A survey on semi-supervised learning**. *Machine Learning*, 109:373–440.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., et Polosukhin, I. (2017). **Attention is all you need**. In *NIPS*.

[Virtanen et al., 2020] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., et SciPy 1.0 Contributors (2020). **SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python**. *Nature Methods*, 17:261–272.

[Vöhringer et al., 2007] Vöhringer, M., Mahrholdt, H., Yilmaz, A., et Sechtem, U. (2007). **Significance of late gadolinium enhancement in cardiovascular magnetic resonance imaging (cmr)**. *Herz Kardiovaskuläre Erkrankungen*, 32:129–137.

[Vyas et al., 2018] Vyas, A., Jammalamadaka, N., Zhu, X., Das, D., Kaul, B., et Willke, T. L. (2018). **Out-of-distribution detection using an ensemble of self supervised leave-out classifiers**. In *European Conference on Computer Vision*.

[Wang et al., 2019] Wang, G., Li, W., Ourselin, S., et Vercauteren, T. (2019). **Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation**. *Frontiers in computational neuroscience*, 13:56.

[Wang et al., 2022] Wang, S., Qin, C., Wang, C., Wang, K., Wang, H., Chen, C., Ouyang, C., Kuang, X., Dai, C., Mo, Y., Shi, Z., Dai, C., Chen, X., Wang, H., et Bai, W. (2022). **The extreme cardiac MRI analysis challenge under respiratory motion (CMRxMotion)**.

[WHO, 2017] WHO, W. H. O. (2017). **Cardiovascular diseases (CVDs)**. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[Wickstrøm et al., 2018] Wickstrøm, K., Kampffmeyer, M. C., et Jenssen, R. (2018). **Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps**. *Medical image analysis*, 60:101619.

[Willemink et al., 2022] Willemink, M. J., Roth, H. R., et Sandfort, V. (2022). **Toward foundational deep learning models for medical imaging in the new era of transformer networks.** *Radiology. Artificial intelligence*, 4 6:e210284.

[Williams et al., 2021] Williams, E., Niehaus, S., Reinelt, J., Merola, A., Mihai, P. G., Roeder, I., Scherf, N., et Hernández, M. d. C. V. (2021). **Quality control for more reliable integration of deep learning-based image segmentation into medical workflows**. *arXiv preprint arXiv:2112.03277*.

[Wilson et al., 2020] Wilson, A. G., et Izmailov, P. (2020). **Bayesian deep learning and a probabilistic perspective of generalization**. *ArXiv*, abs/2002.08791.

[Wu et al., 2021] Wu, K., Peng, H., Chen, M., Fu, J., et Chao, H. (2021). **Rethinking and improving relative position encoding for vision transformer**. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10013–10021.

[Wu et al., 2023] Wu, X., Wagner, P., et Huber, M. F. (2023). **Quantification of uncertainties in neural networks**. *New Digital Work: Digital Sovereignty at the Workplace*, pages 276–287.

[Xie et al., 2017a] Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., et Yuille, A. L. (2017a). **Adversarial examples for semantic segmentation and object detection**. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1378–1387.

[Xie et al., 2017b] Xie, S., Girshick, R., Dollár, P., Tu, Z., et He, K. (2017b). **Aggregated residual transformations for deep neural networks**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

[Xiong et al., 2021] Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., et others (2021). **A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging**. *Medical image analysis*, 67:101832.

[Xu et al., 2021] Xu, G., Wu, X., Zhang, X., et He, X. (2021). **Levit-unet: Make faster encoders with transformer for medical image segmentation**. *ArXiv*, abs/2107.08623.

[Yamashita et al., 2018] Yamashita, R., Nishio, M., Do, R. K. G., et Togashi, K. (2018). **Convolutional neural networks: an overview and application in radiology**. *Insights into imaging*, 9:611–629.

