Karlsruhe Institute of Technology

Challenges and Opportunities of End-to-End Learning in Medical Image Classification

Paul Ferdinand Jäger

This dissertation is submitted for the Degree of Doctor of Engineering

Supervisors:

Prof. Dr. Rainer Stiefelhagen
Prof. Dr. Klaus Maier-Hein

Institute for Anthropomatics and Robotics

April, 2020

Challenges and Opportunities of End-to-End Learning in Medical Image Classification

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

von der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Paul Ferdinand Jäger

Tag der mündlichen Prüfung: 25.06.2020

1. Referent/Referentin: Prof. Dr. Rainer Stiefelhagen
2. Referent/Referentin: Prof. Dr. Klaus Maier-Hein

Acknowledgements

First and foremost I would like to thank Klaus Maier-Hein who has been an excellent supervisor and inspiring mentor throughout the past three and a half years. I am grateful for all the things I learned from you as a researcher and group leader.

I would also like to express my deep gratitude towards Rainer Stiefelhagen for accepting me as an external PhD student at KIT and taking the time to provide valuable and extensive feedback from the very start until the completion of this thesis.

A heartfelt thank you goes to my colleague and friend Fabian Isensee for tutoring me on all things engineering throughout the years and for being my partner-in-crime throughout two projects contributing to this thesis.

I would like to further mention my clinical partners Sebastian Bickelhaupt, Thomas Weikert, and Alexander Sauter for entrusting me with their projects and data and a special thanks to Sebastian for providing precious feedback on the clinical part of this thesis.

I owe deep gratitude to Michal Drozdzal and Adriana Romero, my supervisors at Facebook AI Research, who gave me the great opportunity to experience the workings of core Machine Learning Scientists in a corporate environment. The six months under your supervision have been an invaluable part of my development as a researcher.

Thank you to the students I supervised, Jennifer Kamphenkel and Gregor Ramien, for stimulating discussions and for partnering with me on two projects contributing to this thesis.

A huge thanks goes to my colleagues, friends and co-authors Simon Kohl, Peter Full and Jens Petersen. They say you should separate business from friendship, well, they are wrong, it has been an absolute pleasure working with you. And to Sebastian Wirkert for taking me to lunch on my first day (well not really, in fact you helped me to install CUDA, but that counts).

I would like to express my appreciation for the continuous support of my family, Horst, Anna and Michi. Thank you for always being there, and be it in the form of remote Siedler-sessions throughout the fabrication of this thesis. Special thanks go to Polo for procrastinating with me and always being keen to go out and pee on trees or bark at kids.

Last but not least I want to thank my amazing wife Elsa. Without her support, understanding and putting up with over-hours, work-through weekends, or time spent abroad, none of this would have been possible. Thank you so much!

Abstract (English)

The paradigm of end-to-end learning has revolutionized computer vision in recent years, but clinical application is lagging behind. Image-based computer aided diagnosis systems are still largely based on highly engineered and domain specific pipelines, which consist of independent rule-based models reflecting the subtasks of image classification: Localization of discriminative regions, feature extraction, and decision making. The promise of superior decision making in end-to-end learning derives from removing domain specific prior constraints of limited complexity and instead optimizing all system components simultaneously, directly based on the raw data and with respect to the ultimate task at hand. The reasons for why these advantages have not found their way into clinics yet, i.e. the challenges faced when developing deep learning based diagnosis systems, are manifold: The fact that the generalization ability of learning algorithms scales with how well available training data represents the true underlying data distribution does not play well for medical applications. Annotated datasets in this domain are notoriously small, since labeling involves costly delineation by experts and concatenation of datasets is often hampered by privacy issues and patient rights. Moreover, medical datasets exhibit drastically varying properties with respect to image modalities, acquisition protocols, or anisotropies and the often ambiguous evidence in medical images may propagate to inconsistent or erroneous training annotations. While the data shift between research environment and real life results in diminished model robustness and is thus considered the key obstacle towards clinical translation, this gap is often amplified by nuisance factors such as hardware constraints or granularity of available annotations, which might lead to discrepancies between the modeled task and the underlying clinical question.

This thesis studies the potential of end-to-end learning in clinical diagnosis systems and presents contributions towards some of the key challenges that currently prevent widespread clinical application.

First we attend to the last part of the classification pipeline, the categorization into clinical pathologies. We demonstrate how replacing the current clinical standard of rule-based decision making by large scale feature extraction followed by machine learning-based classification significantly improves breast cancer classification on MRI and accomplishes human-level performance. This approach is further showcased on cardiac diagnosis achiev-

ing the second rank in an international competition. Second, following the paradigm of end-to-end learning, we substitute the biophysical model applied for image normalization in MRI as well as the extraction of handcrafted features with a dedicated CNN architecture and provide an in-depth analysis revealing hidden potential in learned image normalization and a complementary value of learned representations over handcrafted features. While this approach operates on regions of interest and hence relies on manual annotation, in the third part, we include the task of localizing those regions into the learning process to enable true end-to-end diagnosis starting at the raw images. We identify a largely neglected predicament between the strive for evaluating models at clinically relevant scales on one side, and optimizing for efficient training under the burden of data scarcity on the other side. We propose a deep learning model that helps to resolve this trade-off, provide extensive experiments on three medical datasets as well as a series of toy experiments that examines the behavior under limited training data in detail, and open source a comprehensive framework including the first 3D implementations of prevalent object detection models. We identify further leverage points in existing end-to-end learning systems, where domain knowledge can serve as an inductive bias to increase the robustness of deep learning models in medical image classification aiming to pave the way for application in clinical practice. To this end, we address the challenge of erroneous training annotations by substituting the classification component of end-to-end object detection for regression, which enables to train models directly on the continuous scale of underlying pathological processes, thus elevating the models' robustness against rater confusions. Further, we address the challenge of input heterogeneities faced by trained models when deployed across clinical sites by proposing model-based domain adaptation, which enables to recapture the original training domain given altered inputs and thus restores robust generalization. Finally, we address the highly unsystematic, elaborate and subjective trial-and-error process of finding a robust set of hyperparameters for a given task by condensing domain knowledge into a set of key design choices and systematic rules enabling automated and robust deep learning pipeline configuration on a large variety of medical datasets.

To conclude, the work presented here demonstrates the vast potential of end-to-end learning algorithms compared to the clinical standard of compound engineered diagnosis pipelines and presents solutions towards some of the key challenges preventing real life application such as data scarcity, discrepancy between addressed task and underlying clinical question, ambiguity in training annotations, or domain shifts across clinical sites. These contributions tie in to the overarching goal of automating medical image classification - an integral factor of the transformation required to shape the future of health care.

Abstract (German)

Das Paradigma des End-to-End Lernens hat in den letzten Jahren die Bilderkennung revolutioniert, aber die klinische Anwendung hinkt hinterher. Bildbasierte computergestützte Diagnosesysteme basieren immer noch weitgehend auf hochtechnischen und domänenspezifischen Pipelines, die aus unabhängigen regelbasierten Modellen bestehen, welche die Teilaufgaben der Bildklassifikation wiederspiegeln: Lokalisation von auffälligen Regionen, Merkmalsextraktion und Entscheidungsfindung. Das Versprechen einer überlegenen Entscheidungsfindung beim End-to-End Lernen ergibt sich daraus, dass domänenspezifische Zwangsbedingungen von begrenzter Komplexität entfernt werden und stattdessen alle Systemkomponenten gleichzeitig, direkt anhand der Rohdaten, und im Hinblick auf die letztendliche Aufgabe optimiert werden. Die Gründe dafür, dass diese Vorteile noch nicht den Weg in die Klinik gefunden haben, d.h. die Herausforderungen, die sich bei der Entwicklung Deep Learning-basierter Diagnosesysteme stellen, sind vielfältig: Die Tatsache, dass die Generalisierungsfähigkeit von Lernalgorithmen davon abhängt, wie gut die verfügbaren Trainingsdaten die tatsächliche zugrundeliegende Datenverteilung abbilden, erweist sich in medizinische Anwendungen als tiefgreifendes Problem. Annotierte Datensätze in diesem Bereich sind notorisch klein, da für die Annotation eine kostspielige Beurteilung durch Experten erforderlich ist und die Zusammenlegung kleinerer Datensätze oft durch Datenschutzauflagen und Patientenrechte erschwert wird. Darüber hinaus weisen medizinische Datensätze drastisch unterschiedliche Eigenschaften im Bezug auf Bildmodalitäten, Bildgebungsprotokolle oder Anisotropien auf, und die oft mehrdeutige Evidenz in medizinischen Bildern kann sich auf inkonsistente oder fehlerhafte Trainingsannotationen übertragen. Während die Verschiebung von Datenverteilungen zwischen Forschungsumgebung und Realität zu einer verminderten Modellrobustheit führt und deshalb gegenwärtig als das Haupthindernis für die klinische Anwendung von Lernalgorithmen angesehen wird, wird dieser Graben oft noch durch Störfaktoren wie Hardwarelimitationen oder Granularität von gegebenen Annotation erweitert, die zu Diskrepanzen zwischen der modellierten Aufgabe und der zugrunde liegenden klinischen Fragestellung führen.

Diese Arbeit untersucht das Potenzial des End-to-End-Lernens in klinischen Diagnosesystemen und präsentiert Beiträge zu einigen der wichtigsten Herausforderungen, die derzeit eine breite klinische Anwendung verhindern.

Zunächst wird der letzten Teil der Klassifikations-Pipeline untersucht, die Kategorisierung in klinische Pathologien. Wir demonstrieren, wie das Ersetzen des gegenwärtigen klinischen Standards regelbasierter Entscheidungen durch eine groß angelegte Merkmalsextraktion gefolgt von lernbasierten Klassifikatoren die Brustkrebsklassifikation im MRT signifikant verbessert und eine Leistung auf menschlichem Level erzielt. Dieser Ansatz wird weiter anhand von kardiologischer Diagnose gezeigt. Zweitens ersetzen wir, dem Paradigma des End-to-End Lernens folgend, das biophysikalische Modell, das für die Bildnormalisierung in der MRT angewandt wird, sowie die Extraktion handgefertigter Merkmale, durch eine designierte CNN-Architektur und liefern eine eingehende Analyse, die das verborgene Potenzial der gelernten Bildnormalisierung und einen Komplementärwert der gelernten Merkmale gegenüber den handgefertigten Merkmalen aufdeckt. Während dieser Ansatz auf markierten Regionen arbeitet und daher auf manuelle Annotation angewiesen ist, beziehen wir im dritten Teil die Aufgabe der Lokalisierung dieser Regionen in den Lernprozess ein, um eine echte End-to-End-Diagnose baserend auf den Rohbildern zu ermöglichen. Dabei identifizieren wir eine weitgehend vernachlässigte Zwangslage zwischen dem Streben nach der Auswertung von Modellen auf klinisch relevanten Skalen auf der einen Seite, und der Optimierung für effizientes Training unter Datenknappheit auf der anderen Seite. Wir präsentieren ein Deep Learning Modell, das zur Auflösung dieses Kompromisses beiträgt, liefern umfangreiche Experimente auf drei medizinischen Datensätzen sowie eine Serie von Toy-Experimenten, die das Verhalten bei begrenzten Trainingsdaten im Detail untersuchen, und publiziren ein umfassendes Framework, das unter anderem die ersten 3D-Implementierungen gängiger Objekterkennungsmodelle umfasst. Wir identifizieren weitere Hebelpunkte in bestehenden End-to-End-Lernsystemen, bei denen Domänenwissen als Zwangsbedingung dienen kann, um die Robustheit von Modellen in der medizinischen Bildanalyse zu erhöhen, die letztendlich dazu beitragen sollen, den Weg für die Anwendung in der klinischen Praxis zu ebnen. Zu diesem Zweck gehen wir die Herausforderung fehlerhafter Trainingsannotationen an, indem wir die Klassifizierungskompnente in der End-to-End-Objekterkennung durch Regression ersetzen, was es ermöglicht, Modelle direkt auf der kontinuierlichen Skala der zugrunde liegenden pathologischen Prozesse zu trainieren und so die Robustheit der Modelle gegenüber fehlerhaften Trainingsannotationen zu erhöhen. Weiter adressieren wir die Herausforderung der Input-Heterogenitäten, mit denen trainierte Modelle konfrontiert sind, wenn sie an verschiedenen klinischen Orten eingesetzt werden, indem wir eine modellbasierte Domänenanpassung vorschlagen, die es ermöglicht, die ursprüngliche Trainingsdomäne aus veränderten Inputs wiederherzustellen und damit eine robuste Generalisierung zu gewährleisten. Schließlich befassen wir uns mit dem höchst unsystematischen, aufwendigen und subjektiven Trial-and-Error-Prozess zum Finden von robusten Hyperparametern für einen gegebene Aufgabe, indem wir Domänenwissen in ein Set systematischer Regeln überführen, die eine automatisierte und robuste Konfiguration von Deep Learning Modellen auf einer Vielzahl von medizinischen Datensetzen ermöglichen.

Zusammenfassend zeigt die hier vorgestellte Arbeit das enorme Potenzial von End-to-End Lernalgorithmen im Vergleich zum klinischen Standard mehrteiliger und hochtechnisierter Diagnose-Pipelines auf, und präsentiert Lösungsansätze zu einigen der wichtigsten Herausforderungen für eine breite Anwendung unter realen Bedienungen wie Datenknappheit, Diskrepanz zwischen der vom Modell behandelten Aufgabe und der zugrunde liegenden klinischen Fragestellung, Mehrdeutigkeiten in Trainingsannotationen, oder Verschiebung von Datendomänen zwischen klinischen Standorten. Diese Beiträge können als Teil des übergreifende Zieles der Automatisierung von medizinischer Bildklassifikation gesehen werden - ein integraler Bestandteil des Wandels, der erforderlich ist, um die Zukunft des Gesundheitswesens zu gestalten.

Contents

1. Introduction

Radiological images enclose hidden treasures in form of unstructured information. The exclusive value compared to other clinical data sources lies within the non-invasive characterization of tissue or organs and its resolution in time and space. The task of unlocking this potential and making it accessible to clinical decisions is referred to as medical image analysis. Despite tremendous efforts from the computer science community since the 1960s [1, 2, 3], the readout of radiological images in clinical work life remains to be performed by humans. This status quo is problematic in multiple ways: Global digitalization and aging societies drive data growth in health care at an unprecedented speed. While the amount of available data has increased by 878% between 2016 and 2018[4], the amount of human resources for data interpretation has been near constant [5] - an alarming trend, given that halving the time spent per report increases human error rate by 17% [6].

Moreover, the growing abundance of data itself represents a key aspect of the potential hidden in radiological images, which is currently sealed: Manual readouts result in highly subjective and non-quantified free-text reports, which impede to aggregate information beyond single patients [7]. In order to draw general conclusions that are able to drive scientific and medical progress, however, information needs to be combined and clustered across large cohorts, which requires systematic extraction and quantification. This desideratum follows the paradigm of "Precision Medicine", a term coined in the early 2010s to picture the vision of how digitalization will shape the future of health care [8]. Specifically, it describes the scientific idea of leveraging growing data and compute resources ("Big Data") to comprehend medical data in its entirety, including genetic data ("Genomics", where "-omics" signifies "studying the totality of something"), proteomic data ("Proteomics") or metabolic data ("Metabolomics"). The ultimate goal of Precision Medicine is to enable precise treatment tailored towards individual patients by means of a holistic patient biomarker, which is generated by aggregation of the patient's entire

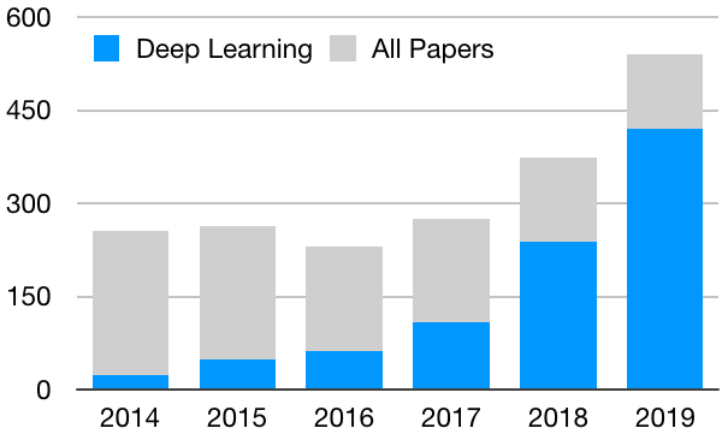

Figure 1.1.: **Increasing number of deep learning related publications at MICCAI 2014-2019.** The international conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) is the largest of its kind in the field of medical image computing. This figure shows how the ratio of deep-learning related contributions has drastically increased in recent years and how this trend effectively doubled the size of the conference.

clinical information and subsequent analysis with respect to vast pools of standardized data cohorts. In 2014, "Radiomics", i.e. the comprehensive and quantitative study of radiological images in form of large-scale visual measurements, was proposed as an addition to the data pool [9, 10, 11]. Examples for the complementary value added by Radiomics are time and space resolution of underlying pathologic processes: The precise location of a tumor and its progress over time or the motion of a beating heart are not accessible from genetic or proteomic data.

Regarding the transformation of raw images into clinical decisions, i.e. the cassification of medical images, however, Radiomics only addresses a fraction of the task: Necessary steps commonly include normalization of images, registration of modalities, localization of Regions of Interest (ROIs), extraction of features and clinical categorization. Thus, while Radiomics is able to generalize the feature extraction process by applying comprehensive and task-agnostic sets of measurements, the associated clinical workflow either still relies on manual effort for the remaining steps or integrates Radiomics as part of a Computer Assisted Diagnosis (CAD) system. These CAD systems, first introduced in the 1980s [12], constitute highly engineered, compound and task-specific pipelines with limited performance and cost efficiency and thus have failed to achieve widespread clinical usage [13, 14].

There is one recent course of events, however, that has been able to re-spark the hope

for successful automation of medical image classification: Starting in 2015, deep learning methods took the field by storm. Figure 1.1 shows, exemplary for the largest conference in medical image analysis, how the ratio of contributions related to deep learning went from 9% in 2014 to 78% in 2019 and how this trend effectively doubled the size of the conference. This development further manifests in the reporting of remarkable image classification results including human-level or even superhuman-level performance of deep learning methods e.g. in breast cancer detection [15], skin cancer detection [16], analysis of electrocardiograms [17] or interpretation of chest x-rays [18]. The overwhelmingly positive research outcomes have spawned high expectations regarding the impact of deep learning on the future of health care, for instance improvement in overall efficiency, the prevention of medical errors that will affect almost every patient during their lifetime [19], or the detection of novel signals of disease that clinicians are unable to perceive [20, 21]. Despite the evident potential, however, deployment in clinical practice is lagging behind: A review from 2019 showed that at the time merely 14 deep learning based systems have been approved by the U.S Food and Drug administration for clinical use [22]. The cautious attitude towards clinical approval follows a disappointing wave of first generation systems such as IBM Watson, that dashed against over-inflated expectations [23]. The discrepancy between successful AI research and stuttering performance in real life applications is a widely discussed issue and generally referred to as the "AI chasm" [24]. In the medical domain, a fraction of this chasm can be related to unrealistic or biased setups of human benchmark evaluations [25, 26], incongruous evaluation metrics [27], or socio-environmental factors in prospective deployment [28], but the question remains as to what are the profound and underlying challenges of deep learning in medical image classification. What impedes this powerful technology from enhancing the live of patients and health care workers?

The answer to this question lies within the very nature of deep learning, i.e. the postulate to fit heavily over-parametrized functions to raw data and thus to optimize all system components simultaneously, in an end-to-end fashion and with respect to the ultimate target. In theory, the superiority of end-to-end learning derives from minimization of inductive biases and the implied ability to generate highly complex decision boundaries that are able to catch even the tiniest peculiarities in the data [29]. In practice, however, and especially in application domains, this advantageous effect is hampered by data scarcity: The annotated data available at training time commonly represents only a fraction of the underlying data distribution, thus generated decision boundaries turn out overly complex and with limited generalization to unseen data. In Medical Imaging, the generalization problem is fueled by domain specific characteristics such as notoriously small datasets available for model development, high cost of manual expert annotations, large image sizes, input shifts due to varying scanner protocols across clinical sites or ambiguities in images that propagate to erroneous annotations [30]. Taken together, these

peculiarities amplify the data shift between research environments and clinical practice resulting in insufficient robustness of deep learning-based medical image classification, hence preventing successful real life deployment [27]. Thus, what are the means to address this issue and which attempts are currently made by research?

As the generalization problem is inherent to machine learning it is widely tackled by the respective community itself. Research on regularization techniques, representation learning, semi-supervised or self-supervised learning, out of distribution modeling or domain adaptation can all be related to the underlying challenge of generalizing beyond training data [29]. There is, however, an additional leverage point that needs to be seized from within the applied research domain: On a broader perspective, the success-story of deep learning is a story of overcoming domain expertise. Knowledge, aggregated over decades and condensed into rule-based systems to solve scientific problems, in many cases has been revealed by deep learning as a constraint that is better left aside. In the context of limited training data, however, the removal of any piece of domain knowledge from a model essentially replaces an inductive bias with additional degrees of freedom during the learning process and hence increases the amount of data required to fit the associated parameters. Vice versa, domain knowledge has the potential to serve as an inductive bias that shortcuts the learning process and alleviates the data burden [31]. Thus, depending on the task and the available data, in every model there exists a sweet spot between inductive biases and learning parameters.

Taken all together, it is crucial for research in applied domains such as medical image analysis to scrutinize the dogma of end-to-end learning, i.e. to carefully counterbalance the power of deep learning with the potential of domain knowledge, in order to accomplish the model robustness required for real life application.

1.1. Contributions

Following the above statement, this thesis studies the challenges and opportunities of end-to-end learning in medical image classification throughout four methodological chapters.

Medical Image Classification based on Handcrafted Features (Chap. 4)
As a first step, we attend to the last part of the classification pipeline, i.e. the categorization of entities into clinical pathologies based on handcrafted features. Specifically, we explore the potential of replacing rule-based decision making as currently deployed in clinical standard, which is often based on a single or few extracted features, with multivariate learning algorithms based on large-scale Radiomics. The main contributions in this chapter are:

- We achieve human-level performance in breast cancer lesion classification on Diffu-

sion Weighted Magentic Resonance Imaging (DWI) by substituting a mono-parametric decision threshold with large-scale Radiomics feature extraction followed by a machine learning classifier [32].

- We provide insights about further parts of the diagnosis pipeline such as an in-depth analysis of Radiomics features or the exploitation of domain knowledge to enhance the image normalization model.
- We further apply the approach of large scale feature extraction followed by machine learning classifiers to cardiac disease classification on MRI, thereby extend the feature set by a novel group of time-series measurements over the cardiac cycle, and achieve the second rank in an international competition [33, 34].

Medical Image Classification based on Learned Representations (Chap. 5)
We further follow the paradigm of end-to-end learning, i.e. the idea that enabling simultaneous optimization of all pipeline components with respect to the ultimate clinical target improves upon compound rule-based diagnosis pipelines. To this end, we substitute the biophysical model applied for image normalization in DWI as well as the extraction of handcrafted (Radiomic) features with respective learning algorithms, which operate on previously annotated Regions of Interest (ROIs) in the image. The main contributions in this chapter are:

- We propose a Convolutional Neural Network (CNN) architecture designed to integrate the biophysical model for image normalization, handcrafted feature extraction as well as clinical categorization, so as to enable ROI-classification of breast lesions on DWI by means of learned representations [35].
- We reveal potential hidden in DWI by demonstrating the benefits of learned image normalizations as compared to the biophysical model currently deployed in clinical research.
- We provide results indicating a complementary value of representations learned in the CNN with respect to handcrafted feature extraction [36].

End-to-end Medical Image Classification (Chap. 6)
When including the task of localizing ROIs into the learning process to enable true end-to-end diagnosis starting at the raw images, there are three current deep learning methodologies to be considered that attend to the problem at three different *levels of granularity*: Whole Image classification for patient level decisions, Object detection for object level decisions, and Semantic Segmentation for pixel-level decisions. These three levels translate to specific model evaluation metrics and in return answer to different clinical questions. We identify a largely neglected predicament between the strive for crossing the AI chasm

by evaluating models at clinically relevant scales on one side, and optimizing for efficient training under the burden of data scarcity on the other side. The main contributions in this chapter are:

- We propose a deep learning model that enables end-to-end object detection and classification on medical images by aligning the model output to the clinically relevant scale while maintaining data efficient training [37].
- We provide an in-depth analysis of the prevalent models from object detection, semantic segmentation and instance segmentation operating in 2D as well as 3D by means of comparative studies on Breast DWI, Lung Computed Tomography (CT) and a series of toy experiments.
- We open source the Medical Detection Toolkit, the first comprehensive framework for object detection on medical images including e.g. modular implementations of all explored models operating in 2D and 3D [38].
- We apply our approach to the task of lung cancer staging on Positron Emission Tomography - Computed Tomography (PET-CT) and perform a sensitivity study under varying clinical training scenarios [39].

Increasing Robustness of Medical Image Classification (Chap. 7)
We continue the path of exploring pitfalls in existing end-to-end learning systems that currently hamper robust generalization. To this end, we identify key leverage points where domain knowledge can be condensed into inductive biases and increase the robustness of end-to-end models in medical image classification aiming to pave the way for application in clinical practice. The main contributions in this chapter are:

- *Robustness against rater confusion:* We address the challenge of erroneous training annotations by substituting the classification component of end-to-end object detection for regression, which enables to train models directly on the continuous scale of underlying pathological processes [40].
- *Robustness against input variations:* When trained models are deployed across clinical sites they commonly face performance drops due to input domain shifts such as missing or altered modalities. We inject domain knowledge in form of a biophysical model that recaptures the original training domain from altered inputs and thus restores robust model generalization. [41].
- *Robust Hyperparameters:* We address the highly unsystematic, cumbersome and subjective trial-and-error process of finding a robust set of hypereparameters for a given task by condensing domain knowledge into a set of key design choices and systematic rules thus enabling automated and robust deep learning pipeline configuration on a large variety of medical datasets. [42].

1.2. Outline

The outline of this thesis is as follows: Chapter 2 introduces the fundamentals of medical imaging techniques, current procedures of cancer diagnosis and the State of the Art in medical image classification. Chapters 4-7 represent methodological chapters as outlined in Section 1.1. Finally, Chapter 8 provides concluding thoughts and an outlook on the future of automation in medical image classification and its impact on health care.

2. Clinical Foundations

2.1. Medical Imaging Techniques

This section is to large parts based on the book "Medizinische Physik: Grundlagen - Bildgebung - Therapie - Technik" by Schlegel et al. [43].

2.1.1. Magentic Resonance Imaging

Since the beginning of the 1980s Magnetic Resonance Imaging (MRI) has become a widely used medical imaging technique, which is able to image a wide variety of anatomical regions and to provide answers to a wide range of morphological and functional questions [43]. MRI allows to generate high-resolution, good-contrast medical images without exposure to harmful radiation, while drawbacks include long acquisition times, relatively high hardware costs and agnosticism to bone and calcium. MRI measures the magnetic resonance of nuclei after excitation by an external magnetic field. Due to the high abundance of hydrogen in the human body and its advantageous magnetic interaction properties, most medical applications of MRI are targeted towards measuring magnetic resonance of hydrogen nuclei. Specifically, the inherent property of a quantum-mechanic spin in nuclei and the fact that this spin couples to external magnetic fields is exploited by applying radio frequency pulses to the tissue in order to excite the nuclear spin onto a higher energy level. The subsequent relaxation emits an electromagnetic signal ("resonance"), which is measured in a receiver coil in terms of time and frequency resolution. By means of an inverse Fourier Transformation and due to a multi-coil setup inducing linear field gradients along spatial axes, the spatial information of magnetic resonance can be reconstructed from these measurements in hindsight. Thereby, different pulse sequences and timings ("contrasts") result in varying image modalities, such as a "*T1-weighted*" (focused on measuring the resonance parallel to the external magnetic field) or "*T2-weighted*" (focused on measuring the resonance transverse to the external magnetic field).

Contrast Enhanced MRI

The contrast in MRI, for example between neighboring tissues, can be additionally increased by application of appropriate MRI contrast agents [43]. The mode of action of such contrast medium in MRI is not based on direct imaging of a substance as such, but its magnetic interaction with the environment. Complementary clinical information in contrast-enhanced images derives from the fact that metabolic changes in diseased tissue affect the absorption capabilities of contrast medium. Most clinically used MRI contrast agents affect the measurement by changing the T1 relaxation time after excitation. During the last 30 years there has been rapid improvement of existing and development of new MRI contrast agents. Currently, there exist different, largely overarching classifications of MRI contrast agents that are based on a variety of features such as the nature of their core, their effect on the surrounding space or their chemical or magnetic properties. MRI contrast agents may be administered by injection into the blood stream or orally, depending on the subject of interest. Recent studies found that injection of contrast agents might be harmful, e.g. via allergic reactions, the rarely occurring nephrogenic systemic fibrosis, or intracerebral deposits, even though no evidence for short-term sequelae has been demonstrated so far [44, 45, 46, 47]. Further, contrast enhancement considerably increases acquisition time compared to conventional MRI [43].

Diffusion Weighted MRI

A light-weight alternative to invasive and time consuming contrast enhanced MRI is Diffusion Weighted Magentic Resonance Imaging (DWI). Given an appropriate pulse sequence, MRI is able to measure molecular diffusion, i.e. the Brownian motion of water molecules. The diffusion of water in vivo depends on cellular parameters such as cell dimensionality, compartmentation and transport processes [Posse et al., 1993]. In simplified terms, the mobility of water molecules is expected to decrease with higher cell density, such as in cancerous tissue, rendering DWI a widely utilized technique in cancer imaging: The diffusion signal is measured as the *spin echo* between two pulses, where nuclei changing their location between the pulses result in reduced signal [48]. Thus, voxels representing cancerous tissue are expected to light up in DWI, because the the diffusion and hence the relative signal attenuation is lower in this location. In clinical scenarios the diffusion signal is interfered by non-linear effects involving the localization gradient fields of MRI [49]. Hence, in practice multiple measurements under varying gradient factors b (higher b-values indicate a larger time difference between the two pulses of the spin echo sequence and thus increase the diffusion component of the measurement) are preformed and subsequently fit by a biophysical model to extract the Apparent Diffusion Coefficient (ADC):

$$S(b) = S_0 \ \exp\left(-b \ \mathrm{ADC}\right) \tag{2.1}$$

Here, S_0 denotes the signal before diffusion pulsing and $S(b)$ the signal after diffusion pulsing depending on the gradient factor b. Note that on ADC maps the relation to diffusivity is inverted compared to raw b-value images, i.e. lower Brownian motion corresponds to lower intensity voxels.

While the conventional DWI model is based on the assumption that water diffusion follows a Gaussian behavior such that water molecules diffuse without any restriction, in living tissue, diffusion is commonly restricted by tissue microstructure and shows non-Gaussian phenomena. Recently, Jensen et al. proposed Diffusion Kurtosis Imaging (DKI) to account for this non-Gaussian component in diffusion behavior by extending the DWI model (Equation 2.1) by an Apparent Kurtosis Coefficient (AKC), which indicates the presence of diffusion restricting barriers and tissue heterogeneity [50, 51]:

$$S(b) = S_0 \ \exp\left(-b \ \mathrm{ADC} + \frac{1}{6} \ b^2 \mathrm{ADC}^2 \mathrm{AKC}\right) \tag{2.2}$$

Previous studies found that DKI substantially increases sensitivity compared to conventional DWI on gliomas [52, 53], hepatocellular carcinomas [54], prostate cancers [55, 56], and breast lesions [57]. While employing biophysical mdoels like in Equation 2.1 or 2.2 is the current clinical standard, Chapter 5 will present a model that integrates this process into a learning algorithm and provides evidence that learning signal exploitation on DWI yields superior performance on downstream tasks such as classification.

2.1.2. Computed Tomography

In Computed Tomography (CT), tissue is X-rayed from multiple angles to generate 3D images of local densities [43]. Compared to MRI, CT comes with high resolutions, short acquisition times, no constraint on molecular response such as to hydrogen, and broad availability of scanners. At the same time, patients absorb considerate doses of radiation throughout acquisition. The functionality of X-rays is based on emitting photons and counting them after traversal of matter. Reduction compared to the initial photon count can be related to absorption or scattering along the way and hence to the density of the radiated matter. CT scanners are built with the x-ray source and the detector rotating around the subject, such that measurements are successively acquired from many different angles. The line integrals resulting from X-ray absorption in each measurement can be inverted and analytically transformed to 3D Cartesian coordinates under certain assumptions on the scanner geometry. In contrast to MRI, the resulting 3D density map of the probe denotes a quantitative measure, i.e. intensities range on an absolute scale that is quantified in terms of Houndsfield Units (HU).

Positron Emission Tomography - Computed Tomography

Positron Emission Tomography (PET) is a nuclear medicine functional imaging technique, which depicts the spatial distribution of metabolic or biochemical activity in the body. Therefore, a radioactive tracer is injected into the patient's blood circulation prior to image acquisition. In oncology, tracers are designed to exhibit glucose-like properties, thus expected to be taken up by fast-growing cancerous cells with high energy consumption. As the radioisotopes of the tracer undergo positron emission decay, the emitted positron travels in tissue for a short distance until interaction with an electron. This encounter produces a pair of annihilation (gamma) photons moving in approximately opposite directions, which are subsequently detected by the PET scanner.

Positron Emission Tomography - Computed Tomography (PET-CT) combines a PET scanner and an x-ray computed tomography (CT) scanner in a single gantry, in order to acquire sequential images from both devices in the same session. PET-CT has revolutionized medical diagnosis in many fields, due to the complementary clinical information obtained from functional PET imaging, while ensuring spatial alignment to the highly resolved CT scan. Many diagnostic imaging procedures in oncology, surgical planning, radiation therapy and cancer staging have been changing rapidly under the influence of PET-CT availability, and centers have been gradually abandoning conventional PET devices and substituting them by PET-CTs. Drawbacks of PET-CT include high cost regarding the combined scanner as well as the radioactive tracer of PET in general [43].

2.2. The Diagnosis of Cancer

Cancer is the second leading cause of death worldwide [58] and represents an increasing burden under aging populations and higher life expectancies. Large scale studies throughout the last decades have shown that early detection is crucial, as it enables curative treatment and significantly improves patient outcomes [59]. However, this is not true for all cases. Many cancers are over-diagnosed, e.g. benign or non-invasive cancers that will not be clinically relevant for many years or even a patient's entire lifetime. In such cases, diagnosis leads to unnecessary costs for the patient such as anxieties or invasive follow-up procedures with associated discomfort, health risks, or impaired subsequent life quality as as well as to unnecessary financial costs and increased workload for the health care system. Under-diagnosis, i.e. delayed or missing detection of clinically relevant cancers, on the other hand, increases mortality and morbidity of patients. Thus, the careful tuning of diagnosis pipelines towards optimal patient outcome remains a challenging task.

This section depicts the chain of clinical procedures currently performed for diagnosis of breast cancer including the ongoing discussions on which diagnoses truly improve patient outcome. Breast cancer as the example of choice allows to provide background

information for the studies presented in Chapters 4 - 7, but also constitutes the most common cancer worldwide with 11.6% of all diagnosed cases (en par with lung cancer), which amounts to 1 in 8 women developing invasive breast cancer over the course of their lifetime [58]. Further, since breast cancer can arise at a relatively young age, it is the leading cause of death in women under 50 [60] and the second leading cause of death from cancer in women in general [58]. Moreover, early detection and thus a thoroughly designed diagnosis pipeline are imperative in breast cancer, as it increases the 5 year survival rate from 27% for distant cancer (already spread to other body parts when diagnosed) to 99% for cancer that is localized (limited to the original site) at the time of diagnosis [61]. Since breast cancer is 100 times more likely to occur in women than in men [58], the remainder of this section will focus on female breast cancer.

2.2.1. Breast Cancer Anatomy

As a mammary gland, the female breast is composed of differing layers of tissue. Glandular tissue is structured in multiple lobes, where the milk is produced and subsequently drained via lactiferous ducts to the nipple (see Figure 2.1 left). The lobes are held in place by surrounding fibrous tissue and the remaining space is filled with fatty tissue. The proportion between fibroglandular and fatty tissue is referred to as the density of the breast and constitutes a relevant factor for breast cancer increasing the risk by up to 4-5 times [62]. This is because cancer commonly begins inside the glandular tissue, i.e. in a lobe or a duct (see Figure 2.1). Since Lobular Carcinoma in Situ (LCIS) (in situ stands for "in its original place") are neither visible in imaging diagnostics nor clinically relevant, their detection is mostly incidental during biopsy of neighboring lesions. Ductal Carcinoma in Situ (DCIS), on the other hand, are visible in some imaging techniques but not yet clinically relevant, fueling controversial discussions about whether they should be detected by diagnostic procedures or not [60]. In situ cancers are not life-threatening, but commonly increase the risk of developing invasive cancer later on. Clinical relevance is given as soon as the cancer starts to invade neighboring tissue, which is referred to as Invasive Lobular Carcinoma (ILC) or Invasive Ductal Carcinoma (IDC). In the subsequent stages, cancer cells can break away from the original tumor in the breast and travel via the lymph system or blood system to other parts of the body, such as the liver, brain, bones, or lungs, where secondary tumors, also referred to as metastases, are developed.

2.2.2. Current Diagnosis Steps

Cancer Diagnosis is performed as a chain of subsequent procedures, where each step aims to resolve ambiguities of the previous step, while typically increasing the level of cost for the patient and the health care system. The traditional trajectory of breast cancer diagnosis is initiated by symptomatic cases, where e.g. a lump is detected in the breast during physical examination by the patient or a physician. Such suspicious findings are further examined by means of diagnostic mammography, a low-dose X-ray image acquisition, where

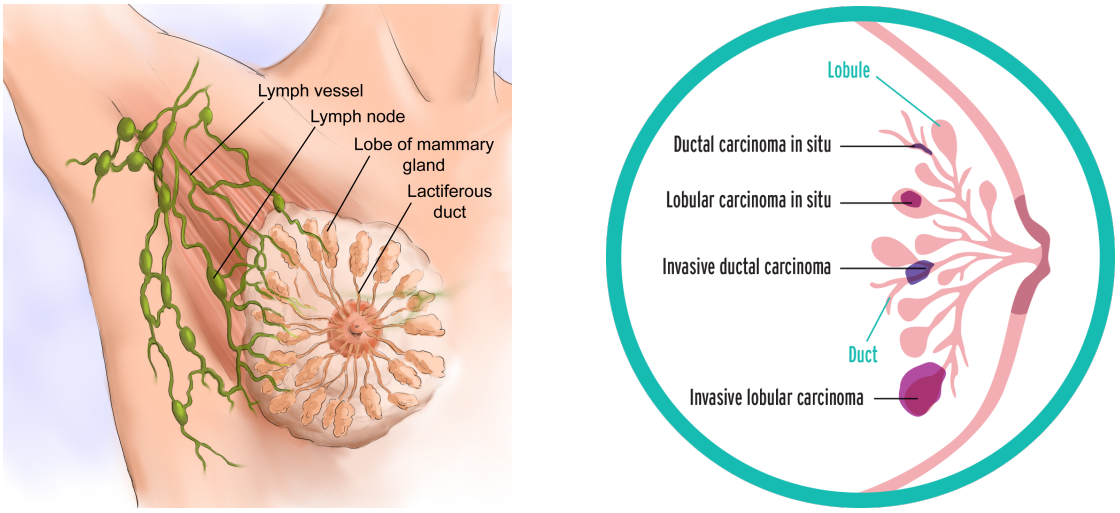

Figure 2.1.: **Breast Anatomy and Types of Breast Cancer.** Left: Glandular tissue is structured in multiple lobes, where the milk is produced and subsequently drained via lactiferous ducts to the nipple. The remaining space is filled with fatty tissue (bright). Image from [63]. Right: Cancer begins in either a lobe or a duct, where it can remain unnoticed for long times ("in situ"). If the cancer starts to invade neighboring tissue, it becomes clinically relevant. Image from [64].

the breast is compressed between two parallel plates to reduce the tissue thickness penetrated by X-rays, which is often experienced as painful by patients. The interpretation of mammographic images follows a standardized protocol, the Breast-Imaging Reporting and Data System (BI-RADS), which allows for concise and unambiguous understanding of patient records between multiple doctors and medical facilities [65]. The BI-RADS scale comprises the numerical codes 0-6 indicating the malignancy of the suspicious lesion (0: Incomplete evidence, 1: Negative, 2: Benign, 3: Probably benign, 4: Suspicious, 5: Highly suggestive of malignancy, 6: Known biopsy with proven malignancy (only occurs in follow-up mammography for treatment response assessment)). Patients with BI-RADS 0 or 3 are typically called for follow-up mammography, while patients with BI-RADS 4 or 5 are sent to biopsy, which is typically performed as a minimal-invasive procedure, where a sample of tissue or fluid is extracted from the suspicious area by means of a hollow needle and histopathologically examined under a microscope for accurate assessment of malignancy. This ultimate stage of the diagnosis chain is associated with high discomfort and anxiety among patients [66].

The traditional "symptomatic" diagnosis setup has several flaws, one of them being that clinically relevant cancer is often visible on radiological images before symptoms are developed (referred to as "lead time") [60]. Since pre-symptomatic detection has been shown to win precious time for curative treatments and improve patient outcomes, many developed

nations have implemented large-scale mammography screening programs [67, 68], where every women above the age of 40 or 50 receives X-ray mammography in intervals of one or two year (depending on country) aiming for a presumptive identification of unrecognized disease in an apparently healthy and asymptomatic population. While randomized controlled clinical trials provide evidence that mammographic screening does reduce mortality from breast cancer [69, 70, 71], the resulting diagnosis chain still suffers from high amounts of underdiagnosed and overdiagnosed cases [72]. Both can be related to the sensitivity profile of X-ray mammography, i.e. the distribution of sensitivity levels across different tumor biologies, which is not fully suitable to the clinical task of breast cancer detection [60]: While an ideal diagnosis procedure exhibits high sensitivity in clinically relevant, aggressive and high-grade cancer, and low-sensitivity in clinically irrelevant, benign, or low-grade lesions (diagnosis of low-grade disease may not contribute to a true survival benefit [73]), the sensitivity profile of X-ray mammography appears to behave inverse. One explanation is that mammography relies on the depiction of architectural and morphological features that reflect pathophysiological processes associated with slow growth and in turn with cancers of often limited clinical relevance such as DCIS [60]. The resulting overdiagnosis manifests in 50% of mammographic findings being identified as benign by subsequent biopsies, revealing half of the performed biopsies as unnecessary [74, 75, 76] and indicating a tremendous burden for patient life quality as well as the health care system [77]. At the same time, aggressively growing and invasive cancer exhibits imaging phenotypes such as roundish shapes or smooth borders rendering them often indistinguishable from ubiquitous benign lesions in X-ray mammography [60]. The resulting underdiagnosis is further amplified by the fact that dense breast tissue considerably reduces sensitivity in X-ray mammography [78], and constitutes the prevalent reason behind breast cancer remaining a major cause of death for women [79]. While various attempts are made at improving the interpretation of X-ray mammograms with deep learning algorithms [15, 80, 81], it is an ongoing discussion whether the flaws associated with this imaging technique can be compensated for, or whether the future of breast cancer diagnosis might require a fundamental shift towards other imaging modalities.

2.2.3. MRI Screening - The Future of Breast Cancer Diagnosis?

Due to lack of availability, long acquisition time and high associated cost, MRI is currently not used for broad screening of the population, but reserved for a small group of high risk patients. Further, MRI is performed at later stages of the diagnostic chain, such as for resolving ambiguities of previous findings or assessing treatment response. The commonly used MRI sequence for this task is Contrast-enhanced T1-weighted MRI. Despite the lack of large-scale randomized clinical trials of MRI screening, studies indicate that MRI is able to double or even triple sensitivity compared to mammography independent of risk factor [82, 83, 84]. Specifically, in contrast to X-ray mammography the sensitivity profile of MRI appears to suit the task of breast cancer detection, with high sensitivities for aggressive

and invasive high-grade cancer and low sensitivities for low-grade disease. This can be explained by the unsurpassed soft-tissue contrast in MRI, the fact that breast density does not affect image evidence such as in mammography, as well as the correlation between angiogenic activity (which drives the take-up of contrast agent) and cancer growth [60]. The reduced sensitivity of low-grade DCIS has been used as an argument against MRI as a standalone screening technique, but recent understanding of DCIS related implications on overdiagnosis, i.e. the assumption that the non-detection of such low-grade DCIS might be clinically desirable, might call for reconsideration of this argument [73].