[Yang et al., 2022] Yang, C., Wang, Y., Zhang, J., Zhang, H., Wei, Z., Lin, Z., et Yuille, A. (2022). **Lite vision transformer with enhanced self-attention**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11998–12008.

[Yang et al., 2021] Yang, J., Zhou, K., Li, Y., et Liu, Z. (2021). **Generalized out-of-distribution detection: A survey**. *ArXiv*, abs/2110.11334.

[Yang et al., 2020] Yang, S., et Wang, X. (2020). **A hybrid network for automatic myocardial infarction segmentation in delayed enhancement-MRI**. In *M&Ms and EMIDEC/STACOM@MICCAI*.

[Yang et al., 2017] Yang, X., Bian, C., Yu, L., Ni, D., et Heng, P.-A. (2017). **Class-balanced deep neural network for automatic ventricular structure segmentation**. In *STACOM@MICCAI*.

[Yu et al., 2019] Yu, L., Wang, S., Li, X., Fu, C.-W., et Heng, P.-A. (2019). **Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation**. *ArXiv*, abs/1907.07034.

[Zabihollahy et al., 2018] Zabihollahy, F., White, J. A., et Ukwatta, E. (2018). **Myocardial scar segmentation from magnetic resonance images using convolutional neural network**. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 105752Z. International Society for Optics and Photonics.

[Zhang et al., 2006] Zhang, H., Cholleti, S. R., Goldman, S. A., et Fritts, J. E. (2006). **Meta-evaluation of image segmentation using machine learning**. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 1:1138–1145.

[Zhang, 2020] Zhang, Y. (2020). **Cascaded Convolutional Neural Network for Automatic Myocardial Infarction Segmentation from Delayed-Enhancement Cardiac MRI**. *arXiv preprint arXiv:2012.14128*.

[Zhang et al., 2022a] Zhang, Y., Meng, Y., et Zheng, Y. (2022a). **Automatically segment the left atrium and scars from LGE-MRIs using a boundary-focused nnU-Net**. In *Challenge on Left Atrial and Scar Quantification and Segmentation*, pages 49–59. Springer.

[Zhang et al., 2022b] Zhang, Z., Zhang, H., Zhao, L., Chen, T., Arik, S. Ö., et Pfister, T. (2022b). **Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding**. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3417–3425.

[Zhao et al., 2022] Zhao, Y., Yang, C., Schweidtmann, A. M., et Tao, Q. (2022). **Efficient bayesian uncertainty estimation for nnU-Net**. *ArXiv*, abs/2212.06278.

[Zhou et al., 2016] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., et Torralba, A. (2016). **Semantic understanding of scenes through the ade20k dataset**. *International Journal of Computer Vision*, 127:302–321.

[Zhou et al., 2017] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., et Torralba, A. (2017). **Scene parsing through ade20k dataset**. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130.

[Zhou et al., 2019] Zhou, Y., Shi, C., Lai, B., et Jimenez, G. (2019). **Contrast enhancement of medical images using a new version of the world cup optimization algorithm**. *Quantitative imaging in medicine and surgery*, 9(9):1528.

[Zhou et al., 2020] Zhou, Y., Zhang, K., Luo, X., Wang, S., et Zhuang, X. (2020). **Anatomy prior based u-net for pathology segmentation with attention**. *ArXiv*, abs/2011.08769.

[Zhuang et al., 2019] Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M. P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et others (2019). **Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge**. *Medical image analysis*, 58:101537.

[Zhuang et al., 2022] Zhuang, X., Xu, J., Luo, X., Chen, C., Ouyang, C., Rueckert, D., Campello, V. M., Lekadir, K., Vesal, S., RaviKumar, N., et others (2022). **Cardiac segmentation on late gadolinium enhancement MRI: a benchmark study from multi-sequence cardiac MR segmentation challenge**. *Medical Image Analysis*, 81:102528.