Efforts to improve availability and bring down costs to establish MRI as a frontline tool in breast cancer diagnosis range from tailoring abbreviated sequencing protocols to designing new scanners tuned for high patient throughput acquisition [60]. Arguably, the most promising recent development towards a widely applicable, light-weight and highly accurate MRI technique for breast cancer diagnosis is the acquisition of DWI (see Section 2.1.1). This method refrains from administration of contrast agent, which is an invasive procedure associated with health risks (as elaborated on in Section 2.1.1) and thus suffers from diminished acceptance among the population. Moreover, DWI is able to bring down acquisition time to around 7 minutes [85]. First studies have shown that measuring diffusion is superior in assessing the malignancy of breast lesions compared to mammography and as a consequence enables drastic reduction of false positive findings [86, 85, 87]. An explanation is that diffusivity is correlated to integrity of cellular structure, which in turn is related to speed of cellular growth. Such micro-scale information is unaccessible for X-ray imaging, which describes density of cell patterns on a macro-scale (see Section 2.1.2). Moreover, DWI might be able to reduce false positive findings compared to Contrast-enhanced MRI [88, 89]. Chapters 4-7 present studies on how learning algorithms can automate and improve breast cancer diagnosis based on this emerging imaging technique.

In summary, breast cancer diagnosis is a fast developing field of research with heterogeneous opinions and vibrant discussions on defining the right steps towards improved patient outcome. One concept that large parts of the community seem to agree on is the importance of "risk stratification", i.e. the step away from "one-fits-all" screening programs towards personalized diagnostic procedures by means of prior risk assessment of individual patients in terms of e.g. family history, gene sequencing or breast density. This way, resources of current health care systems could be thoughtfully allocated to where they are needed most: For patients with low-risk and low breast density X-ray mammography is expected to suffice, patients with intermediate risk or high breast density could undergo light-weight MRI such as DWI, while high-risk patients could be treated by means of the entire palette of diagnostic tools. This transformation ties in nicely with the concept of "precision medicine" (see Chapter 1) as a vision for the future of digitalized health care.

3. State of the Art in Medical Image Classification

Classification of Images, i.e. their interpretation and allocation to semantic categories, roughly involves three tasks:

1. Localization of discriminative regions (or Regions of Interest (ROIs)) in the image, i.e. areas containing evidence related to the task at hand such as objects or image patterns.
2. Delineation of ROIs in terms of measurements, effectively transforming spatial information regarding precise image localization into semantic information in form of quantized features.
3. Application of decision thresholds in the mono- or multivariate features space aiming to assign the underlying image to predefined semantic categories.

In the medical domain, the three tasks have traditionally been projected into single components of engineered, and task-specific classification pipelines involving high amounts of domain knowledge such as feature sets tailored towards application in different parts of the human body. Since then, the paradigm of end-to-end learning and the implied decrease of relevant domain knowledge has lead to a drastic generalization of methodologies, which in turn has caused a shift of research focus away from the engineering of task-specific systems towards the successful adaptation of existing deep learning methods. As a consequence, the methodological State of the Art (SotA) in Medical Image Classification has become increasingly aligned with the corresponding SotA in deep learning research. This Section will provide a short overview of current SotA algorithms relevant to Medical Image Classification while sensitizing the reader to two major challenges that currently prevent end-to-end learning systems from clinical application: Model robustness under limited training data and the clinical relevance of evaluation scales.

3.1. Medical Image Classification based on Handcrafted Features

Before the era of end-to-end learning systems, conventional machine-learning techniques were limited in their ability to process natural data in their raw form essentially only covering the last of the three subtasks involved in image classification outlined above. Thus, for decades, constructing a pattern-recognition system required careful engineering and considerable domain expertise to design an upstream feature extractor that transformed the raw data (such as the pixel values of an image) into a suitable internal representation or feature vector, from which downstream learning system could detect or classify patterns in the input. This section will introduce Radiomics, an effort from the medical domain in the early 2010s to get away from engineering features highly tailored to specific tasks and towards standardized large-scale measurements, as well as two types of machine learning classifiers relevant for the scope of this thesis: Random Forests and Neural Networks.

Radiomics Feature Extraction As briefly mentioned in Chapter 1, Radiomics describes the comprehensive understanding and quantification of information in radiological images and ties into an overarching data pool of a patient's quantized clinical information with the ultimate goal to improve patient care by enabling personalized treatment [8]. Thereby, Radiomics comprises large-scale extraction of features from images regarding intensity statistics, shape, or texture [9, 10, 11]. Radiomics studies have been able to reveal insightful correlations between imaging biomarkers and underlying clinical or genetic profiling e.g. for categorization of cancer [90], early metastases [91] or protein expression [92]. Due to the fact that Radiomic features describe local image patterns such as associated with single objects, their extraction commonly requires upstream image processing like localization of Regions of Interest (ROIs) in form of manual annotation. Chapter 4 will showcase successful applications of Radiomics features combined with machine learning classifiers for breast cancer classification on DWI and cardiac diagnosis on MRI.

Random Forest Random forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees [93]. A trained decision tree constitutes a combination of if-then statements, which are expressed as nodes in a tree-like structure, where each node represents a linear cut on a single variable. New inputs are classified by propagation through the tree until reaching an end node ("leaf node"), which has been assigned to a certain class during training and thus represents the classification decision. The tree structure is built during training, where new nodes are successively generated with associated rules by optimizing for class separation until a stopping criterion is met and the current node is defined as a leaf node.

Neural Network Feedforward Neural Networks, also referred to as Multilayer Perceptron (MLP)s when implemented as shallow classifiers, are built of processing units ("neurons") each representing a nonlinear transformation by means of internal parameters ("weights"). Neurons are arrayed parallely, i.e. several independent neurons processing the same input

vector, forming a network layer, while a neural network is composed of multiple layers stacked on top of each other. In MLPs, the stacking of layers is implemented in a "fully connected" fashion, i.e. each of a layer's neurons has a connection (i.e. network weight) to each of the previous layer's neurons. In classification tasks, the ultimate network layer is commonly modeled as a categorical softmax distribution over class targets Y, which allows to describe the neural network as a conditional probability distribution $P(Y|X, W)$ depending on some input vector X and the weights W. The network weights are iteratively updated in a supervised fashion, i.e. by sampling mini-batches of input-target-tuples (X, Y) from the training data distribution while minimizing some loss function [29].

3.2. Medical Image Classification based on learned Representations

Bias Variance Trade-off The advantages of substituting handcrafted feature extraction with data-driven learning of representations are manifold: Integrating feature extraction and classification into a single neural network architecture allows for optimizing both system components simultaneously, based on task-specific observations, and with respect to the ultimate target. Moreover, learning representations enables dynamic adaptation of context scale, i.e. is not restricted to local image patterns such as Radiomics features, but allows to encode global dependencies. While Neural Networks are universal function approximators [94], i.e. in theory able to learn arbitrarily complex relations between high dimensional inputs and associated targets, the ability of original MLP architectures to process raw data such as images is limited mainly due to a phenomenon referred to as the *bias-variance-tradeoff* [95]. The inductive bias of a model is the set of assumptions driving a model's predictions on unseen data, where erroneous assumptions may lead to so-called *bias error*. In order to achieve the ability to generalize to unseen data, this bias error needs to be counterbalanced with a model's variance, i.e. the error from modelling noise in the training data rather than underlying patterns with causal relations to the target ("overfitting"), such as in over-parametrized models. The tradeoff between the two errors is commonly controlled by model complexity, i.e. the number of model parameters [1] , but depends on further external factors such as the amount and diversity of training data and the general complexity of the task (such as the level of target-related evidence in the input). As a consequence, if an MLP architecture were to be scaled up to operate on high-dimensional inputs such as raw images, the resulting increase in model parameters (typically by several orders of magnitude) would render the model prone to overfitting and in return hamber the ability to generalize to unseen data.

Convolutional Neural Networks The watershed for representation learning in image

[1] There is recent evidence suggesting the definition of model complexity might be more complex e.g. excludes certain topological factors of the network architecture [96].

classification by means of neural networks was the introduction of Convolutional Neural Network (CNN)s in the late 1980s [97], which derive their success from drastically reducing the amount of network parameters compared to MLPs. This is achieved by condensing domain knowledge regarding translational invariance of objects in images into an informed inductive bias, i.e. one that does not increase bias error while reducing model variance. The additional bias effectively constraints the space of possible solutions during generalization, which is implemented by replacing the fully connected layering of MLPs with convolutional filters, i.e. local operators that successively scan over the image, enabling to learn the detection of edges, motifs, or objects in one set of weights shared across spatial locations. This economical dealing with model parameters allows for stacking of large numbers of layers in a network without overshooting the model's variance, which eventually coined the term deep learning. CNNs for whole image classification are implemented as encoding architectures, where convolutional layers are intermitted by pooling layers, which reduce the spatial size of the array of filter outputs ("feature maps"), effectively trading in spatial precision for semantic richness in learned representations. This effect is amplified by the fact that the "effective receptive field", i.e. the area of input pixels being connected to a single neuron via consecutive convolution operations, increases throughout the network architecture allowing for capturing higher-level abstractions. The concept of learning hierarchical representations has proven very effective in images, because it exploits the property that evidence in images often comes in compositional hierarchies, in which higher-level features are obtained by composing lower-level ones, e.g. local combinations of edges form motifs, motifs assemble into parts, and parts form objects. The delayed success of CNNs for natural image recognition tasks starting with AlexNet in 2012 [98] and later VGG in 2015 [99] is attributed to insufficient compute and data resources at the time of their introduction.

Medical Application While CNN encoders as described above technically allow for implicit localization of ROIs and thus can be trained in end-to-end settings, i.e. operating on the raw images without previous annotations (see Section 3.3 for SotA and examples), this scenario is sometimes not feasible in the medical domain: Assigning the training signals derived from scalar class targets to high dimensional inputs becomes a highly ambiguous task, if the ROIs containing task-related evidence comprise only a small fraction of this input. This issue is commonly amplified in the medical domain by notoriously limited training data, and ultimately leads to high model variance and poor generalization ability. In contrast, datasets of classification tasks on natural images such as ImageNet [98] often depict a single centric object of considerable size, rendering implicit localization of ROIs feasible. In compensation for the described challenges in the medical domain, there exists an intermediate step in medical image classification aiming to make use of learned representations while still relying on previous detection of ROIs (by clinicians or upstream detection systems), by applying CNNs on cropped patches en-

compassing these ROIs. This setup is also referred to "false positive reduction" (bearing on previously annotated ROI candidates) and is often performed in tasks involving small ROIs in large images such as classification of lung nodules in chest CTs [100, 101] or breast lesions in mammography [102]. One focus of this thesis is to study the potential of end-to-end learning in clinical diagnosis systems by starting with compound classification pipelines as applied in clinical standard and successively substituting components for learning algorithms ultimately deploying end-to-end models. Thereby, Chapter 5 showcases the intermediate step of ROI-based false positive reduction, which we will refer to as "Roi-to-end" in the remainder of this thesis, so as to distinguish it from true end-to-end training on raw images, on breast cancer classification in DWI, where associated lesions (i.e. ROIs) exhibit particularly small sizes, since images were acquired as part of an early detection screening program.

3.3. End-to-end Medical Image Classification

End-to-end learning in image classification describes the integration of all subtasks, i.e. localization of ROIs, extraction of semantic features, and decision making, into one single learning algorithm and training all components simultaneously, directly on the raw data, and with respect to the ultimate target. This is a highly desirable scenario for the medical domain, since it refrains from any task-specific engineering and alleviates the workload of clinicians by automating the process of annotation of ROIs. While CNN encoders (see Section 3.2) model the localization of ROIs implicitly, there exist further end-to-end learning algorithms tailored to solving this task explicitly by employing spatially resolved class targets and generating spatially resolved predictions. Thereby, the spatial resolution of annotations and predictions comes at different *granularity levels* which implies an unfolding of the classification task onto varying scales: Encoding architectures reduce the high-dimensional image input to a single distribution over target class probabilities (as elaborated on above), i.e. assign one single category to an entire image, hence address the classification task on image level (or *patient level* in medical terms). Object detection models, on the other hand, output multiple predictions per image on *object level*, each composed of coordinates and per-object class distributions. Finally, semantic segmentation models output class distributions on *pixel-level* , i.e. in form of segmentation maps, essentially categorizing each individual pixel. Accurate differentiation of these levels is highly important in medical image classification, because each addresses different tasks and comes with specific evaluation metrics essentially counting correctly and incorrectly classified instances of the respective granularity level, i.e. patients, objects, or pixels, respectively. Thus, the granularity level of a model's output predictions directly answers to a specific clinical questions (while the inverse mapping from clinical task to suitable granularity level of classification is less straightforward due to several nuisance factors discussed in Section 3.4). This section introduces the algorithmic SotA associated

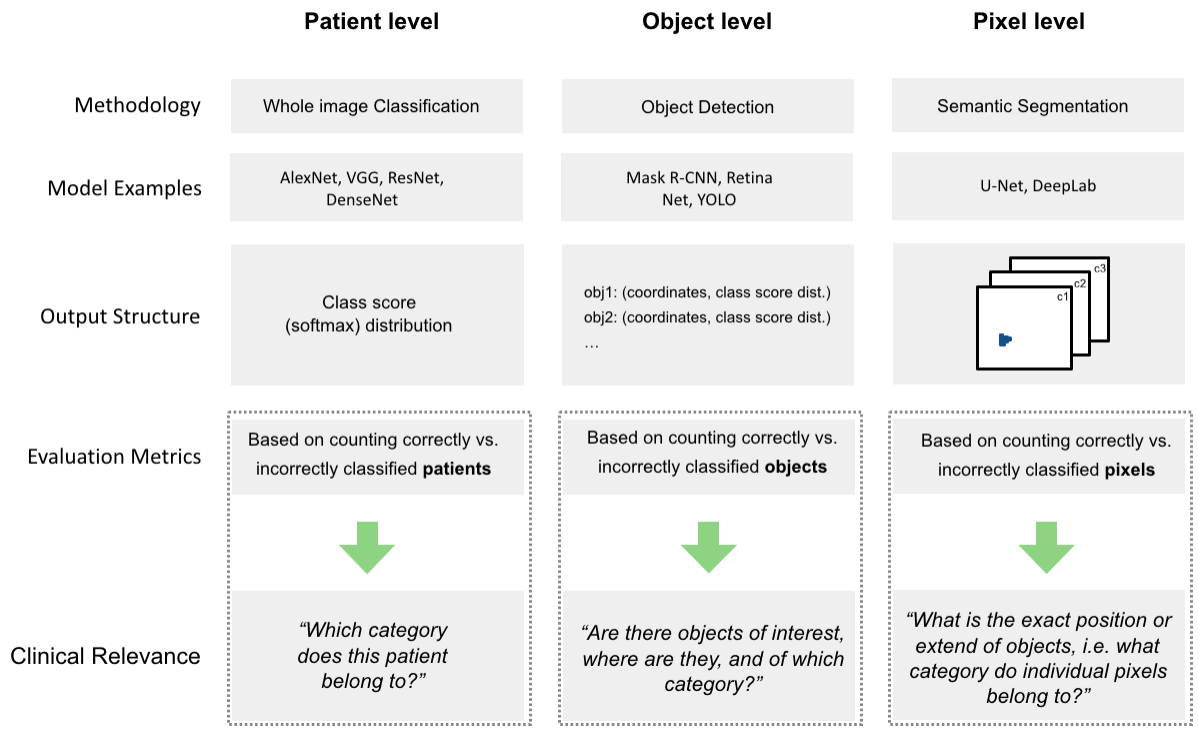

Figure 3.1.: **Granularity Levels in Medical Image Classification.** This figure visualizes the three levels of granularity in classification of medical images, the associated computer vision methodology, model examples, and output structure. Each scale comes with specific evaluation metrics, that determine which clinical question is addressed. References for model examples: AlexNet [98], VGG [99], ResNet [103], DenseNet [104], Mask R-CNN [105], Retina Net [106], YOLO [107], U-Net [108], DeepLab [109].

with different levels of granularity in image classification as well as associated clinical applications. An introduction to evaluation metrics in image classification is provided in Appendix A.

Patient Level: Whole Image Classification

Image classification on patient-level, i.e. the task of assigning a single class label to an entire image, is commonly modeled by encoding architectures as described in Section 3.2. While the fundamental architecture of CNN encoders has not changed since the introduction in the late 1980s [97], research since focused on controlling gradient instabilities observed when stacking many layers by means of *residual connections* [103] or *dense connections* [104]. Additionally, feature map normalization schemes such as batch-norm or instance-norm have demonstrated considerable improvements [110, 111].

Patient-level classification is required for any clinical question that is ultimately interested in whether a patient is healthy or not, and in case of the latter which type of clinical condition is present. This includes for instance the diagnosis of primary cancer [70, 112, 113]. As described in Section 3.2, the implicit modeling of ROI localization is hampered in practice due to severe credit assignment problems of scalar training labels in high-dimensional inputs ("which part of the input image should the scalar training signal be assigned to during backpropagation?"), which renders models prone to overfitting. Nevertheless, recent studies have reported human-level performance of such models in clinical tasks, while overcoming the credit assignment issue by either training on enormous datasets [18] (roughly 760000 chest radiographs), operating on images with high task-related evidence, i.e. comprising one centric ROI of considerable size such as in dermoscopy [16], or enhancing the training signal by employing annotations of higher spatial resolution [15, 114, 115]. While employing annotations of higher levels of granularity commonly requires subsequent aggregation of model predictions in order to evaluate on the clinically relevant level, Chapter 6 will present a solution that enables end-to-end training on multiple levels thus removing the need for aggregation.

Pixel Level: Semantic Segmentation

In contrast to encoding architectures, pixel-level predictions require models with spatially resolved output. To this end, the first Fully Convolutional Networks replaced the densely connected classification component at the end of SotA encoding architectures at the time such as AlexNet [98] or VGG [116] by 1×1 convolutions enabling subsequent upsampling of feature maps [117]. These first attempts focused on exploiting existing encoders pre-trained on large natural image datasets and neglected the upsampling part of the model, predicting at resolutions 32 times smaller than the input thus resulting in highly asymmetric architectures. The current SotA in biomedical image segmentation is the U-Net [108] architecture, which was presented shortly after and introduced a principled way of combining the high-level semantic information of later layers with the fine-grained localization of earlier layers. Instead of generating spatially resolved output by means of ad-hoc upsampling operations, the U-Net adds a decoder part to the network, which successively combines coarser features with more localized features in convolutional layers, i.e. in a *learned fashion*, essentially mirroring the encoder part of the network. The U-Net is complemented by *skip connections*, which shortcut signals of each resolution stage in the encoder to the corresponding resolution stage in the decoder, allowing to successively recombine features of different scales. The resulting encoder-decoder architecture generates otuput at full spatial resolution with respect to the input.

Pixel-level predictions are relevant, where the exact extend or shape of structures is of interest such as in radiotherapy planning [118], intra-operative support [119], tumor growth monitoring [120]. Moreover, segmentation maps are often used as intermediate represen-

tations for subsequent measurements such as in cardiac diagnosis [34] or diagnosis (see Section 4.1.2) or retinal disease [121].

Object Level: Object Detection

While prediction at full resolution as performed by segmentation models might seem as the richest possible output representation, it comes with several flaws: Models operating solely on pixels have no notion about what constitutes an object or different instances of classes in an image. As a consequence, any such computer vision task interested in classification of multiple objects in an image, commonly referred to as object-detection tasks, would require aggregation of pixel predictions in ad-hoc postprocessing steps. Perhaps unsurprisingly, the respective field of research converged to architectures similar to the U-Net, the Feature Pyramid Network (FPN), where instead of upsampling predictions to full resolution, coarser representations at earlier stages of the decoder commonly associated with motifs of higher abstraction such as objects are extracted for subsequent detection and classification. The current SotA in object detection is roughly divided into two groups, which differ in how the detection and classification task based on the FPN features is implemented: So-called two-stage detectors first discriminate objects from background ("class agnostic") while simultaneously regressing bounding box coordinates by means of a Region Proposal Network (RPN) [105, 122, 123]. Subsequently, proposals are categorized after resampling them to a fixed grid ("RoiAlign"), thus ensuring scale-invariance for categorization. One-stage detectors, on the other hand, have been proposed to perform class-aware categorization directly inside the RPN [106, 107, 124]. In general, prediction of multiple objects per image is implemented by generating predictions at all spatial locations of the feature map in a fully convolutional fashion, i.e. by scanning the RPN over the feature map while sharing weights. The excessive amount of resulting predictions is reduced in a subsequent filtering process referred to as Non-maximum Surpression (NMS), where predictions are clustered according to IoU and all but the prediction with highest confidence score per cluster are discarded. Since the scales of objects may vary drastically, i.e. range from few pixels to areas covering most parts of the image, the RPN generates predictions based on multiple resolution stages (or "pyramid levels") of the FPN while sharing weights between stages. Moreover, additional guidance in form of priors on the predicted box coordinates ("anchor boxes") has shown empirical improvements. Thereby, multiple predictions per spatial locations corresponding to different anchor boxes are generated while not sharing weights, essentially aiming at training different classifier experts for varying box ratios. These anchor boxes are implemented by not predicting the coordinates of ground truth boxes directly, but instead modifying the coordinate targets during training to represent the delta between the target coordinates and the most overlapping prior box. If additional per-instance pixelwise predictions are generated and evaluated, the task is referred to as *instance segmentation*. This additional output can e.g. be achieved by implementing an additional head network on the extracted

FPN features trained to predict pixel-maps per object proposed by the RPN, such as in Mask R-CNN [105].

Object level classification is clinically relevant, when studying a disease in more detail such as the spread or stage of cancer, i.e. when the clinical interest lies within the existence, rough position and category of all individual objects of interest such as in delineation of multiple sclerosis lesions [125], lung cancer staging [126], or multiple myeloma staging [127]. Object-level outputs are further generated as a proxy for patient-level tasks, when hardware constraints prevent end-to-end patient-level training (detailed discussion in Section 6.7).

3.4. Discussion

While in rule-based decision making, prior knowledge is condensed into predefined constraints that serve as the model's inductive bias, i.e. the set of assumptions used to generate predictions on unseen data, in learning algorithms task-specific inductive biases are captured by training on observations [31] [2]. Thus, a learning algorithm's ability to generalize to unseen data scales with the amount and quality of training data. This is a great advantage, because, in contrast to rule-based systems, performance is not sealed by the complexity of decision boundaries. However, if the data available at training time denotes a poor representation of the underlying data distribution, such as commonly occurring in medical imaging tasks, the inductive biases resulting from training do not enable the model to robustly generalize to unseen data. End-to-end learning algorithms, where an entire data processing pipeline is learned mostly from observations, carries the described data dependency to extremes. This challenge is amplified in the medical domain by the fact that medical datasets are notoriously small and suffer from distribution shifts due to varying scanner types, scanner sequences, scanner protocols or ambiguous annotations, and renders model robustness towards unseen data characteristics the prevalent obstacle in the strive for clinical application [27, 30, 25, 22]. To this end, the work presented in Chapters 4-6 will examine the effects when successively replacing rule-based components of clinical classification pipelines by learning algorithms, i.e. benchmark performances and draw conclusion about the potential of learning algorithms under data constraints. Moreover, the idea behind CNNs to condense prior knowledge about the domain into inductive biases in order to reduce model variance and in return improve the generalization ability of deep learning models will serve as blueprint for the work presented in Chapters 6 and 7. Thereby, we move beyond mere architectural modifications and explores leverage points in all components representing a model's inductive bias: Training data, architecture, loss function, and optimization [128]. We introduce new constraints regarding architecture

[2] The inductive biases derived from training are to be considered as an addition to initial inductive bias given by the model architecture such as in Convolutional Neural Network (CNN)s (see Section 3.2)

in Section 7.2, regarding loss function in Chapter 6 and Section 7.1, and finally touch upon all four components by systematizing the process of hyperparameter configuration in Section 7.3.

Another key obstacle preventing end-to-end learning systems from clinical application is the discrepancy between model evaluation in research environments and actual clinical requirements [27, 129, 130]. While Figure 3.1 might convey the impression that the choice of granularity level for a clinical task is straightforward, there are in fact several nuisance factors, that may introduce a discrepancy between the researchers objectives and the clinical question to be addressed, such as hardware constraints or vast differences in training efficiency associated to annotations at different levels. Chapter 6 will address this challenge and present a solution that maintains data efficient training while accounting for clinical relevance of evaluation scale.

4. Medical Image Classification based on Handcrafted Features

The current clinical standard of computer-aided image classification is mostly embodied by engineering task-specific compound pipelines. We study the potential of end-to-end learning for clinical diagnosis by starting with current practice systems and gradually substituting single pipeline components for learning algorithms while carefully tracking the effects. As a first step, we attend to the last part of the classification pipeline, i.e. the categorization of entities into clinical pathologies based on handcrafted features. Specifically, we explore the potential of replacing univariate rule-based decisions with machine learning classifiers. The main contributions in this chapter are:

- We achieve human-level performance in breast cancer lesion classification on DWI by substituting a mono-parametric decision threshold with large scale Radiomics feature extraction followed by a machine learning classifier [32].
- We provide insights about further parts of the diagnosis pipeline such as an in-depth analysis of Radiomics features or the exploitation of domain knowledge to enhance the image normalization model.
- We apply the approach of large scale feature extraction followed by machine learning classifiers to cardiac disease classification on MRI, thereby extending the feature set by a novel group of time-series measurements over the cardiac cycle, and achieve the second rank in an international competition [33, 34].

As kindly permitted by the Radiological Society for North America (RSNA), the sections of this chapter regarding breast lesion classification reproduce parts of the following publication:

Bickelhaupt, S.*, Jaeger, P. F.*, Laun, F. B., Lederer, W., Daniel, H., Kuder, T. A., Wuesthof L., Paech D., Bonekamp D., Radbruch A., Delorme S., Schlemmer H. P., Steudle F. H., Maier-Hein, K. H. (2018). Radiomics Based on Adapted Diffusion Kurtosis Imaging Helps to Clarify Most Mammographic Findings Suspicious for Cancer. Radiology, 287(3), 761–770. * equal contribution

The study on cardiac diagnosis was submitted to an international competition [34] and further published in the associated proceedings [33].

4.1. Problem Statements

4.1.1. Breast Lesion Classification on DWI

As elaborated on in Section 2.2.2, current breast cancer diagnosis and large-scale mammography screening in particular suffer from high amounts of false positive findings: Around 50% of women who are sent to biopsy are over-diagnosed, i.e. sent based on a false positive mammography finding [74, 75, 76]. The invasive procedure of biopsy and the associated waiting time for results constitute a significant mental and physiological burden for these patients. While parts of the breast cancer radiology community suggest to improve breast cancer diagnosis by shifting towards MRI as a standalone screening method replacing mammography [60], such considerations seem ambitious in context of current availability of MRI and the pace at which fundamental transformations in health care systems typically proceed. In this context, integration of additional lightweight and noninvasive imaging such as DWI (see Section 2.1.1 into the current diagnositc chain has been proposed as a reasonable first step towards widespread MRI-based breast cancer diagnosis [131]. In this scenario, women with suspicious lesions (BI-RADS 4 or 5, see Section 2.2) are sent to additional DWI acquisition aiming to filter out false positive findings before sending patients to biopsy.

In current clinical standard, malignancy of suspicious lesions on Diffusion Weighted Magentic Resonance Imaging (DWI) is assessed by first manually annotating the ROIs, subsequently fitting a biophysical model to each individual voxel to extract the Apparent Diffusion Coefficient (ADC) or more recently the Apparent Kurtosis Coefficient (AKC) (see Section 2.1.1), computing the median coefficients per ROI over voxels, and finally applying simple cut-based thresholds on those [132]. On the other hand, there is a growing research community applying Radiomics to extract supplementary computational information on the imaged tissue by using either MR imaging or conventional x-ray mammography [133, 134, 135, 136, 9]. Previous Radiomics studies mainly were focused on revealing correlations between a radiomics signature and the underlying clinical or genetic profiling (eg, OncotypeDX categories [90], early metastases [91] and protein expression [92]), but not used for clinical diagnosis directly.

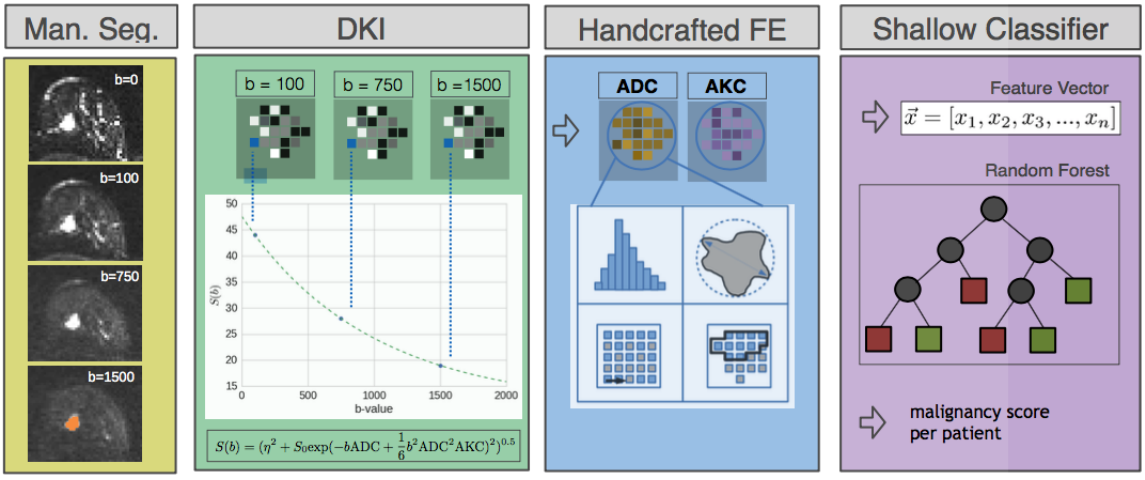

Figure 4.1.: **Proposed Radiomics Pipeline for Clarification of Breast Lesions suspicious for Cancer based on Diffusion weighted Imaging.** First, a DKI model is fit to the b-values of individuals pixels in the previously annotated ROI. The resulting ADC and AKC coefficient maps are the basis for large-scale Radiomics feature extraction. Finally, a random forest performs multivariate classification to reduce the dimension of the feature vector into a binary decision regarding the malignancy of lesions.

In this study, we propose to enhance the current clinical research practice in 3 steps (See Figure 4.1):

1. Extend the DKI model by a vector $\theta(b)$ as a patient specific calibration utilizing a separatly annotated fat area.
2. Replace the simple median aggregation of voxel coefficients by extracting a set of Radiomics features from the coefficient maps based on first order statistics, shape and texture.
3. Train a Random forest for classification based on extracted features to account for multivariate dependencies between Radiomics features.

4.1.2. Cardiac Disease Classification on MRI

We further applied the approach of Radiomics feature extraction followed by machine learning classifiers to a public competition on cardiac diagnosis, the Automated Cardiag Diagnosis Challenge (ACDC) [34]. The goal of this two-part challenge, where first three cardiac structures had to be automatically segmented from MRI and secondly patients had to be categorized into 5 disease categories, is to improve upon the clinical practice of cardiac diagnosis. In current practice, first cardiac structures are segmented by applying

semi-automatic systems followed by rule-based decision making based on shape-related measurements of segmented structures [137, 138, 139]. While it was theoretically possible to address the challenge using end-to-end classification by means of a CNN that directly maps raw input images to pathologies, the challenge dataset clearly did not provide sufficient training data (100 training cases, 20 per disease category) for this approach. As described in Section 3.2, especially the mapping of high dimensional images to scalar values is suffering from weak training signals, i.e. a credit assignment problem that renders the model prone to overfitting. Hence we decided against end-to-end learning and opted for a domain knowledge-guided approach, which we adjudged feasible given the encountered data burden. Specifically, we applied an ensemble of UNet inspired architectures [108] (see Section 3.3) for segmentation of cardiac structures including the Left Ventricular Cavity (LVC), Right Ventricular Cavity (RVC) and Left Ventricular Myocardium (LVM) on each time instance of the cardiac cycle. Subsequently, information was extracted from the segmented time-series in form of comprehensive features handcrafted to reflect diagnostic clinical procedures. Thereby, we extended the set of standard Radiomics features by introducing additional measurements describing volume behavior over time. Based on all extracted features we trained an ensemble of heavily regularized Multilayer Perceptrons (MLP) and a Random Forest Classifier (see Section 3.1) to predict the pathologic target class. Figure 4.2 shows an overview of the proposed pipeline [33].

Since the ground truth of cardiac pathologies in this dataset is determined by applying rule-based cuts on shape features based on the manual annotations of cardiac structures, benchmarking our machine-learning based classification against corresponding rule-based approaches does not allow for comparison between them, but merely gives indication of robustness against variations in segmentations (deviations of rule-based predictions from the ground truth could be attributed to deviations of underlying predicted segmentations from the original manual ground truth segmentations). We generally expect increased robustness against variations in segmentations when applying machine-learning classifiers over rule-based decision making, since the latter is commonly based on merely one or two features and thus more prone to be affected by segmentation bias [34]. Further, learning-based decision making can easily be "calibrated" to a new dataset (which often come with deviating annotation styles) by re-training, while explicit rules are considered global domain knowledge and therefore require careful justification when altered. However, as argued above such studies involve quality assurance of automated segmentation, and thus go beyond the scope of this thesis. As a consequence, this study intents to serve as a vivid showcase for a successful machine learning-based diagnosis pipeline (we achieved the first rank in the segmentation task and the second rank in the classification task) rather than to draw scientific conclusions regarding the superiority over rule-based decision making.

Clinical Motivation and Related Work

Analysis of cardiac function plays an important role in clinical cardiology for patient management, disease diagnosis, risk evaluation, and therapy decision [137, 140, 138]. The progressive course of heart failure can be associated with cardiac remodeling, which results in poor prognosis for the patient due to diminished contractile systolic function, reduced stroke volume or malignant arrhythmia. Clinical manifestations comprise changes in size, mass, geometry, regional wall motion and function of the heart [141], which can be assessed temporally and monitored non-invasively by cardiac magnetic resonance imaging (CMRI). In today's clinical routine, the huge benefits of comprehensive quantitative measurements are still not exploited due to the associated labour time, subjective biases and lack of reproducibility. Accurate automatic approaches for simultaneous multi-structure segmentation and CAD are thus desirable assets for a large spectrum of cardiac diseases. CAD approaches originate from the field of lesion detection and classification [142], thus primarily used to focus on texture information to discriminate healthy from pathological tissue. Medrano-Gracia et al. investigated global shape variations of the left ventricle in a large cohort of an asymptomatic population [143]. They found the major principal modes of shape variation to be associated with known clinical indices of adverse re-modelling, including heart size, sphericity and concentricity. Later, Zhang et al. used a supervised method to extract the most discriminatory global shape changes associated with remodeling after myocardial infarction [144]. The resulting shape model was able to discriminate patients from asymptomatic subjects with 95% accuracy. However, to the best of our knowledge, a comprehensive CAD system for different cardiac re-modelling pathologies and myocardial infarct patients has not been proposed before.

4.2. Utilized Datasets

Breast Lesion Classification on DWI

To the best of our knowledge, the dataset utilized in this study constitutes the largest of its kind at the time of publication. DWI data of 222 consecutive patients examined in two study centers in 2014–2016 were included in this analysis [32, 85, 132, 145, 146]. 95 patients were examined at the first study site and included for method development and training (mean age: 58.6 years ±6.6; 61 malignant and 34 benign). 127 patients from the second study site were included in the independent test set (mean age: 58.2 years ±6.8; 61 malignant and 66 benign lesions). Women included in the study were those who received a final indication for histopathologic analysis of the lesion by means of breast biopsy because of a BI-RADS category 4 or 5 finding (see Section 2.2) on an x-ray screening mammogram. After receiving this final indication for breast biopsy, patients were invited to participate in the study, with the MR imaging examination being performed before the biopsy. Exclusion criteria were general exclusion criteria for performing

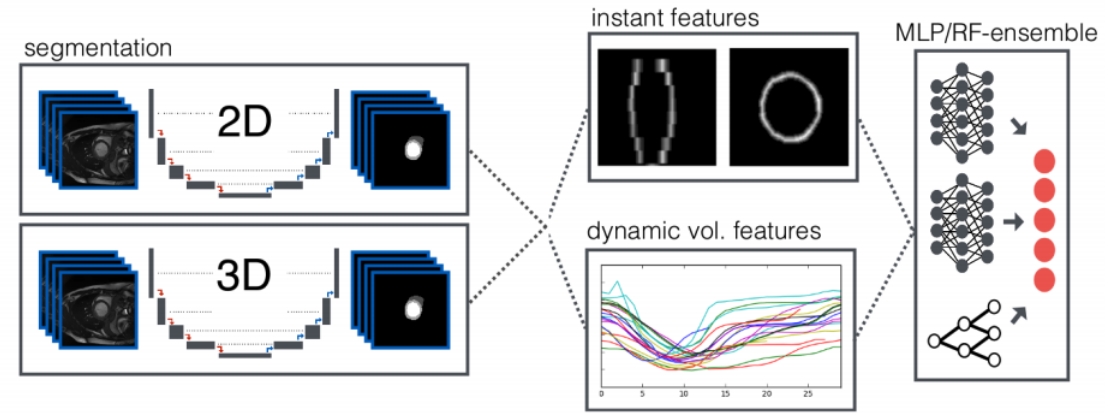

Figure 4.2.: **Overview of the Proposed Pipeline for Automated Cardiac Diagnosis on Cine-MRI**. Segmentation predictions from a 2D and a 3D model are averaged and used to extract instant and dynamic volume features, which are fed into an ensemble of classifiers for disease prediction.

an MR imaging examination (eg, ferromagnetic implants, severe claustrophobia, allergies to contrast agents and others) (n = 1) or unwillingness to participate in the study (n = 2).

MR images were acquired by using 1.5-T MR imaging devices from two different vendors (for explanation of the following protocol data see Section 2.1.1). At the first study site the imaging device (Aera; Siemens Erlangen, Germany) was used with an 18-channel breast coil in 95 patients for the training set. At the second study site the imaging device (Ingenia; Philips, Best, The Netherlands) was used with a two channel breast coil with additive elements on the MR imaging table for the independent test set of 127 patients. All women were imaged in the prone position with the breast not compressed but softly supported by foam material. All MR imaging examinations were performed before biopsy. A full diagnostic breast MR imaging protocol was performed as previously described, including DWI sequences. DWI was performed with transverse echo-planar imaging by using a single-shot technique at the first study site and multishot with readout segmentation at the second site. Section orientation was axial and section thickness was 3 mm at both study sites. b values were set to 0, 100, 750 and 1500 $\mathrm{s\,mm^{-2}}$. Imaging time for the DWI sequence was around 4 minutes at the first site and around 7 minutes at the second site.

After acquisition, all images were read out by two expert radiologists to identify all index lesions in consensus using the DWI source images with b values of 750 $\mathrm{s\,mm^{-2}}$ and 1500 $\mathrm{s\,mm^{-2}}$ and the T2-weighted morphologic images. Information on index lesion localization was given in the x-ray screening mammographic report. The abbreviated MR

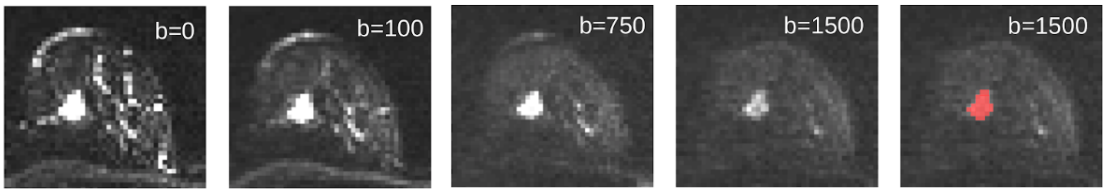

Figure 4.3.: **Set of diffusion-weighted images at different b-values for one patient.** Additional example segmentation on the $b = 1500\,\mathrm{s\,mm^{-2}}$ image.

imaging protocol provided to the reader consisted of a T2-weighted morphologic sequence and the DWI sequence. The index lesions as described in the x-ray mammography screening report were located by using the complementary information of both the DWI and T2-weighted sequences and were manually segmented in the DWI sequence on all sections visible, resulting in a three-dimensional volume image of the lesion. Lesions were segmented by using the inner border of the lesion to minimize partial volume effects. For this purpose, the DWI image with the highest b value that provided visibility of the lesion was used. Thus, lesions clearly visible with all b values were segmented on sections with a b value of $1500\,\mathrm{s\,mm^{-2}}$, while lesions not visible on sections with a b value of $1500\,\mathrm{s\,mm^{-2}}$ were segmented on sections with a b value of $750\,\mathrm{s\,mm^{-2}}$.

Subsequent biopsy, i.e. cancer burden and histopathologic evaluation of the lesions revealed 122 (54.9%) malignant lesions in the group of 222 patients, with the most common malignant lesion being invasive ductal carcinoma in 90 (73.77%) patients. Among the benign lesions, which were found in 100 (45.04%) patients, the most common lesions described were fibrosis, in 21 (21%) patients, and fibroadenoma in 20 (20%) patients. The described histopathological information is utilized as ground truth for training the machine learning model in this study. Figure 4.3 shows a DWI series for an exemplary patient. Extended information on histopathology of lesions and size statistics is provided in Appendix B.

Cardiac Disease Classification on MRI

The ACDC dataset [34] comprises short-axis cine-MRI of 150 patients acquired at the University Hospital of Dijon using two MR scanners of different magnetic strengths (1.5 T and 3.0 T). Each time-series is composed of 28 to 40 3D volumes, which partially or completely cover the cardiac cycle. As typical for CMRI, the data is characterized by a high in-plane resolution ranging from 0.49 to $3.69\,\mathrm{mm^2}$ and a low resolution in the direction of the long axis of the heart ($5 - 10\,\mathrm{mm}$ slice thickness). Note that some data exhibit severe slice misalignments, which originate from different breath hold positions between slice stack acquisitions. Structures of interest, namely Left Ventricular Cavity (LVC),

Left Ventricular Myocardium (LVM) and Right Ventricular Cavity (RVC) were segmented manually by clinical experts on End Dyastolic (ED) and End Systolic (ES) phase instants. Four pathological groups and one group of healthy patients are evenly distributed in the dataset: patients with previous Myocardial Infarction (MINF), Dilated Cardiomyopathy (DCM), Hypertrophic Cardiomyopathy (HCM), Abnormal Right Ventricle (ARV) and normal (healthy) subjects (NOR). Additional information for all patients is provided in form of height and weight. The dataset has been split by the challenge organizers into 100 training and 50 test patients. Segmentation and classification ground truth is provided only for the 100 training cases. All reported test set results were obtained by submitting our predictions to the online evaluation platform.

4.3. Methodology

4.3.1. Breast Lesion Classification on DWI

Enhancing the Biophysical Model by Fat Area Calibration

In recent clinical research, Diffusion Kurtosis Imaging (DKI) is performed in the annotated regions of interest on the raw b-value data to extract the ADC and AKC biomarkers (as described in Section 2.1.1). While the literature uses a constant value for the background signal intensity level as in equation 2.2, we introduce a data-dependent intensity level calibration vector $\theta(b)$. This vector is determined as the intensity of an additionally segmented fat area extracted from all b values so as to normalize the background signal variation across patients:

$$S(b) = \left(\theta(b)^2 + \left(S_0 \ \exp\left(-b \ \mathrm{ADC} + \frac{1}{6} \ b^2 \mathrm{ADC}^2 \mathrm{AKC} \right) \right)^2 \right)^{1/2} \tag{4.1}$$

The effect of this factor on classification performance is studied in a subsequent analysis by comparing the final results when using θ as described versus setting θ equal to 0. The signal intensities for $b = 0$ are omitted during the fit because of drastic instabilities, instead, they are utilized as a third free-fit parameter. DKI is performed on each individual voxel in the ROI. The resulting ADCs are required to be within $0 - 3.5\,\mu\mathrm{m}^2\,\mathrm{ms}^{-1}$ and AKCs within $0 - 3$. Voxels yielding values outside of these intervals are excluded from the ROI and considered background for the remainder of the analysis. These requirements are applied to exclude outliers caused by artifacts such as flow or motion without distorting correctly determined values. The upper limit of ADCs is chosen as the expected ADC value of $3.25\,\mu\mathrm{m}^2\,\mathrm{ms}^{-1}$ for free water, and the upper limit of AKCs was based on common assumptions and observations from the literature [51, 57, 147]. Examples for the resulting normalized image maps containing ADCs and AKCs are displayed in Figure 4.4).