[Zotti et al., 2017] Zotti, C., Luo, Z., Humbert, O., Lalande, A., et Jodoin, P.-M. (2017). **GridNet with automatic shape prior registration for automatic MRI cardiac segmentation**. In *STACOM@MICCAI*.

[Zou et al., 2023] Zou, K. Y., Chen, Z., Yuan, X., Shen, X., Wang, M., et Fu, H. (2023). **A review of uncertainty estimation and its application in medical imaging**. *ArXiv*, abs/2302.08119.

[Çiçek et al., 2016] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., et Ronneberger, O. (2016). **3D U-Net: Learning dense volumetric segmentation from sparse annotation**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

IV

Appendix

A

MICCAI Challenge papers

A.1/ Using Polynomial Loss and Uncertainty Information for Robust Left Atrial and Scar Quantification and Segmentation

Automatic and accurate segmentation of the left atrial (LA) cavity and scar can be helpful for the diagnosis and prognosis of patients with atrial fibrillation. However, automating the segmentation can be difficult due to the poor image quality, variable LA shapes, and small discrete regions of LA scars. In this paper, we proposed a fully-automatic method to segment LA cavity and scar from Late Gadolinium Enhancement (LGE) MRIs. For the loss functions, we propose two different losses for each task. To enhance the segmentation of LA cavity from the multi-center dataset, we present a hybrid loss that leverages Dice loss with a polynomial version of cross-entropy loss (PolyCE). We also utilize different data augmentations that include histogram matching to increase the variety of the dataset. For the more difficult LA scar segmentation, we propose a loss function that uses uncertainty information to improve the uncertain and inaccurate scar segmentation results. We evaluate the proposed method on the Left Atrial and Scar Quantification and Segmentation (LAScarQS 2022) Challenge dataset. It achieves a Dice score of 0.8897 and a Hausdorff distance (HD) of $16.91mm$ for LA cavity and a Dice score of 0.6406 and sensitivity of 0.5853 for LA scar. From the results, we notice that for LA scar segmentation, which has small and irregular shapes, the proposed loss that utilizes the uncertainty estimates generated by the scar yields the best result compared to the other loss functions. For the multi-center LA cavity segmentation, we observe that combining the region-based Dice loss with the pixelwise PolyCE can achieve a good result by enhancing the segmentation result in terms of both Dice score and HD. Furthermore, using moderate-level data augmentation with histogram matching improves the model's generalization capability. Our proposed method won the Left Atrial and Scar Quantification and Segmentation (LAScarQS 2022) Challenge.

This work was presented at the STACOM - MICCAI 2022 conference as part of the Left Atrial and Scar Quantification and Segmentation (LAScarQS 2022) Challenge [Arega et al., 2022b]:

Arega, T. W., Bricq, S., & Meriaudeau, F. (2022). **Using Polynomial Loss and Uncertainty Information for Robust Left Atrial and Scar Quantification and Segmentation**. *In Challenge on Left Atrial and Scar Quantification and Segmentation (pp. 133-144), MICCAI 2022*. Cham: Springer Nature Switzerland. [Arega et al., 2022b]

A.2/ Automatic Quality Assessment of Cardiac MR Images with Motion Artefacts Using Multi-task Learning and K-Space Motion Artefact Augmentation