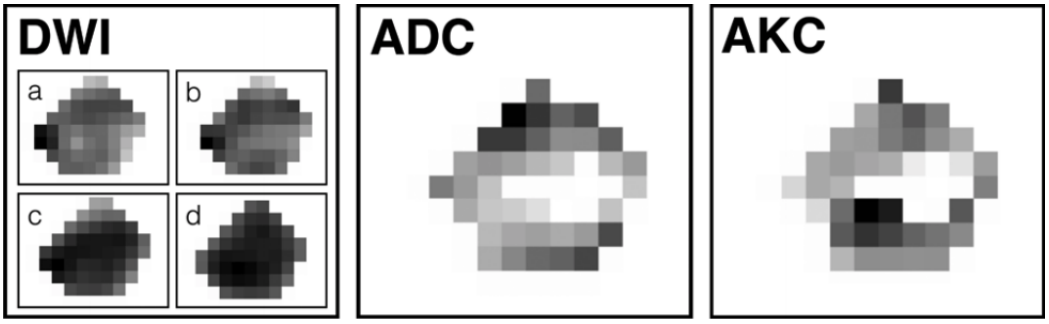

Figure 4.4.: **Example regions of interest from raw DWI images and resulting ADC and AKC maps.** Two-dimensional sections of a three-dimensional acquisition of images of a malignant tumor. Diffusion Weighted Magentic Resonance Imaging with b values of (a) 0, (b) 100, (c) 750, and (d) 1500 $\mathrm{s\,mm^{-2}}$. Apparent Diffusion Coefficient map and Apparent Kurtosis Coefficient map show the resulting pixel values after Kurtosis fitting. Notably, white pixels inside lesion constitute background after exclusion for not matching the fit criteria.

Feature Extraction

In contrast to computing the mean or median over coefficient maps, in this study we propose to perform large scale Radiomic feature extraction on the ADC and AKC maps by using the Medical Image Interaction Toolkit [148]: First, four Daubechies wavelet transforms are performed on the two input maps, yielding a total of 10 image variations. These wavelet decompositions of the original images evaluate the radiomics features at varying spatial frequencies and resolutions with pronounced focus on edge information, along the three spatial directions. Subsequently, the following features are extracted:

- 21 firstorder statistics calculated from the histogram of voxel intensities using first-order statistics
- 17 volume and shape features including diametral, volumetric and surface measurements, as well as shape features such as compactness or sphericity.
- 321 texture features so as to characterize the topography of intensity distribution and periodicity in the tumor volume as well as co-occurrence, run-length, size-zone, and neighborhood gray level based features.

A detailed description of utilized features is provided in Appendix C. The obtained feature vectors of the 10 image variations are concatenated, yielding a final vector containing 3590 features. Finally, the age of the patients is added as an additional clinical parameter.

Random Forest Training

A Random Forest Regressor (see Section 3.1) is trained for dimensionality reduction, i.e., mapping the Radiomics feature vector to a one-dimensional malignancy score. Biopsy results of the 222 patients are utilized as ground truth information during training. Since the true class balance between benign and malignant lesions for this clinical scenario is estimated to be around 0.5 (following the 50% false positives in mammography [149]), the Random Forest output can be interpreted as a probability of malignancy for each lesion. A Regression Forest is preferred over a Classification Forest here, because it comes with a continuous output and hence enables ROC analysis, i.e. empirical choosing of working points for deployment on the individual test set. No feature selection is applied to the data beforehand, since Random Forests perform intrinsic feature selection. For training, a five-fold cross validation is performed on the training set. Besides the model parameters, the hyper-parameters of the model, e.g., number of estimators, maximal depth of the trees or the random state were optimized during validation. This was done using grid search and yielded the following results: number of estimators = 300, maximum depth = 5. Thereby, it is crucial to fix the hyper parameters before unblinding the test dataset to avoid information leakage, which can lead to overfitting on the test set. Finally, the multivariate model is fixed and deployed on the independent test data. Note, that no changes are made to the model at this point and no ground truth information of the test data has been used during previous model development. For further enhancement of the predictions, an ensemble of 100 Random Forests is trained each using differently shuffled subsample splits of the training set during cross-validation. The final prediction for each patient is then given by the average over the predictions of all 100 models.

4.3.2. Cardiac Disease Classification on MRI

Segmentation of Cardiac Structures

For the fist step, the segmentation of cardiac target structures, we resample all volumes to $1.25 \times 1.25 \times 10\,\text{mm}$ per voxel (for 3D UNet and feature extraction) and $1.25 \times 1.25 \times Z_{orig}$ mm per voxel (for 2D UNet) to account for varying spatial resolutions. The grey level information of every image was normalized to zero-mean and unit-variance. We tackle the segmentation using an ensemble of modified 2D and 3D UNets [108, 150] (see Section 3.3). We carefully adapted the architecture to cope with specific challenges of CMRI: Due to low z-resolution of the input, pooling and upscaling operations are carried out only in the x-y-plane. Context in the z-dimension is solely aggregated through the 3D convolutions. Each feature extraction block consists of two padded $3 \times 3 \times 3$ convolutions, followed by batch normalization and a leaky ReLU nonlinearity. Due to the shallow nature of the network (18 layers) no residual connections are utilized. The initial

number of 26 feature maps is doubled (halved) with each of the 4 pooling (upscaling) operations, resulting in a maximum of 416 feature maps at the bottom of the U-shape. Deep supervision (as in [151]) is implemented by generating low resolution segmentation outputs via $1 \times 1 \times 1$ convolutions before each of the last two upscaling operations, which are upscaled and aggregated for the final segmentation.

The 3D model was trained for 300 epochs in a 5-fold cross validation using the ADAM solver and a pixel-wise categorical cross-entropy loss. The initial learning rate of $5 \cdot 10^{-4}$ was decayed by 0.98 per epoch, where an epoch was defined as 100 batches, each comprising four training examples. Training examples were generated as random crops of size $224 \times 224 \times 10$ taken from randomly chosen training cases and phase instances (ED/ES). The 2D model's architecture is equivalent to the 3D approach except for 2D convolutions. Due to the lower memory requirements, we increased the number of initial feature maps to 48. The network is trained with a batch size of 10 and input patches of size 352×352 pixels using a multi-class dice loss:

$$\mathcal{L}_{\text{dc}} = -\frac{2}{|K|} \sum_{k \in K} \frac{\sum_i u_i^k v_i^k}{\sum_i u_i^k + \sum_i v_i^k}, \tag{4.2}$$

where u is the softmax output of the network and v denotes a one hot encoding of the ground truth segmentation map. Both u and v are of size $i \times k$ with i being the number of pixels in the training patch and $k \in K$ being the classes. To accomplish the training of a well generalizing model on limited data, we used a broad range of data augmentation techniques, such as mirroring along the x and y axes, random rotations, gamma-correction and elastic deformations. Due to the low z-resolution all data augmentation was performed only in the x-y-plane. To account for the presence of slice misalignments, we artificially increased the number of misaligned slices by motion augmentation for the training of the 3D model: All slices within the training batch were perturbed with a probability of 10% and a random offset drawn from $\mathcal{N}(0, 20)$. To obtain the final segmentations, softmax outputs of both networks were resampled to the original voxel resolution of the input image and then averaged.

Feature Extraction

We extract two sets of features from the previously segmented structures to perform disease classification. All features were designed to quantify the traditional assessment procedures of expert cardiologists by describing static and dynamic properties of the structures of interest (see Table 4.1).

Instant features Extracted from the two labeled ED and ES time instants as provided by the ACDC dataset, these features cover local and global shape information (circumference, circularity, LVM thickness, etc.), local variations (size of RVC at the apex, LVM

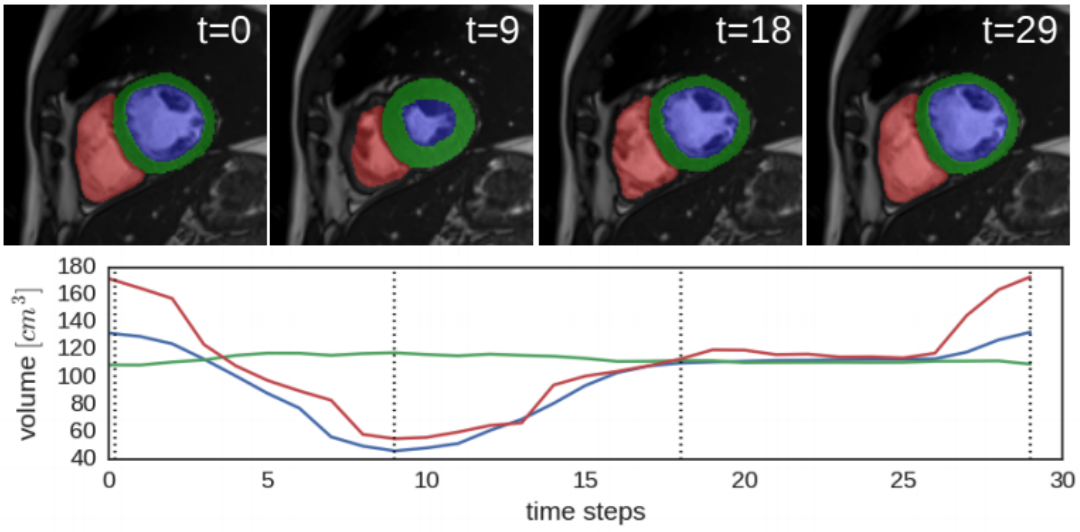

Figure 4.5.: **Time-series segmentation of cardiac structures on cine-MRI**. Time-series segmentation for RVC (red), LVM (green), LVC (blue) and their corresponding volume dynamics. The example shows the central slice in z direction of a healthy patient (NOR).

thickness between RVC and LVC), simple texture descriptors (mass) as well as additional meta information (body mass index, weight, height). Notably, all thicknesses, circumferences and circularities are computed on the individual x-y-planes and aggregated over the z-dimension. The body surface is estimated from weight and height using the Mosteller formula.

Dynamic volume features We deployed the trained segmentation model to predict the anatomical structures in all time steps of the CMRI (see Figure 4.5). This allows for exploitation of volume dynamics throughout the entire cardiac cycle independent of the predefined ED/ES. These volume dynamics are quantified in form of first order statistics (median, standard deviation, kurtosis, skewness) complemented by characteristics of the cardiac cycle's minimum and maximum volumes: We found the time instants of these extrema to not match the predefined ED/ES instants in the majority of patients. This finding is accounted for by computing volume, volume ratios and ejection fractions based on the determined actual minimum (v_{min}) and maximum (v_{max}) volume of the cardiac cycle. Finally, the synchrony of contraction between LVC and RVC is measured in form of the time step differences between their corresponding v_{min} and v_{max}.

Classification

The features described in section 4.3.2 were used to train an ensemble of 50 Multilayer Perceptron (MLP)s and a Random Forest for pathology classification. The MLP's ar-

Table 4.1.: **List of extracted features for cardiag diagnosis task.** The two sets of features extracted for disease classification and the corresponding cardiovascular structure (RVC, LVM, LVC). All instant features (except for additional patient information) are extracted on both ED and ES.
*this feature was calculated in the x-y-plane and aggregated over slices in z.
** $v_{min,LVM}$ was determined at $t(v_{min,LVC})$.

instant features	**RVC**	**LVM**	**LVC**
max thickness*		x	
min thickness*		x	
std thickness*		x	
mean thickness*		x	
std thickness of LVM between LVC and RVC*			
mean thickness of LVM between LVC and RVC*			
mean circularity*	x	x	
max circumference*	x	x	
mean circumference*	x	x	
RVC size at most apical LVM slice*			
RVC to LVC size ratio at most apical LVM slice*			
volume per m^2 body surface	x	x	x
mass		x	
patient weight			
patient height			
patient body mass index			

dynamic volume features	**RVC**	**LVM**	**LVC**
v_{max}	x	x	x
v_{min}	x	x**	x
dynamic ejection fraction	x	x**	x
volume median	x	x	x
volume kurtosis	x	x	x
volume skewness	x	x	x
volume standard deviation	x	x	x
volume ratio $v_{min,LVC}/v_{min,RVC}$			
volume ratio $v_{min,LVM}/v_{min,LVC}$			
volume ratio $v_{min,RVC}/v_{min,LVM}$			
time step difference $t(v_{min,LVC})$-$t(v_{min,RVC})$			
time step difference $t(v_{max,LVC})$-$t(v_{max,RVC})$			

chitecture consists of four hidden layers, each containing 32 units, followed by batch normalization, leaky ReLU nonlinearity and a Gaussian noise layer ($\sigma = 0.1$). Each MLP was trained on a random subset of 75% of the training data, while the remaining 25% were used for epoch selection. Further regularization was provoked by only presenting a random subset of 2/3 of the features to each MLP. We trained all MLP for 400 epochs (with a patience of 40 epochs) using the ADAM solver with an initial learning rate of $5 \cdot 10^{-4}$, decayed by 0.97 per epoch. An epoch was defined as a set of 50 batches containing 20 patients each. Additionally, we trained a random forest with 1000 trees. During testing,

the softmax outputs of all MLPs were averaged to obtain an overall MLP score, which was averaged subsequently with the random forest output to obtain the final ensemble prediction.

4.4. Results

4.4.1. Breast Lesion Classification on DWI

Evaluation metrics

Since this study is intended to enter the clinical workflow as a follow-up on mammography, we are interested in decreasing the number of false-positive decisions, while not producing new false-negative results. Hence, we choose our sensitivity working point, such that a sensitivity of at least 98% is retained, with the aim of providing sensitivity that is comparable to that of core-needle biopsy, for which sensitivities ranging from 87% to greater than 97% have been reported [152, 153, 154, 155, 156]. Subsequently, this threshold is applied to the test set to evaluate the quality of decisions. In case a lesion contained too few voxels (< 4) for Radiomics analysis, the associated prediction was automatically set to be malignant. If no lesion was observed on the raw images during annotation, the sample was predicted to be benign. Differences in performance were evaluated by comparing the AUC using the method of DeLong et al. [157]. Since this study focuses on the clinically relevant section of the AUC where the sensitivity was predefined to exceed 98%, a pAUC is evaluated in a bootstrapping approach. Differences in specificity are evaluated by means of the McNemar test. The model was additionally evaluated at the sensitivity and specificity thresholds of a radiologist's performance, who given the full protocol, i.e. additional MRI modalities including contrast enhancement.

Feature importance

Table 4.2 shows a ranking of the 20 most important Radiomics features determined as the mean decrease in impurity, which calculates feature importance as the sum over the number of tree splits (across all tress) that include the feature, proportionally to the number of samples it splits. Advanced features based on texture or shape were shown to be less meaningful for the classification than simple first-order statistics.

Statistical Analysis

The thresholds applied on voxel coefficients resulting from DKI reduced the median number of voxels per lesion by 36.6%. However, because of a heterogeneous distribution of

Table 4.2.: **Radiomics Features Ranked by the Mean Decrease in Impurity of the Random Forest.** Further description of Radiomics features can be found in Appendix C. Data in parentheses are number of discrete grey values. N, L, V correspond to different wavelet transformations (N = no transformation). RMS denotes Root Mean Square, co-occ denotes co-occurrence

Rank	Feature Name	Mean Decrease in Impurity (%)
1	N/ADC/FirstOrder 10th Percentile	0.1599
2	N/AKC/FirstOrder 90th Percentile	0.0587
3	N/AKC/FirstOrder RMS	0.0534
4	N/AKC/FirstOrder Median	0.0494
5	N/AKC/FirstOrder Energy	0.0336
6	V/AKC/SizeZone (64) HighGreyLevelZoneEmphasis	0.0251
7	V/AKC/SizeZone (64) SmallZoneLowGreyLevelEmphasis	0.0185
8	V/AKC/SizeZone (128) SmallZoneHighGreyLevelEmphasis	0.0156
9	N/AKC/FirstOrder Maximum	0.0154
10	L/AKC/FirstOrder RMS	0.0147
11	D/ADC/FirstOrder Median	0.0146
12	V/AKC/FirstOrder RMS	0.0137
13	N/ADC/SizeZone (128) SmallZoneLowGreyLevelEmphasis	0.0129
14	V/AKC/SizeZone (256) HighGreyLevelZoneEmphasis	0.0116
15	V/AKC/SizeZone (256) SmallZoneLowGreyLevelEmphasis	0.0107
16	V/AKC/SizeZone (128) SmallZoneLowGreyLevelEmphasis	0.0103
17	Age	0.0087
18	D/AKC/FirstOrder 90th Percentile	0.0085
19	V/AKC/SizeZone (64) SmallZoneHighGreyLevelEmphasis	0.0084
20	N/AKC/co-occ (1) Autocorrelation Means	0.0082

Table 4.3.: **Final results of evaluation comparing the radiomics model to the univariate Parameters in the independent test set.** Data in parentheses are 95% confidence intervals (CIs), and data in brackets are numerators and denominators. Partial AUC is with sensitivity set to at least 98%. All P values are reported with respect to the corresponding radiomics performance. NA denotes "not available".

Parameter	AUC	P Value	Partial AUC	*P* Value	Sensitivity (%)	Specificity (%)	*P* Value
Overall diagnostic performance (including nonvisible lesions)							
Radiomics	0.911 (0.861, 0.962)		0.841 (0.787, 0.894)		98.4 [60/61]	69.7 [46/66]	
BI-RADS 4a	0.948 (0.885, 1.000)		0.870 (0.809, 1.000)		100 [9/9]	74.0 [37/50]	
BI-RADS 4b	0.775 (0.603, 0.950)		0.798 (0.663, 0.933)		100 [29/29]	60 [9/15]	
BI-RADS 5	0.870 (NA)		...		95.7 [22/23]	.001 [0/1]	
ADC median	0.841 (0.772, 0.910)	< .001	0.681 (0.625, 0.773)	< .001	96.7 [59/61]	48.5 [32/66]	< .001
AKC median	0.874 (0.813, 0.936)	.167	0.752 (0.694, 0.839)	.003	98.4 [60/61]	53.0 [35/66]	.007
Radiologist using full diagnostic protocol	...		...		91.8 [56/61]	74.2 [49/66]	
Performance without nonvisible lesions							
Radiomics	0.870 (0.799, 0.941)		0.776 (0.708, 0.843)		98.4 [60/61]	55.6 [25/45]	
ADC median	0.767 (0.670, 0.863)	.025	0.533 (0.505, 0.656)	< .001	96.7 [59/61]	24.4 [11/45]	< .001
AKC median	0.815 (0.728, 0.903)	.165	0.636 (0.578, 0.742)	.002	98.4 [60/61]	31.1 [14/45]	.006

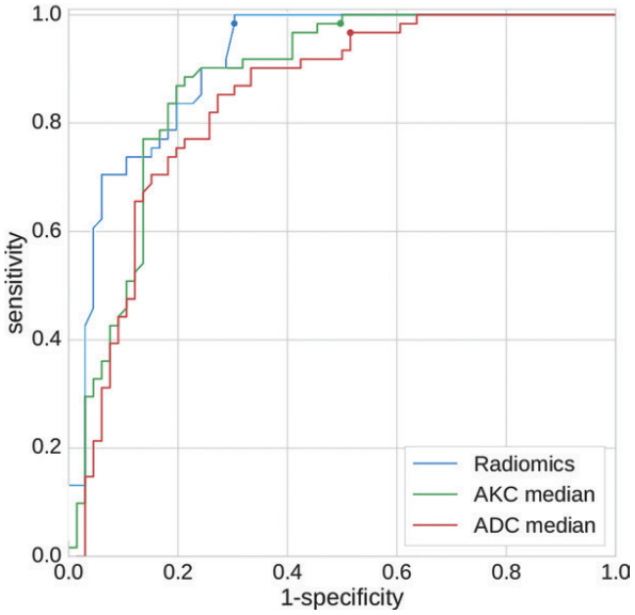

Figure 4.6.: **Receiver Operating Characteristic Curve for the Radiomics model, AKC median and ADC median.** Dots illustrate the working points resulting from the applied sensitivity threshold $\geq 98\%$.

excluded voxels within the lesions, the overall lesion sizes were reduced only marginally. To compare the performance of the radiomics analysis to univariate quantitative assessment, we studied the medians of the fit output parameters computed in every individual ROI on the independent test set (ADC median and AKC median). Table 4.3 shows that the Radiomics model performs best (AUC 0.911; 95% confidence interval [CI]: 0.861, 0.962) in comparison with the ADC median (0.841; 95% CI: 0.772, 0.910) and AKC median (0.874; 95% CI: 0.813, 0.936). For the full AUC analysis, this difference was only significant when Radiomics was compared to ADCs ($P < 0.001$). In the clinically relevant section of the AUC (pAUC), Radiomics performed significantly better than both univariate approaches (Radiomics versus AKC: $P = 0.02$, radiomics versus ADC: $P < 0.001$). Figure 4.6 visualizes these results.

Applying the decision rule of sensitivity ≥ 0.98 on the training set, i.e. choosing a working point on the ROC, yields the threshold value $t_c = 0.46$ on the Random Forest output to differentiate between benign and malignant lesions. Sensitivity and specificity values resulting from deploying t_c on the test set are shown in Table 4.3. The Radiomics model reduced the amount of false positive findings from 66 to 20 (specificity 0.697 [46 of 66]) while only one true positive result was missed (sensitivity 0.984 [60 of 61]). In the comparison of specificity values at equal sensitivity values, the Radiomics approach showed significant superiority with respect to ADC (0.485, $P < 0.001$) and AKC (0.530,

Table 4.4.: **Results of standard ADC Fit and Subsequent Radiomics Analysis on the Independent Test Set.** Data are percentages with 95% CIs in brackets and numerators/denominators in parentheses. All P values are reported with respect to the corresponding Radiomics performance. Abbreviations: AUC = area under the curve, se = sensitivity, sADC = standard ADC

	AUC	pAUC ($se \geq 0.98$)	Sensitivity	Specificity	t_c
sADC median	0.867 (0.805, 0.930)	0.776 (0.718, 0.851)	0.984 [60/61]	0.561 [37/66]	≤ 1.34
Radiomics (based on sADC)	0.855 (0.791, 0.912)	0.752 (0.697, 0.826)	0.951 [58/61]	0.561 [37/66]	≥ 0.345

$P = 0.02$). Evaluation of the diagnostic performance of an experienced breast radiologist who used the full diagnostic protocol (including the contrast-enhanced MR imaging sequences) on the test set revealed a sensitivity of 0.918 (56 of 61) and specificity of 0.742 (49 of 66). To provide a comparison to the Radiomics model, we adjusted the working point of the Radiomics model by either setting the specificity or the sensitivity identical to the radiologist. The Radiomics model provided identical diagnostic indexes for the identical sensitivity (resulting specificity, 74.2% [49 of 66]) and specificity (resulting sensitivity, 91.8% [56 of 61]) compared to the radiologist. As expected, BI-RADS 5 lesions had a much higher malignancy rate in the test set (95.8% [23 of 24]) when compared with BI-RADS 4a and 4b lesions (36.9%, [38 of 103]). Analysis of the Radiomics model shows that all clarified false-positive results were BIRADS 4a or 4b and the single benign lesion, which was wrongly categorized as BI-RADS 5 in previous mammography, was not identified correctly. Details on the sensitivity and specificity of the Radiomics model for each subgroup are demonstrated in Table 4.3. To evaluate the effect of the proposed fat suppression factor $\theta(b)$, we repeated the Radiomics analysis while setting $\theta(b) = 0$. We found a significant performance loss when taking out $\theta(b)$, for AUC (0.845; 95% CI: 0.780, 0.911; $P = 0.001$) and specificity (51.5% [34 of 66]; $P < 0.001$). We also compared our model to the performance of a standard ADC fit (as opposed to DKI), which is obtained by setting AKC=0 in equation 4.1, in order to evaluate the benefits of the AKC.

We observe slightly better performance for standard ADC compared to DKI, but a large decrease in performance when utilizing it as input for the radiomics analysis (see Table 4.4).

To isolate the credit for improvement to the model itself rather than entangled with the benefits of DWI compared to mammography, we tested if reported significance tests still hold when manual predictions for invisible lesions (two in the training set and 21 on the test set) were excluded from the analysis. Table 4.5 shows that all reported significances still hold. As expected, specificity and accuracy decreased with all methods because of the proportional increase of misclassified lesions in the dataset. The Radiomics model yields

Table 4.5.: **Final Results when Omitting Nonvisible Lesions Comparing Performances of the Test Set to Cross Validation Results.** Data are percentages with 95% CIs in brackets and numerators/denominators in parentheses. All P values are reported with respect to the corresponding Radiomics performance. Abbreviations: AUC = area under the curve, se = sensitivity

	AUC	pAUC ($se \geq 0.98$)	Sensitivity	Specificity	t_c
Training set					
Radiomics (*cross validation results*)	0.887 (0.822, 0.952)	0.703 (0.621, 0.810)	0.984 [60/61]	0.438 [14/32]	≥ 0.46
ADC median	0.832 (0.746, 0.918)	0.588 (0.526, 0.740)	0.984 [60/61]	0.313 [10/32]	≤ 2.23
AKC median	0.831 (0.745, 0.917)	0.523 (0.500, 0.779)	0.984 [60/61]	0.281 [9/32]	≥ 0.64
Test set					
Radiomics	0.870 (0.799, 0.941)	0.776 (0.708–0.843)	0.984 [60/61]	0.556 [25/45]	
ADC median	0.767 (0.670, 0.863) *P* = .025	0.533 (0.505, 0.656) *P* < .001	0.967 [59/61]	0.244 [11/45] *P* < .001	
AKC median	0.815 (0.728, 0.903) *P* = .165	0.636 (0.578, 0.742) *P* = .002	0.984 [60/61]	0.311 [14/45] *P* = .006	

an AUC of 0.87 (95% CI: 0.79, 0.94) with sensitivity of 0.984 (60 of 61) and specificity of 0.556 (25 of 45). In order to assess the dependency of results on human noise induced in the ROI annotation process ("inter-reader variability"), we deployed the Radiomics model on a test set with alternative ROIs manually segmented by a second reader. No significant changes were observed, only a marginal improvement in AUC (0.918; 95% CI: 0.870, 0.910) and slightly higher sensitivity of 100% [61 of 61] at equal specificity (69.7% [46 of 66]).

4.4.2. Cardiac Disease Classification on MRI

We trained the classification ensemble on the ACDC training data using the features described in Section 4.3.2. In a five fold cross-validation, a classification accuracy of 94% was achieved. The individual performance of the MLP ensemble and Random Forest were 93% and 92%, respectively. The test set accuracy was 92%. Confusion matrices are provided in Figure 4.7, indicating equal perfomance among classes in the cross-validation, and difficulties in distinguishing DCM from MINF patients on the test set. Feature computation took 15 s for instant features and less than one second for the dynamic volume features.

4.5. Discussion

We presented a machine learning model based on Radiomics features extracted from Diffusion Kurtosis Imaging of microstructural breast tissue that reduces false positive results

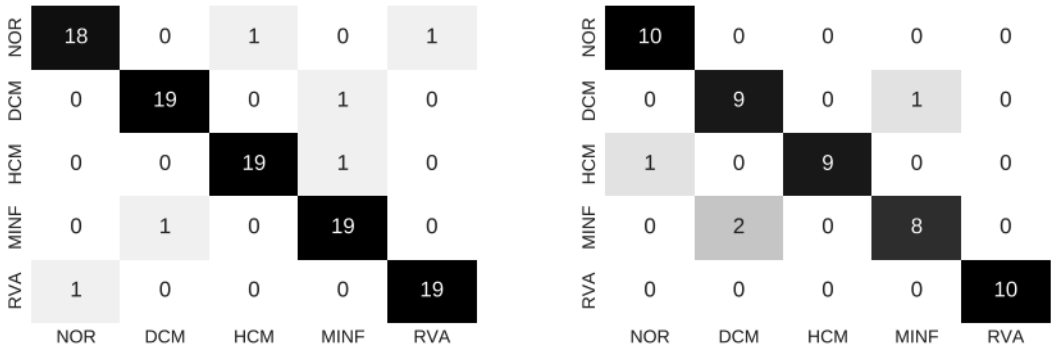

Figure 4.7.: **Cross validation and test set results for classification task of cardiac diagnosis competition.** Confusion matrices of the ensemble predictions from cross-validation on the training set (left) and on the test set (right). Rows correspond to the predicted class and columns to the target class, respectively.

by 70% in lesions classified as BI-RADS 4 or 5 at screening mammography while retaining sensitivity greater than 98%. Our model was evaluated on an independent test set with differing MR imaging machines from a separate study sites in order to simulate clinical application. We achieve human-level performance on an independent test set comprising images from a different clinical site acquired from a different scanner, when comparing to a radiologist that was allowed to consult the full imaging protocol including contrast-enhanced sequences. When ranking all features extracted with Radiomics, we found that advanced features based on texture or shape were less significant for classification than simple first-order statistics. This was presumably due to the relatively small regions of interest in our datasets (median size: 21 voxels) and indicates the utilized dataset as a difficult setup for Radiomics, since a significant amount of the associated features describes non-local voxel dependencies. Sun et al. [87] reported higher AUCs of 0.974 for AKC compared to our results. However, their study was performed on a dataset without independent validation and with predominately large lesions (benign: 1.9 cm ±1.0, malignant: 2.4 cm ±1.0) as opposed to our dataset (see Table B.2). Furthermore, their ratio of DCIS, which are notoriously hard to categorize based on imaging, was relatively low (8.8%, [5 / 57]) compared to our dataset (22.9%, [14 of 61], see Table B.1). Since DWI of the breast is challenged by high tissue heterogeneity with intermittent fat and glandular tissue, we enhanced current kurtosis fitting strategies by introducing a fat tissue correction term, $\theta(b)$, in analogy to the air volume correction term suggested by Jensen et al. [51]. In comparison to standard kurtosis, this adjusted model significantly improved differentiation between malignant and benign lesions. In the independent test set of 127 women, the Radiomics analysis allowed 70% (46 of 66) of the false-positive findings to be detected by means of noninvasive imaging, while retaining sensitivity of 98.4% (60 of 61) for malignant lesions. This is similar to the reduction of false-positive results as

reported for full contrast-enhanced MR mammography when added to BI-RADS 4 lesions detected with other imaging modalities [158]. In comparison with full contrast-enhanced MR imaging mammography, however, the approach as presented here might be of special value, because no gadolinium based contrast agents must be given to patients, thus reducing the risk of adverse effects and costs for the examination [159]. Costs might be a further aspect to be considered, because acquisition time for the DWI sequence used for the kurtosis fitting was less than 7 minutes, allowing for an increased number of patients examined with the method as previously suggested for DWI sequences [85]. This study has several limitations: The model depends on the quality of manual segmentation, thus lesions that were incorrectly segmented, either because of the limited experience of the reader or because of limited visibility on the images, could be missed in malignancy classification. Inter-reader variability was assessed by deploying the fixed Radiomics analysis model on the test set, with alternative regions of interest manually segmented by an independent reader, demonstrating similar diagnostic performance, and further indicating the robustness of the model. Some lesions considered benign with the approach were classified as benign because of their non-visibility on DWI, thus the incremental benefit of the Kurtosis model was not only due to the Radiomics analysis but also to the underlying DWI sequence, which must be considered when interpreting our results. We also showed, however, that results of all applied significance tests remained significant when non-visible lesions were excluded from the analyses. The rate of malignant lesions can be considered atypical, with more malignant lesions than expected in clinical practice on the test set for the BI-RADS subgroups. However, the values of sensitivity and specificity did not decrease on the test set. Another limitation is that the choice of θ may not have correctly reflected the fat signal intensity fractions in the actual lesion regions of interest. Indeed, it is very challenging to correctly obtain fat signal intensity fractions in each single voxel with MR imaging artifacts spreading the fat signal in different areas. The improved performance for differentiating benign and malignant lesions by using our choice of θ seems to confirm that this was a first step in the right direction to account for residual fat signal intensity. In conclusion, a radiomics breast cancer model based on DWI with adapted kurtosis fitting allowed for improved differentiation between malignant and benign breast lesions in both training and independent test datasets acquired by using MR imaging machines from different vendors at different institutions. Our results support further evaluation of the use of optimized Diffusion Weighted Magentic Resonance Imaging (DWI) with multivariate analysis based on Radiomics features to differentiate between malignant and benign breast lesions.

We further presented a fully automatic processing pipeline for Cardiac Disease Classification on MRI. There, we started by developing an accurate multistructure segmentation method trained solely on ED and ES phase instances, but capable of processing the entire cardiac cycle. Our approach revolves around the use of both a 2D and 3D

model, leveraging their respective advantages through ensembling. The resulting pipeline is robust against slice misalignments, different CMRI protocols as well as various pathologies. We achieved the first place in the segmentation part of the challenge. Based on the segmentations generated by our model, geometrical features are extracted and utilized by an ensemble of classifiers to predict the diagnosis, yielding promising outcomes. We ranked second in the classification part of the challenge with an accuracy of 92% [34]. Our fully automatic processing pipeline constitutes an attractive software for clinical decision support due to the visualization of intermediate segmentation maps, the comprehensive quantification of cardiologic assessment and the rapid processing speed of less than 40 s. Possible future improvements of the model concern data augmentation and the architecture of the segmentation network as well as a regularization objective as used in [160]. Given continuously growing datasets, training the cardiac diagnosis task directly on the raw images might soon be feasible and surpass the performance of the presented multi-step pipeline following the premise of end-to-end learning.

5. Medical Image Classification based on Learned Representations

We further follow the dogma of end-to-end learning, i.e. the idea that enabling simultaneous optimization of all pipeline components with respect to the ultimate clinical target improves upon compound rule-based diagnosis pipelines. To this end, we substitute the handcrafted feature extraction (Radiomics) deployed in Chapter 4 by a learning algorithm allowing to learn representations directly on crops of annotated Regions of Interest in the images ("Roi-to-end", see Section 3.2). The presented study showcases this approach by means of breast lesion classification on DWI. Thus, a second focus of this chapter is set on further integrating the biophysical model applied for image normalization in DWI (see Section 2.1.1) into the learning algorithm and study the potential hidden in operating on the raw Diffusion Weighted Magentic Resonance Imaging (DWI) signal as opposed to operating on ADC or AKC maps. The main contributions in this chapter are:

- We propose a Convolutional Neural Network (CNN) architecture designed to integrate the biophysical model for image normalization, handcrafted feature extraction as well as clinical categorization, so as to enable ROI-based classification of breast lesions on DWI [35].
- We reveal potential hidden in DWI by demonstrating the benefits of learned image normalizations as compared to the biophysical model currently deployed in clinical research.
- We provide results indicating a complementary value of representations learned in the CNN with respect to handcrafted feature extraction [36].

The parts of this chapter regarding evaluation of learning DWI image normalization model were presented at MICCAI 2017 and published in the respective proceedings [35]. The

comparison against Radiomics was presented at the annual meeting of the International Society for Magnetic Resonance Imaging in Medicine (ISMRM) and is available at [36].

5.1. Problem statement

In DWI, the signal behavior at different diffusion gradients is quantitatively characterized by fitting biophysical models to the signal and inferring apparent tissue properties from them (see Section 2.1.1). The state of the art method in breast cancer DWI is DKI, where the ADC and AKC are extracted representing Gaussian and non-Gaussian diffusion, respectively [51, 50]. Using DKI, state of the art results for breast lesion classification have been reported recently [161, 87]. Methods relying on biophysical models for image normalization, however, are simplified approaches to physical processes, making them prone to partial information loss and dependent on explicit prior physical knowledge resulting in potential fitting instabilities and limited generalization abilities. These shortcomings have led to an emergence of a broad spectrum of signal and noise models designed under different assumptions. Recent studies in brain imaging have shown how deep learning can circumvent some of the disadvantages related to classical model-based approaches in diffusion MRI data processing [162, 163, 164, 165]. However, the currently existing learning-based approaches cannot be more knowledgeable since they utilize the classical model-based approach as ground truth during training. Thus, the performance of existing model-free approaches is currently sealed, and the main benefit of machine learning application in this domain so far was found in the reduction of requirements on the input data side, e.g. saving acquisition time.

In this study, we show in a first clinical scenario how model-free diffusion MRI can be integrated into a deep learning algorithm, thus directly relating clinical information to the raw input signal. By backpropagating this information through an integrative CNN architecture, simultaneous and target-specific optimization of image normalization, signal exploitation, representation learning and classification is achieved.

5.2. Utilized Dataset

The dataset is identical to the one in Section 4.2. However, as this project represents a shift towards methodological research compared to the approach from Chapter 4, we prioritized the number of training cases over the clinically interesting aspect of testing under a input domain shift. Thus, we opted for mixing patients from both study sites into one dataset and run evaluation in a nested cross-validation over all cases (details on the evaluation scheme can be found in Section 5.4).

5.3. Methods

The recently proposed q-space Deep Learning method [162] uses neural networks to imitate model-based approaches like DKI by training them on model-derived parameters, i.e. the networks are trained to imitate the biophysical models. In contrast to that method, we aim to replace the model-based approaches by not using any model-related parameters as training target. Instead we train our approach directly on targeted clinical decision. By integrating the data processing pipeline from annotated ROI to clinical decision into a CNN, this valuable information is backpropagated through the network optimizing all pipeline components on the specific clinical task. This enables our approach to yield performances beyond model-based methods. The proposed architecture consists of four modules:

Input and Image Normalization Module The proposed CNN architecture is developed to operate directly on the diffusion-weighted images as input, where each of the four b-value images is assigned to a corresponding input channel of the network. For the task of lesion classification every image is cropped to a bounding box around the segmented ROI and voxels outside of the ROI are set to 0. Intensity calibration over the data cohort can be essential when working with raw MRI signal intensities. To facilitate this step, we measure the mean signal intensity of an additional ROI placed in a fat area of a breast in each image. The measured value is arrayed to match the shape of the corresponding lesion ROI and provided to the CNN as a fat intensity map in an additional input channel. Note, that this method corresponds to the calibration vector θ introduced in Section 4.3.1.

Signal Exploitation Module This CNN component is designed to mimic the biophysical model fit. The input is processed by layers of 1x1 convolutions, which only convolve the signals in each separate voxel across the input channels. This method is equivalent to applying an independent multilayer perceptron to each voxel, like it is done in q-space Deep Learning, i.e it enables the network to exploit the information contained in the differently weighted signals for independent voxels. The additional input channel for image normalization extends the set of differently weighted signals in every voxel by the corresponding value of the fat intensity map, thus transferring normalizing information about the image into all 1x1 convolutions. Three layers of 1x1 convolutions are applied transforming the input data into 512 feature maps. In analogy to model-based diffusion coefficients, we term these representations Deep Diffusion Coefficient (DDC), where each of the 512 feature maps corresponds to one coefficient.

Representation Learning Module Feature extraction is an integral part of the diagnosis pipeline. While in Chapter 4 we replaced the clinical practice of computing median as a single feature by large scale Radiomics feature extraction, we now follow the dogma of end-to-end learning and integrate this component into the CNN by means of a representation learning module (see Chapter 3.2). To this end, the DDC maps from the previous

Signal Exploitation Module are processed by two blocks of three 3×3 convolutional layers, while downsampling the input sizes between the blocks using 2×2 max pooling. In principal, this component can be repeated arbitrarily often, but is limited in the case of lesion classification by the relatively small input sizes of the ROIs.

Binary Classification Module The final convolutional layer containing the learned representations in form of feature maps is followed by a global average pooling layer, which aggregates the representations by transforming each feature map into a single mean value. Note, that this all convolutional architecture allows for variable input sizes by avoiding any classical dense layers [103], which we exploit by processing ROIs of different sizes through the same network. The output is a vector with the length of the number of feature maps containing the representations. This feature vector is used as input for a softmax layer transforming the features into class probabilities, which the binary classification is performed on using a categorical cross entropy loss function. By training the proposed network architecture in an "Roi-to-end" fashion, image calibration, signal exploitation and representations are learned simultaneously and optimized directly for the classification problem.

5.4. Experimental Setup

Designing an appropriate experimental setup is crucial in this study, in order to be able to assign credit for experimental results to the correct modifications of the diagnosis pipeline. To investigate the potential improvement of the proposed approach for clinical decision making, a two step evaluation is performed. First, the performance of model-free signal exploitation is assessed by a DDC-score against the means of the model-based coefficients ADC and AKC. The DDC-score is generated by applying global average pooling directly on the 512 DDC feature maps and running a linear classifier on top of the resulting 512-dimensional vector to obtain a single scalar score. In a second experiment, the Roi-to-end (R2E) approach, i.e. the simultaneous optimization of all DWI data processing components, is evaluated by comparing it against a model-based CNN method. For this benchmark, the R2E architecture is modified by feeding parametric maps of ADC and AKC into the representation module instead of the learned DDC feature maps. For simplicity, the benchmark method is referred to as the Fit-to-end method (F2E). All experimental setups and the detailed network architecture are shown in Figure 5.1.

Training Details

Experiments were run using 10-fold cross validation (CV) with 80% training data, 10% validation data, and 10% inner loop test data, where the validation data was used for hyper parameter search and the corresponding loss as a stopping criterion. Batches were generated by randomly sampling 25 slices of each target class. The size of the input layer

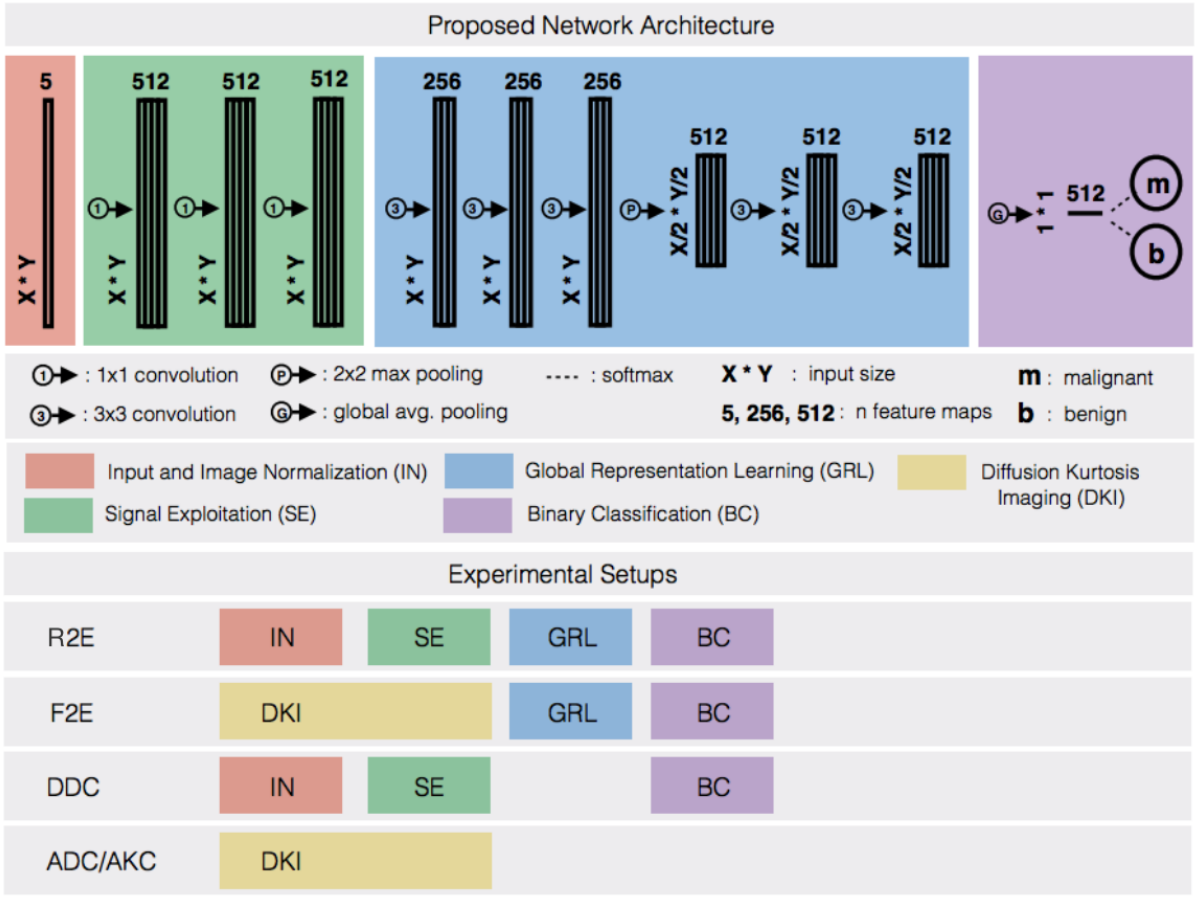

Figure 5.1.: **Proposed CNN Architecture and Experimental Setup.** Detailed network architecture for the different experimental setups explored in this paper. All convolutional layers are followed by ReLU activation functions. The 5 input channels receive the diffusion-weighted images at four b-values plus the fat intensity map. R2E is the proposed ent-to-end Convolutional Neural Network (CNN) (not counting the reliance on manual lesion localization in form of ROI annotations.), F2E is the fit-to-end baseline, where a CNN is run on the output of the biophysical model fit parameters. DDC denotes a linear classifier run directly on top of the learned signal fit outputs, the Deep Diffusion Coefficient (DDC). ADC and AKC are medians over the respective coefficient maps resulting from the biophysical model fit turned into univariate decision rules as applied in clinical practice.

was chosen according to the size of the largest lesion in the batch, while zero padding the smaller lesions. Note, that this method results in batches of variable shape, which is accounted for by the model's all convolutional architecture. All images were masked according to the segmented lesion, i.e. voxels outside the ROI were set to 0. Data augmentation was performed batch-wise by randomly mirroring or rotating. Dropout was applied to all convolutional layers with $p = 0.5$. The learning rate was initialized at $lr = 0.0005$ and decreased each epoch by a factor 0.985. The model was trained using categorical cross-entropy loss over 12 epochs, processing 100 batches per epoch. Inference was done by processing each slice j of a patient i individually and weighting the obtained predictions $p_{i,j}$ with the number of voxels $v_{i,j}$ in the slice against the overall number of voxels in the lesion v in order to obtain the prediction p_i for a patient:

$$p_i = \frac{1}{v} \sum_{j=1}^{s} p_{i,j} * v_{i,j} \tag{5.1}$$

An ensemble of fifteen networks was trained for each fold of the CV and the resulting p_i for one patient were averaged for the final ensemble prediction.