The movement of patients and respiratory motion during MRI acquisition produce image artefacts that reduce the image quality and its diagnostic value. Quality assessment of the images is essential to minimize segmentation errors and avoid wrong clinical decisions in the downstream tasks. In this paper, we propose automatic multi-task learning (MTL) based classification model to detect cardiac MR images with different levels of motion artefact. We also develop an automatic segmentation model that leverages k-space based motion artefact augmentation (MAA) and a novel compound loss that utilizes Dice loss with a polynomial version of cross-entropy loss (PolyLoss) to robustly segment cardiac structures from cardiac MRIs with respiratory motion artefacts. We evaluate the proposed method on the Extreme Cardiac MRI Analysis Challenge under Respiratory Motion (CMRxMotion 2022) challenge dataset. For the detection task, the multi-task learning-based model that simultaneously learns both image artefact prediction and breath-hold type prediction achieved significantly better results compared to the single-task model, showing the benefits of MTL. In addition, we utilized test-time augmentation (TTA) to enhance the classification accuracy and study aleatoric uncertainty of the images. Using TTA further improved the classification result as it achieved an accuracy of 0.65 and Cohen's kappa of 0.413. From the estimated aleatoric uncertainty, we observe that images with higher aleatoric uncertainty are more difficult to classify than the ones with lower uncertainty. For the segmentation task, the k-space based MAA enhanced the segmentation accuracy of the baseline model. From the results, we also observe that using a hybrid loss of Dice and PolyLoss can be advantageous to robustly segment cardiac MRIs with motion artefact, leading to a mean Dice of 0.9204, 0.8315, and 0.8906 and mean HD95 of 8.09 mm, 3.60 mm and 6.07 mm for LV, MYO and RV respectively on the official validation set. On the test set, the proposed segmentation method was ranked in second place in the segmentation task of CMRxMotion 2022 challenge.

This work was presented at the STACOM - MICCAI 2022 conference as part of the Extreme Cardiac MRI Analysis Challenge under Respiratory Motion (CMRxMotion 2022) challenge [Arega et al., 2022a]:

Arega, T.W., Bricq, S., & Mériaudeau, F. (2022). **Automatic Quality Assessment of Cardiac MR Images with Motion Artefacts Using Multi-task Learning and K-Space Motion Artefact Augmentation**. *In International Workshop on Statistical Atlases and Computational Models of the Heart (pp. 418-428) STACOM@MICCAI 2022*. Cham: Springer Nature Switzerland. [Arega et al., 2022a]

A.3/ Using MRI-specific Data Augmentation to Enhance the Segmentation of Right Ventricle in Multi-disease, Multi-center and Multi-view Cardiac MRI

Accurate segmentation of the right ventricle (RV) from cardiac MRI is essential to evaluate the structure and function of the RV and to further study cardiac disorders. However, it is a difficult task due to its complex crescent shape and the presence of wall irregularities in its cavity. As part of the multi-disease, multi-center, and multi-view RV segmentation in cardiac MRI challenge (M&Ms-2), we propose to solve the problem using a fully automatic deep learning method that employs different data augmentation techniques. More specifically, we applied MRI-specific based, intensity and spatial data augmentation techniques to reduce the variation among the multi-center images with various cardiac pathologies. MRI-specific data augmentation are transformations that simulate image artifacts specific to MRI such as random bias field, random ghosting and random motion artifacts. We evaluate the proposed method in the validation set of the challenge. Among the data augmentation techniques applied, the MRI-specific based data augmentation enhanced the segmentation results of both long-axis and short-axis images in terms of Dice coefficient and Hausdorff Distance (HD). From the experiments, it shows us that the usage of MRI-specific transformations alongside intensity and spatial transformations in cardiac MRI can increase the variety of the training dataset and further help to improve the generalization capabilities of the models in multi-center, multi-disease cardiac MRI images. The proposed method ranked second in the M&Ms-2 challenge.

This work was presented at the STACOM - MICCAI 2021 conference as part of the Multi-Disease, Multi-View & Multi-Center Right Ventricular Segmentation in Cardiac MRI (M&Ms-2) challenge [Arega et al., 2021b]:

Arega, T. W., Legrand, F., Bricq, S., & Meriaudeau, F. (2021, September). **Using MRI-specific data augmentation to enhance the segmentation of right ventricle in multi-disease, multi-center, and multi-view cardiac MRI**. *In International Workshop on Sta-*

tistical Atlases and Computational Models of the Heart (pp. 250-258), MICCAI 2021. Cham: Springer International Publishing. [Arega et al., 2021b]