Statistical Evaluation

Models were compared by evaluating the AUC on the test set. The decision threshold t_c was chosen at sensitivity = 0.96. This relatively high threshold matches the sensitivity of core-needle biopsy as reported in literature [154], thus ensuring the integrative character of DWI as a follow-up study of mammography. The resulting specificities at t_c, i.e. the percentage of removed false positives from the test set, were tested for significance using the McNemar-Test. Note, that statistics were calculated across all CV folds, i.e. test set predictions of each fold were collected and fused to a final test set containing all patients.

5.5. Results

Table 5.1 shows a comparison among all methods explored in this study. On the studied dataset with 100 benign lesions (false positive mammographic findings) and 122 malignant lesions (true positive mammographic findings), the decision threshold is set for all methods to a sensitivity of 0.967, which corresponds to correctly identifying 118 out of the 122 true positives. The R2E approach shows best performances with an accuracy of 0.815 ± 0.0260 and a specificity at t_c of 0.630 ± 0.0480, correctly identifying 63 of the 100 false-positives. This significantly (p-value < 0.01) improves the clinical decisions with respect to the F2E method, which has an accuracy of 0.743 ± 0.0292 and a specificity at t_c of 0.470 ± 0.0496, correctly identifying 47 of the 100 false-positives. Comparing classification performances of the coefficients without additional representation learning, the DDC show the highest accuracy of 0.770 ± 0.0281 and the highest specificity at t_c of 0.530 ± 0.0496 outperforming

Table 5.1.: Test **Data Results of the CNN for Lesion Classification Including All Methods Explored in this Study.** AUC denotes Area Under the Receiver Operating Characetristic Curve, Spec. denotes specificity, Sens. denotes sensitivity, and t_c denotes the threshold value on the output softmax probability. R2E is the proposed ent-to-end Convolutional Neural Network (CNN) (not counting the reliance on manual lesion localization in form of ROI annotations.), F2E is the fit-to-end baseline, where a CNN is run on the output of the biophysical model fit parameters. DDC is a linear classifier run directly on top of the learned signal fit outputs, the Deep Diffusion Coefficient (DDC). ADC and AKC are medians over the respective coefficient maps resulting from the biophysical model fit turned into univariate decision rules as applied in clinical practice.

Method	AUC	Spec. (Sens.) at t_c	t_c
R2E	$\mathbf{0.907} \pm 0.038$	$\mathbf{0.630} \pm 0.048$ (0.967)	≥ 0.40
F2E	0.886 ± 0.043	0.470 ± 0.050 (0.967)	≥ 0.34
DDC	$\mathbf{0.868} \pm 0.043$	$\mathbf{0.530} \pm 0.050$ (0.967)	≥ 0.29
ADC	0.827 ± 0.056	0.450 ± 0.050 (0.967)	≤ 1.83
AKC	0.799 ± 0.056	0.450 ± 0.050 (0.967)	≥ 0.85

Table 5.2.: **Test Data Results of the CNN vs. Radiomics study on Breast Lesion Classification.** CNN corresponds to the proposed Roi-to-end Convolutional Neural Network (CNN) identical to R2E in Table 5.1. Radiomics denotes the pipeline from Chapter 4, where a Random Forest is trained on top of Radiomics features extracted from the DKI model. Ensemble denotes the ensemble of the two.

Method	AUC	Spec. (Sens.) at t_c	t_c
Ensemble	$\mathbf{0.927} \pm 0.032$	$\mathbf{0.690} \pm 0.048$ (0.967)	≥ 0.42
CNN (R2E)	0.907 ± 0.038	0.630 ± 0.048 (0.967)	≥ 0.40
Radiomics	0.905 ± 0.041	0.600 ± 0.048 (0.967)	≥ 0.33

the model-based coefficients ADC and AKC, which both show an accuracy of 0.734±0.0295 and a specificity at t_c of 0.450 ± 0.0495.

Comparison against Radiomics

While the previous experiments focus on revealing potentials sealed by the biophysical model fit by means of a CNN (corresponding to R2E in Table 5.1), the question remains as to whether the introduction of representation learning into the model yields performance superior to the handcrafted Radiomics feature extraction such as in Chapter 4. While theory suggests, that feasibility of learning representations scales with the amount of provided training data, in contrast to a fix set of measurements such as in Radiomics, it is unclear whether the dataset utilized in this study is sufficiently large in order to observe the expected effects. Further, we are interested in whether the representations learned from the task-specific data contain complementary value over the handcrafted set of task-agnostic measurements. The Radiomics pipeline is identical to the one in Section 4.3.1, i.e. a Random Forest trained on top of Radiomics features extracted from the DKI model, except for retraining on the updated train test splits described in Section 5.2. In order to study the complementary value of learned representations, we compare against an ensemble of CNN and Radiomics, where we train the two methods separately and subsequently average the two corresponding output scores for each patient. Results are shown in Table 5.2. Output histograms of all compared methods are shown in Figure 5.2 and ROC curves are shown in Figure 5.3. Both in AUC and in specificity at a predefined cut-off sensitivity $> 96\%$, the ensemble performed superior to Radiomics and CNN. On the same metrics, no significant differences were found between Radiomcis and CNN, however, at low sensitivities $(0 - 50\%)$ the CNN showed higher separation power than Radiomics, while at a sensitivity spectrum of $80 - 90\%$, separation power was lower.

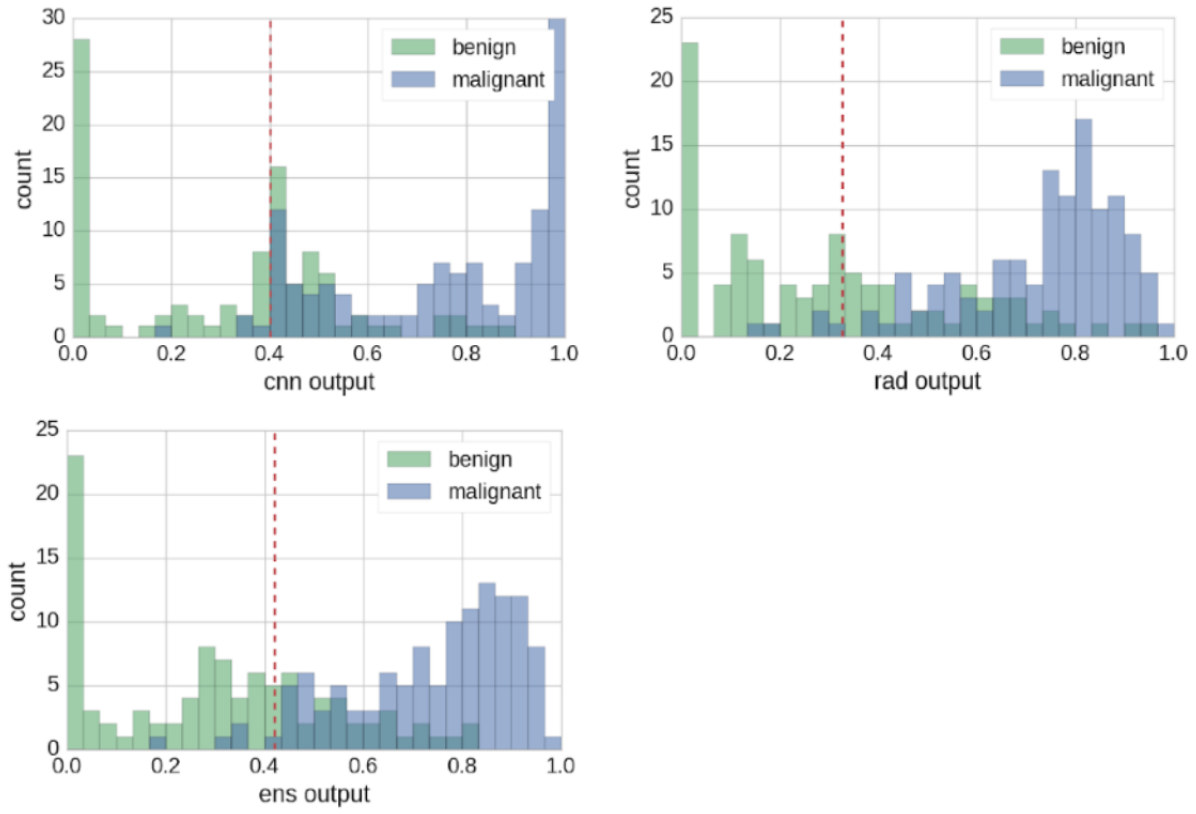

Figure 5.2.: **Output Histograms of CNN vs. Radiomics Experiments.** Output histograms of Radiomics (rad), the CNN (=R2E) and the ensemble of the two (ens) including the resulting cut-off values t_c (red dashed lines).

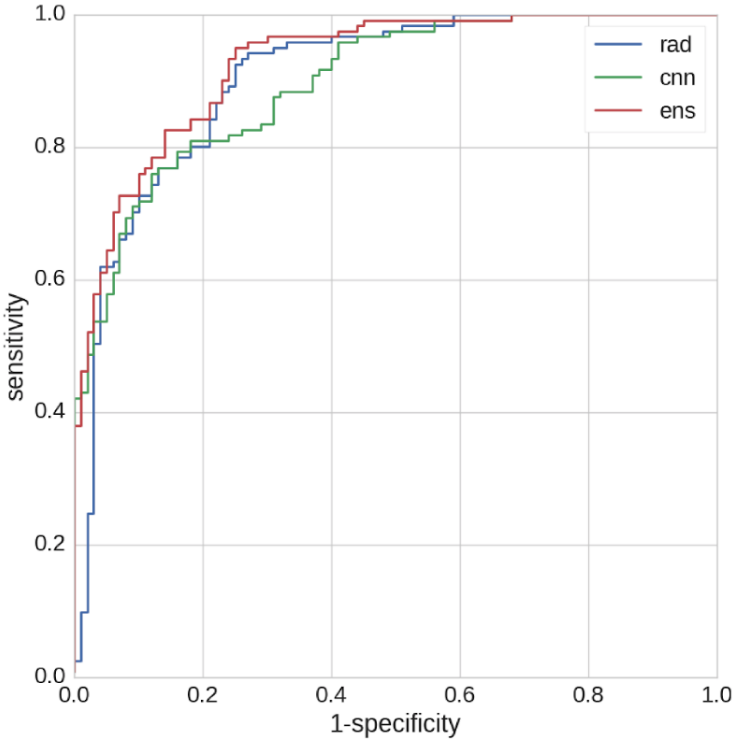

Figure 5.3.: **ROC curve for CNN vs. Radiomics Experiments.** ROC curve displaying sensitivity (true positive rate) over 1-specificity (false positive rate) for Radiomics (rad), CNN (=R2E) and the ensemble of the two (ens).

5.6. Discussion

This study's results show that our CNN approach significantly improves clinical decision making compared to the current clinical state of the art of fitting biophysical models to the data. We first demonstrate, how data-driven signal exploitation in DWI, i.e. our learned Deep Diffusion Coefficient (DDC), outperforms the current model-based methods and show in a second step how this approach can be integrated into a CNN architecture. On the dataset of breast DWI, the Roi-to-end training is able to prevent an additional 16 out of 100 women from overdiagnosis with respect to the benchmark method, where we trained a CNN on top of model-based coefficients ADC or AKC. This benchmark aims at a clear credit assignment for improvement of our approach to the data-driven signal exploitation and its integrability to joint optimization. In contrast to recent data-driven methods like q-space deep learning, which are trained to mimic model-related parameters, our Roi-to-end training is optimized directly with respect to the targeted clinical decision. This enables our approach to optimize all components of the data processing pipeline simultaneously on a specific task, thus not being limited by model assumptions. A limitation to our approach is the dependence on manual segmentation of lesions, which will be addressed in Chapter 6. The multi-centric character of the utilized dataset hints upon the generalization and normalization abilities of the method across different input characteristics.

As for the comparison against the Radiomics approach from Chapter 4 (see Figure 4.1), the performance differences across sensitivity spectra indicate a complementary nature of learned CNN representations w.r.t Radiomics features. This results in a considerably superior performance for the ensembled model. The similar standalone performances of CNN and Radiomics do not confirm the findings of previous studies conducted on mammograms [166] and contrast enhanced MRI [167], where Radiomics outperformed CNN by far. The irritating results in these studies might be related to the fact, that neither of the two studies adapted the CNN architecture to the specific problem, but either used a large fully connected model [166], which is prone to overfitting on small datasets, or deployed an AlexNet pretrained on ImageNet for feature extraction [167]. The latter is a transfer learning appraoch, which are subject to a controversial discussion in the field of medical image analysis: There is strong evidence, that the domain gap between medical datasets and natural iamge datasets such as ImageNet are too large in order for generalizing effects to occur such as aimed for with transfer learning [168]. Remarkably, the achieved AUC performances of all explored models are superior to the respective models of the contrast enhanced MRI study [167], which hints upon the great potential of DWI in breast imaging. Possible reasons for the relative improvement of our CNN over Radiomics compared to the mentioned studies are the Roi-to-end training from scratch, the adapted light-weight network architecture, but also dataset specific properties hampering Radiomics perfor-

mance on this dataset such as adversely impact of small lesion sizes to advanced feature groups such as shape or texture. The fact that the CNN does not considerably outperform the Radiomics baseline hints upon the limited training data available in this study. Following the premises of end-to-end learning, we expect the CNN performance to scale with growing datasets and thus naturally outperform Radiomics baselines in the future.

6. End-to-end Medical Image Classification

While the previous two chapters studied the potential of learning algorithms in clinical diagnosis systems, one important part of the pipeline was not taken into account: The localization of Regions of Interest (ROIs) in the image. The Random Forest model based on Radiomics features presented in Chapter 4 as well as the CNN presented in Chapter 5 relied on previous manual annotation of lesions (or previous automated annotation of cardiac structures in the case of cardiac diagnosis, see Section 4.1.2). This scenario, however, fails to significantly alleviate the current workload related to human readout of medical images, which is to a large extend assigned to location of ROIs [169]. Further, framing localization as an extra step is not fully in line with the dogma of end-to-end learning, meaning in this case localization is not part of the simultaneous optimization process with respect to the ultimate clinical task of clinical diagnosis.

When including the task of localizing ROIs into the learning process to enable true end-to-end diagnosis starting at the raw images, there are three current deep learning methodologies to be considered that attend to the problem at three different levels of granularity: Whole Image classification for patient level decisions, object detection for object level decisions, and semantic segmentation for pixel-level decisions (see Section 3.3). These three levels of image classification translate to specific model evaluation metrics and in return answer to different clinical questions. We identify a largely neglected predicament between the strive for crossing the AI chasm by evaluating models at clinically relevant scales on one side, and optimizing for efficient training under the burden of data scarcity on the other side. The main contributions in this chapter are:

- We propose a deep learning model that enables end-to-end object detection and classification on medical images by aligning the model output to the clinically relevant scale while maintaining data efficient training [37].

- We provide an in-depth analysis of prevalent models from object detection, semantic segmentation and instance segmentation operating in 2D as well as 3D by means of comparative studies on Breast DWI, CT of the Lung and a series of toy experiments.
- We open source the Medical Detection Toolkit, the first comprehensive framework for object detection on medical images including e.g. modular implementations of all explored models operating in 2D and 3D [38].
- We apply our approach to the task of lung cancer staging on Positron Emission Tomography - Computed Tomography (PET-CT) and perform a sensitivity study under varying clinical training scenarios [39].

This work was presented as an oral at the Machine Learning for Health Workshop at Neurips 2019 and is scheduled to be published in the Proceedings of Machine Learning Research [37]. The associated code framework, the Medical Detection Toolkit, is available at `https://github.com/mic-dkfz/medicaldetectiontoolkit` . The applied study on Lung PET-CT was awarded a Research Seed Grant on behalf of the European Society of Radiology and is currently in submission [39].

6.1. Problem Statement

When transforming raw images to clinical decisions, it is important to consider that there exist three *levels of granularity* in classification of medical images. This section builds on the introduction of the different levels provided in Section 3.3 and discusses the associated implications for researchers and clinicians. Figure 6.1 visualizes the described setup and is recommended to be followed along with the text.

While choosing the appropriate model type for a specific task might seem obvious by simply following the classification scale of the associated clinical interest (such as depicted in Figure 3.1), in practice, this decision is distorted by nuisance factors such as the scale of annotations in the provided training data. Since the three model groups are defined by the scale of their output, the scale of associated training annotations is inherent to the respective group. Thus, whole image classification is trained with single scalar labels, object detectors is trained with bounding boxes and respective category scores, and semantic segmentation models are trained with pixel-wise annotation maps. As a result, there exists a ranking in data efficiency between the three model types following the spatial resolution of respective annotations: While pixel-wise annotations provide explicit supervision for the task of localizing objects in images, bounding boxes only represent rough location estimates, and scalar labels provide no spatial information and thus leave the localization task implicitly to the learning process, which drastically increases the amount of required training data (see Section 3.2).

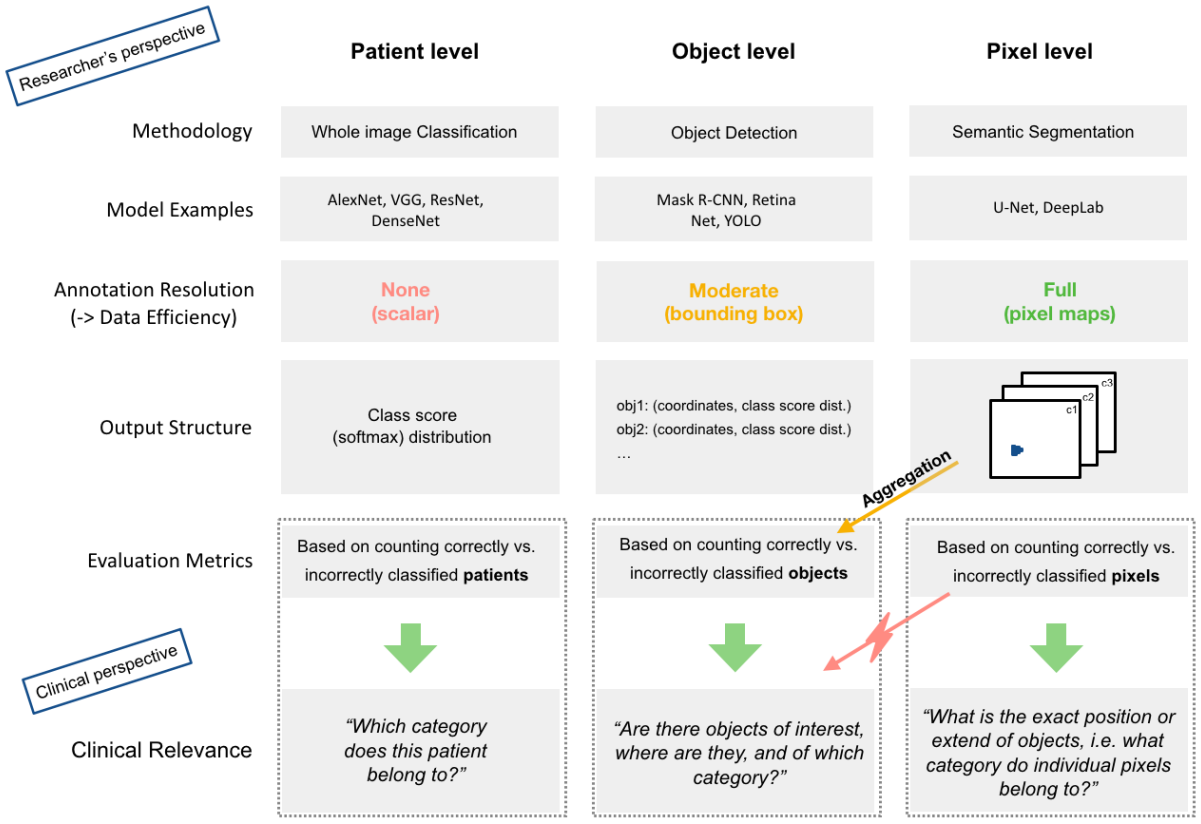

Figure 6.1.: **Conflicts of Choosing a Level of Granularity for Classification in Medical Images.** This figure builds upon Figure 3.1 and visualizes the conflict between research factors and clinical factors regarding the questions as to which classification scale a task should be addressed at. While the clinical scale is directly derived from a underlying medical interest, there are several nuisance factors influencing the researcher's decision such as available annotations and the associated impact on data efficiency for the learning process. As a consequence, the clinical interest is often ignored (red arrow), which leads to models being proposed in literature that are evaluated with metrics operating on clinically irrelvant scales (see dashed boxes tying evaluation metrics to clinical questions of the respective scale). A valid workaround in the form of output aggregation, on the other hand, introduces brittle aggregation heuristics (orange arrow) into otherwise end-to-end learning models. References for model examples: AlexNet [98], VGG [99], ResNet [103], DenseNet [104], Mask R-CNN [105], Retina Net [106], YOLO [107], U-Net [108], DeepLab [109].