B

A Simple Uncertainty-based Quality Control for Cardiac MR Images Segmentation

Deep learning-based methods have achieved state-of-the-art results for cardiac MR segmentation. However, inaccuracies from these methods can lead to wrong clinical decisions in subsequent tasks. To identify the incorrect segmentations, experts manually inspect the segmentation results, but this is a very tiresome and time-consuming task. In this work, we propose a simple uncertainty-based quality control (QC) that estimates the quality of segmentation results. It utilizes image-level uncertainty features as input to a random forest-based classifier to determine the quality of the segmentation outputs. First, the cardiac structures are segmented from the cardiac MR images using the deep ensemble-based Bayesian segmentation model. Then the quality of the segmentation output is assessed using the uncertainty-based QC. The Random Forest (RF) classifier uses four image-level uncertainty features as an input to determine the segmentation quality. The four image-level uncertainty features include Dice agreement within deep ensemble samples (DiceWithinSamples), HD agreement within deep ensemble samples (HDWithinsamples) as well as mean of sample variance and mean of predictive entropy. We evaluated the QC methods on cardiac MR segmentation using the Automated Cardiac Diagnosis Challenge (ACDC) and Extreme Cardiac MRI Analysis Challenge under Respiratory Motion (CMRxMotion) datasets. The segmentation models were trained on the ACDC dataset and tested on the ACDC and CMRxMotion datasets to evaluate the performance of the QC methods on the segmentation results of these two public datasets. From the results (Table B.1), our method outperformed other state-of-the-art uncertainty-based QC methods that are based on image, segmentation, and uncertainty maps in detecting bad quality segmentation results. This enhancement was shown in both ACDC and CMRxMotion segmentation results by achieving an F1-score of 84.1% and AUC of 89.34% on the ACDC segmentation results (n=40) and an F1-score of 84.6% and AUC of

92.2% on the CMRxMotion segmentation results (n=139). From the results, we showed that training a classifier using simple inputs that are derived from uncertainty metrics can determine segmentation quality better than the ones that directly use the image, segmentation, and uncertainty map.

Table B.1: Comparison of different uncertainty-based QC methods and the proposed QC method in terms of F1-score and area under the receiver operating characteristic curve (AUC) on ACDC [Bernard et al., 2018] and CMRxMotion cardiac MRI datasets [Wang et al., 2022]. The best results are in bold.

QC Methods	**ACDC**		**CMRxMotion**	
	F1-Score (%)	**AUC (%)**	**F1-Score (%)**	**AUC (%)**
Img-Seg-Uncert [Devries et al., 2018b, Chen et al., 2020c]	80.8	83.8	78.5	85.4
Seg-Uncert [Williams et al., 2021]	81.3	85.9	80.1	86.4
Proposed method	**84.1**	**89.34**	**84.6**	**92.2**

This work has been accepted as part of the first edition of the French Colloquium on Artificial Intelligence in Biomedical Imaging (IABM) 2023. It is an extension of our proposed QC method, which is described in Chapter 5.

Arega, T.W., Bricq, S., & Mériaudeau, F. (2023). **A Simple Uncertainty-based Quality Control for Cardiac MR Images Segmentation**. *Colloque Français d'Intelligence Artificielle en Imagerie Biomédicale (IABM 2023).*

Document generated with LaTeX and:
the LaTeX style for PhD Thesis created by S. Galland — http://www.multiagent.fr/ThesisStyle
the tex-upmethodology package suite — http://www.arakhne.org/tex-upmethodology/

école doctorale **sciences pour l'ingénieur et microtechniques**

Title: Uncertainty-based Deep Learning Methods for Robust and Reliable Cardiac MRI Segmentation and Analysis

Keywords: Cardiac MRI Segmentation, Myocardial scar, LGE-MRI, T1 mapping MRI, ECV, Uncertainty Estimation, Quality Control, Out-of-distribution (OOD) detection

Abstract:

Deep learning-based segmentation methods have shown promise in automating the segmentation of cardiac MRI images, but they still face challenges in robustly segmenting small, ambiguous regions with irregular shapes like myocardial scars. Additionally, these models struggle with domain shifts and out-of-distribution (OOD) samples, which makes them unreliable and limits their usage in clinical practice. To improve the segmentation of myocardial scars, a segmentation model is proposed that integrates uncertainty information into the learning process. By utilizing Monte-Carlo dropout-based Bayesian neural networks, uncertainty is estimated and incorporated into the loss function, resulting in improved segmentation accuracy and probability calibration. To enhance the reliability of segmentation models, an uncertainty-based quality control (QC) framework is introduced to identify failed segmentations before further analysis. The QC framework utilizes a Bayesian Swin transformer-based U-Net for the segmentation of T1 mapping images and employs image-level uncertainty features to detect poorly segmented images. Experimental results on private and public datasets demonstrate that the proposed QC method significantly outperforms other state-of-the-art uncertainty-based QC methods, particularly in challenging scenarios. Furthermore, a post-hoc OOD detection method is proposed to identify and reject outlier images. This method utilizes the encoder features of the segmentation model and similarity metrics to enhance the trustworthiness of segmentation models. Experimental results demonstrate that the proposed method outperforms state-of-the-art feature space-based and uncertainty-based OOD detection methods across the various OOD datasets.

Titre : Méthodes d'apprentissage profond basées sur l'incertitude pour une segmentation et une analyse robustes et fiables de l'IRM cardiaque

Mots-clés : Segmentation IRM cardiaque, cicatrice myocardique, IRM LGE, IRM de cartographie T1, volume extracellulaire, estimation de l'incertitude, contrôle qualité, détection hors distribution (OOD)

Résumé :

Les méthodes de segmentation basées sur l'apprentissage profond se sont révélées prometteuses pour automatiser la segmentation des images IRM cardiaques, mais elles sont toujours confrontées à des défis pour segmenter de manière robuste de petites régions ambiguës aux formes irrégulières comme les cicatrices myocardiques. De plus, ces modèles sont confrontés aux changements de domaine et aux échantillons hors distribution (OOD), ce qui les rend peu fiables et limite leur utilisation dans la pratique clinique. Pour améliorer la segmentation des cicatrices myocardiques, un modèle de segmentation est proposé qui intègre les informations d'incertitude dans le processus d'apprentissage. En utilisant des réseaux neuronaux bayésiens basés sur une méthode Monté Carlo Drop out, l'incertitude est estimée et incorporée dans la fonction de perte, ce qui améliore la précision de la segmentation et l'étalonnage des probabilités. Pour améliorer la fiabilité des modèles de segmentation, un cadre de contrôle qualité (CQ) basé sur l'incertitude est introduit pour identifier les segmentations ayant échoué avant une analyse plus approfondie. Le cadre CQ utilise un U-Net basé sur un Transformer bayésien Swin pour la segmentation des images cartographiques T1 et utilise des caractéristiques d'incertitude au niveau de l'image pour détecter les images mal segmentées. Les résultats expérimentaux sur des ensembles de données privés et publics démontrent que la méthode de CQ proposée surpasse considérablement les autres méthodes de CQ de l'état de l'art basées sur l'incertitude, en particulier dans des scénarios difficiles. De plus, une méthode de détection OOD post-hoc est proposée pour identifier et rejeter les images aberrantes. Cette méthode utilise les fonctionnalités d'encodeur du modèle de segmentation et les métriques de similarité pour améliorer la fiabilité des modèles de segmentation. Les résultats expérimentaux démontrent que la méthode proposée surpasse les méthodes de détection OOD de l'état de l'art basées sur l'espace des caractéristiques et l'incertitude dans les différents ensembles de données OOD.

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon, France