Thus, incentives in end-to-end medical image classification are currently high to neglect the clinical interest associated with a task and opt for models that are trained on the highest resolution of available training annotations. In our experience, the most common case of a discrepancy between clinical interest and model output scale are segmentation algorithms proposed to address clinical questions at object-level. This observation is in line with the overwhelming popularity of semantic segmentation methods in the biomedical sector [170], which could be related to the fact that delineation of structures in 3D images by radiologists is typically performed on the pixel-level resulting in a large ratio of medical datasets with pixel-wise annotations. Examples for segmentation models applied to object-level tasks can be found in the detection of prostate cancer [171, 172], lung lesions [173, 174], breast cancer [175, 176, 177], liver tumors [178, 179, 173, 180], kidney tumors [173], multiple sclerosis lesions [173], skin lesion [181], or angiodysplasia lesions [182]. To stress the magnitude of this issue, all segmentation models operate and predict on pixel-level, i.e. there is no inherent notion about the concept of objects such as lesion or tumor, which means they are not able to answer questions regarding the existence, location or category of these objects.

The resulting discrepancy of granularity levels in classification between research studies and clinical interest is a vivid manifestation of the AI chasm [24] (as discussed in Chapter 1). While it is technically possible to aggregate model predictions to coarser levels during clinical deployment, i.e. from pixel-level to object-level or from there to patient-level in an additional post-processing step, this only superficially closes the gap, because model evaluation and hence model selection lies in the past and hence remains on the finer scale (see red arrow in Figure 6.1). Notably, each classification scale is tied to evaluation metrics operating on the respective scale (see Appendix A). In this context, and as emphasized by several studies recently, evaluation metrics represent the "glue" that connects research to future real life application and hence should be given the highest priority when designing a research project and the corresponding model [27, 129, 130] (see dashed boxes in Figure 6.1). One valid strategy in order to circumvent the discrepancy of classification scales accurately is to incorporate the aggregation of output predictions into the research study and directly evaluate the model on the scale of clinical interest (see orange arrow in Figure 6.1). However, this additional post-processing step introduces ungraceful and brittle heuristics into an otherwise end-to-end trained learning system. Figure 6.2 shows an exemplary sketch of the ambiguous decisions faced when aggregating pixel-level predictions to an object-level score. In this study, we put forward the hypothesis that learning to predict objects end-to-end yields performance superior to ad hoc heuristics.

Taken all together, we identify a predicament between the strive for end-to-end object prediction when answering to clinical questions at object-level, and the strive for exploitation

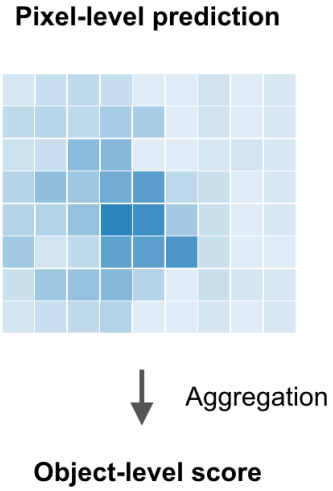

Figure 6.2.: **Aggregation of Pixel-level Predictions to Object-level score.** This sketch is targeted at visualizing the ambiguities faced when attempting to identify connected components (i.e. objects) in a pixel map and assign respective object-level scores (corresponding to the orange arrow in Figure 6.1). In this example, pixels have been classified on a continues scale between foreground (i.e. "belongs to lesion", dark blue) and background (light-blue). The continuous softmax outputs of individual pixels shall now be aggregated to one lesion and given a respective lesion score. What should the intensity threshold be to decide whether a pixel lies inside or outside of the lesion, i.e. how should the connected component be identified? And as a second step, what should the overall object score for this lesion be and how should it be obtained? Assigning the highest value of lesion pixels as the lesion score might lead to a bias towards false-positive predictions over the entire dataset, while assigning the mean or median value across lesion pixels as the lesion score might lead to false negatives. In this study, we hypothesize that there lies a considerable performance gain in learning this process as part of an end-to-end model to circumvent the brittle rule-based aggregation step.

of pixel-wise annotations enabling data efficient training thus alleviating the data burden typically faced in medical imaging. When considering to work around this predicament by aggregating pixel-wise predictions of segmentation models to object scores, we are left with a trade-off between diminished model robustness due to the introduction of brittle rule-based post-processing, or diminished model robustness due to the transformation of pixel-wise annotations to bounding boxes for object detection training.

This study demonstrates a straight-forward yet effective strategy on how to convert available pixel-wise annotations into significant performance gains on object-level detection tasks. To this end, we propose Retina U-Net, which is based on the plain one-stage detector Retina Net [106] (see Section 3.3), complemented by architectural elements of the U-Net, a very successful model for semantic segmentation of medical images [108] (see Section 3.3). From an object detection perspective, the decoding part of Retina Net is complemented by additional high resolution levels to learn the auxiliary task of semantic segmentation and enable data efficient training by exploitation of pixel-wise annotations. From a segmentation perspective, the proposed architecture retrofits the U-Net with two sub-networks operating on coarser feature levels of the decoder part to allow for end-to-end object scoring (and obviating the need for brittle heuristic post-processing). Regarding the choice of a one-stage detector, we argue that the explicit scale invariance enforced by the re-sampling operation in two-stage detectors is not helpful in the medical domain, since unlike in natural images, where object scale is an artifact in form of varying distances between object and camera, in medical images object scale encodes semantic information. We demonstrate the effectiveness of our model on the task of detecting and categorizing lesions on two medical datasets and support our analysis by a series of toy experiments that help shed light on the reasons behind the observed performance gains. Finally, we apply Retina U-Net to the task of lung cancer staging on PET-CT data and study the performance and failure cases under varying training scenarios.

6.2. Related Work

For background information on object detection methodology please see Section 3.3. Since object detection in natural images is increasingly formulated as an instance segmentation problem (see Section 3.3), several two-stage object detectors utilize additional instance-based segmentation labels during training ([105, 183, 122]). However, we argue that this setup does not fully exploit semantic segmentation supervision for the following reasons:

- The mask loss is only evaluated on cropped proposal regions, i.e. context gradients of surrounding regions are not backpropagated.
- The proposal region as well as the ground truth mask are typically resampled to a fixed-sized grid (known as RoIAlign ([105])).

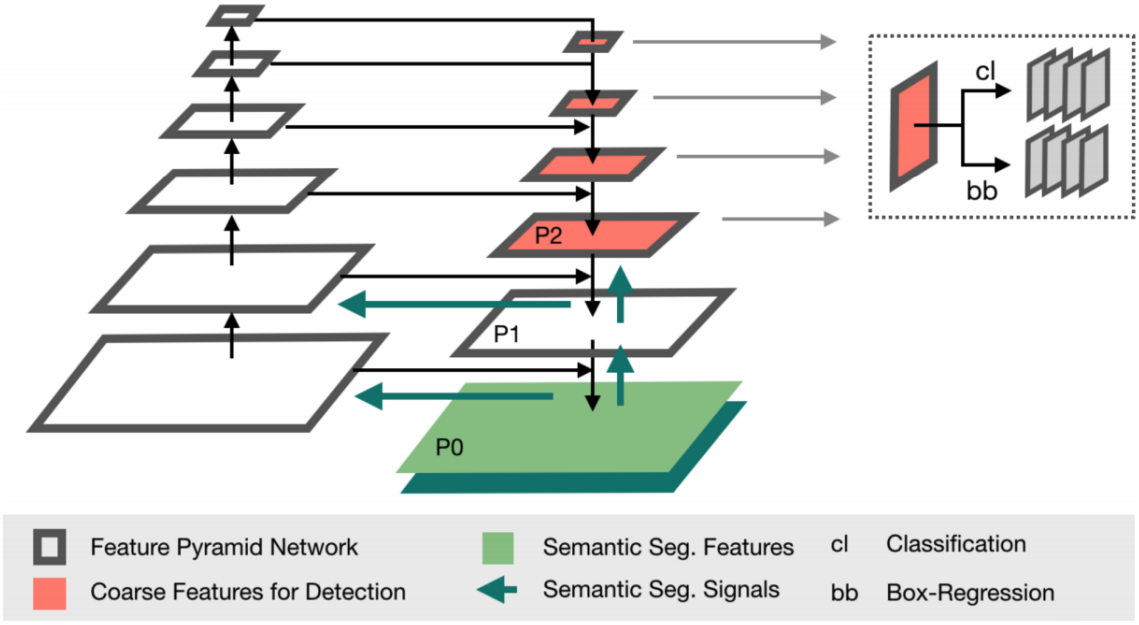

Figure 6.3.: **Retina U-Net Architecture in 2D.** From a Retina Net perspective we add pyramid levels (i.e. feature maps) P0 and P1, so as to backpropagate rich training signals from a full resolution pixel-wise loss into the feature pyramid network, thereby facilitating the learning process in coarser feature maps used for object detection. From a U-Net perspective we enable end-to-end object detection via a head network operating on the coarse pyramid levels P2-P5.

- Only positive matched proposals are utilized for the mask loss, which induces a dependency on the region proposition performance.
- Gradients of the mask loss do not flow through the entire model, but merely from the corresponding pyramid level upwards.

Auxiliary tasks for exploiting semantic segmentation supervision have been applied in two stage detectors with bottom-up feature extractors (i.e. encoders) [184, 185]. In the one-stage domain, the work of [186] performs semantic segmentation on top of a single-shot detection (SSD) architecture for instance segmentation, where segmentation outputs are assigned to box proposals in a post-processing step. [187] propose a similar architecture, but learn segmentation in a weakly-supervised manner, using pseudo-masks created from bounding box annotations.

As opposed to bottom-up backbones for feature extraction, we follow the argumentation of feature pyramid networks [188], where a top-down (i.e. decoder) pathway is installed to allow for semantically rich representations at different scales. This concept is adapted from state-of-the-art segmentation architectures [108, 189] and used in both current one- and two-stage detectors. Recent approaches report to use semantic segmentation as an auxiliary training signal, but apply it to lower resolution feature maps only [190, 191]. [192] learn semantic segmentation at full resolution, but down-sample all feature maps to the bottleneck's spatial resolution before feeding them to the detection module, thereby discarding all multi-scale information in the detection task. In contrast, we propose a one-stage detector based on a Feature Pyramid Network (FPN) performing semantic segmentation on full resolution and object detection on multiple scales, which allows to naturally fuse existing state-of-the-art models from both domains resulting in the simple Retina U-Net architecture. [193] converged to an architecture similar to ours while working on the inverse setup, i.e. enhancing segmentation performance by additionally training on bounding box labels.

6.3. Methods

6.3.1. Retina Net

The basis of our proposed model is the Retina Net, a simple one-stage detector based on a FPN for feature extraction [106], where two sub-networks operate on the pyramid levels P_3-P_6 for classification and bounding box regression, respectively (see Figure 6.4c). Here P_j denotes the feature-maps of the jth decoder level, where j increases as the resolution decreases. In this study, we compare various state-of-the-art one- and two-stage object detectors in both 2D (slice-based) and 3D (volumetric patches). For the sake of unrestricted comparability, all methods including Retina Net are implemented in one framework, using the same backbone FPN [188] based on a ResNet50 [103] as identical

architecture for feature extraction. In our FPN implementation, anchor sizes are divided by a factor of 4 to account for smaller objects in the medical domain resulting in anchors of size $\{4^2, 8^2, 16^2, 32^2\}$ for the corresponding pyramid levels $\{P_2, P_3, P_4, P_5\}$. In the 3D implementation, the z-scale of anchor-cubes is set to $\{1, 2, 4, 8\}$. For Retina Net, two adaptations were made deviating from the original version: To factor in the existence of small object sizes in medical images, we shifted sub-network operations by one pyramid level towards P_2-P_5. This comes at a computational price, since a vast number of dense positions are produced in the higher resolution P_2 level. We further exchanged the sigmoid non-linearity in the classification sub-network for a softmax operation, to account for mutual exclusiveness of classes due to non-overlapping objects in 3D images.

6.3.2. Retina U-Net

Retina Net is complemented with architectural elements from the U-Net, resulting in the proposed Retina U-Net. Training signals for full semantic supervision are added to the top-down path by means of additional pyramid levels P_1 and P_0, including the respective skip connections. The resulting Feature Pyramid resembles the symmetric U-Net architecture (see Figure 6.3), which in the following we refer to as *U-FPN* for clarity. The detection sub-networks do not operate on P_1 and P_0, which keeps the number of parameters at inference time unchanged. The segmentation loss is calculated from P_0 logits. In addition to a pixel-wise cross entropy loss $\mathcal{L}_{\text{CE}}$, a soft Dice loss is applied, which has been shown to stabilize training on highly class imbalanced segmentation tasks e.g. in the medical domain ([173]):

$$\mathcal{L} = \mathcal{L}_{\text{CE}} - \frac{2}{|K|} \sum_{k \in K} \frac{\sum_{i \in I} u_{ik} v_{ik}}{\sum_{i \in I} u_{ik} + \sum_{i \in I} v_{ik}}, \tag{6.1}$$

where u is the softmax output of the network and v is a one hot encoding of the ground truth segmentation map. Both u and v have shape $I \times K$ with $i \in I$ being the number of pixels in the training batch and $k \in K$ being the classes.

6.3.3. Weighted Box Clustering

Medical images typically are comparatively large, due to e.g. very high resolutions as in mammograms, or since they are acquired as 3D volumes like in MRI. Image resolutions are expected to keep rising in the future driven by advances in imaging technologies, such as the recent introduction of 7T MRI scanners. For this reason models are trained on patch crops, resulting in a trade-off between patch size and batch size limited by available GPU memory. If single images exceeds GPU memory, test time inference is performed

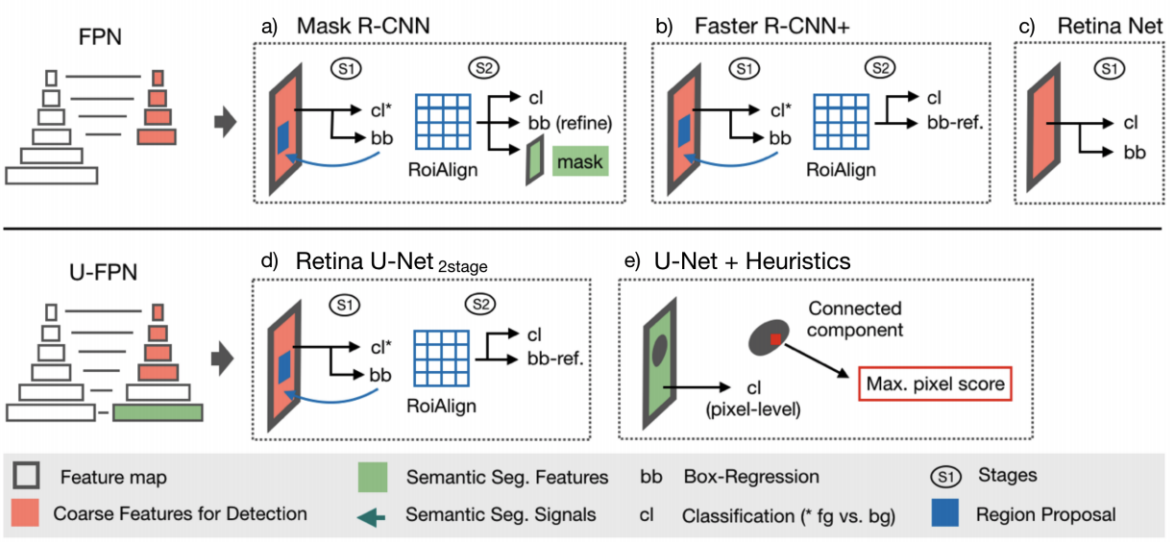

Figure 6.4.: **Baseline Models in Medical Object Detection.** The upper panel shows all baselines utilizing a regular FPN feature extractor while the lower panel depicts baselines that employ a symmetric FPN akin to a U-Net (U-FPN). Subfigures a) - e) show the detection sub networks (heads) that are characteristic of each model and operate on FPN features. All models employ their respective head topology to different decoder scales which are denoted in red. Boxes in green indicate logits that are trained on an auxiliary semantic segmentation task.

patch-wise as well, where tiling strategies are designed to avoid potential artifacts that arise due to effects at the patch boundaries (e.g. by allowing for sufficient overlap between patches). The tiling strategies as well as test time augmentations and model ensembling can amount to a large number of predictions per patch and image (particularly in medical object detection, where validation metrics for model selection are often based on limited validation data, ensembling over multiple selected epochs is able to reduce noise in the process). The resulting predictions for different views of the same image need to be consolidated, which in semantic segmentation is done via simple per-pixel softmax output averaging. For consolidation of predictions in object detection, we propose *weighted box clustering* (WBC): Similar to the commonly used non-maximum suppression algorithm, WBC clusters predictions to be consolidated according to an IoU threshold, but instead of selecting the highest scoring box in the cluster, weighted averages o_c per coordinate and a weighted confidence score o_s per resulting box are computed. Further, the prior knowledge about the expected number of predictions at a position (number of views from ensembling, test time augmentations and patch overlaps at the position) is used to down-weight o_s for views that did not contribute at least one box to the cluster ($n_{missing}$):

$$o_s = \frac{\sum s_i w_i}{\sum w_i + n_{missing} * \bar{w}}, \quad o_c = \frac{\sum c_i s_i w_i}{\sum s_i w_i}, \tag{6.2}$$

where i is the cluster members' index, s and c the corresponding confidence scores and coordinates. $w = f \cdot a \cdot p$ is the weighting factor, consisting of:

- overlap factor f: weights according to the overlap between a box and the highest scoring box (softmax confidence) in the cluster.
- area a: assigns higher weights to larger boxes based on empirical observations indicating an increase in image evidence from larger areas.
- patch center factor p: down-weights boxes based on the distance to the input patch center, where most image context is captured. Scores are assigned according to the density of a normal distribution that is centered at the patch center.

.

6.3.4. Baseline methods.

The following baseline methods were implemented (see Section 3.3 for background details.):

- Mask R-CNN ([105]): Adjustments for the 3D implementation: The number of feature maps in the region proposal network is lowered to 64 to account for increased GPU memory usage. The poolsize of 3D-RoIAlign, a 3D implementation of the

re-sampling operation applied in two stage detectors, is set to (7, 7, 3) for the classification head and (14, 14, 5) for the mask head. The matching IoU for positive proposals is lowered to 0.3 (see Figure 6.4a).

- Faster R-CNN+: In order to single out the Mask R-CNN's performance gain obtained by segmentation supervision from the mask head, we run ablations on the toy datasets while disabling the mask-loss, thereby effectively reducing the model to the Faster R-CNN architecture [194] except for the RoIAlign operation (indicated in the method's name by an additional +) (see Figure 6.4b).

- Retina U-Net$_{2\text{stage}}$: To analyze the gain of additional semantic segmentation supervision in two-stage detectors, we implemented a two-stage variant of Retina U-Net by deploying Faster R-CNN+ on top of U-FPN. (see Figure 6.4d).

- U-Net+Heuristics: Essentially formulating the problem as a semantic segmentation task, as commonly done in medical imaging, we implement a U-Net-like baseline using U-FPN. Therefore, softmax predictions were extracted from P_0 via 1x1 convolution and utilized to identify connected components for all foreground classes. Subsequently, bounding boxes (or cubes) are drawn around connected components and the highest softmax probability per component and class is assigned as object score (see Figure 6.4e). As an alternative to max aggregation, mean aggregation has been implemented without notable changes in performance.

6.4. Experimental Setup

6.4.1. Clinical Studies and Utilized Datasets

Lung Nodule Detection and Classification on CT

A lung nodule detection and categorization task is performed on the publicly available LIDC-IDRI data set [195], consisting of 1035 lung CT scans with pixel-wise lesion annotations and malignancy likelihood scores (1-5) from four experts. We fuse annotations of raters per lesion by applying a pixel-wise majority voting and averaging the malignancy scores. Scores are then re-labelled into benign (1-2: 1319 cases) and malignant (3-5: 494 cases). This is both a difficult and very frequent problem setting in radiology and therefore constitutes a highly-relevant domain of application.

Breast Lesion Detection and Classification on DWI

A breast lesion detection and categorization task is performed on an in-house Diffusion MRI data set of 331 Patients with suspicious findings in previous mammography (an

updated version of the dataset described in Section 4.2, i.e. extended by 109 additional patients). Pixel-wise annotations of lesions are provided by experts. Categorisation labels are given by subsequent pathology results (benign: 141 cases, malignant: 190 cases). Both clinical datasets require the detection and categorisation (*benign* vs. *malignant*) of lesions.

Lung Cancer Detection on PET-CT

While promising results have been reported recently for deep learning-based early detection of lung cancer on large high-resolution CT datasets [196, 197], the detection of advanced tumors, when performed solely based on CT, remains challenging [198]. In clinical practice, it is well recognized that the positron emission tomography (PET) using 18F-fluorodeoxyglucose (FDG) as radiotracer contains unique metabolic information that enables tumor characterization and staging [199, 200] (see Section 2.1.2). Therefore, FDG-PET/CT is currently the status quo for the diagnostic workup of patients with lung cancer as it combines this metabolic information with anatomy and morphology derived from CT [201]. However, to the best of our knowledge to date there has been no study on localizing and classifying lung cancer with end-to-end learning models using the combined FDG-PET/CT protocol. Thus, we aim to explore the performance of Retina U-Net for the detection of T (primary tumor), N (cancerous lymph node) and M-lesions (metastasis, secondary tumor) in patients with primary lung cancer on a FDG-PET/CT dataset of 364 patients, 1913 lesions including pixel-wise annotations, and histologically confirmed tumors of all stages. While the dataset is certainly large for medical standards, the task of sub-categorization of lesions drastically decreases the training cases per category and hence is considered difficult. Thus, for this initial study, we opt for the simpler task of general lesion detection, i.e. grouping lesions into one foreground class to be distinguished from background. This serves as a preparatory study for prospective clinical trials of our algorithm planned at the university hospital Basel. The resulting model could be integrated in to the diagnostic workflow by flagging ROIs to the radiologist for subsequent analysis. Since manual annotation of datasets of the provided size are expensive, in this preliminary study we are interested in whether the annotated training examples of primary tumors (T) can be enhanced by additionally exploiting annotated secondary tumors such as tumorous lymph nodes (N), or metastases (M). Specifically, we aim to examine how the sensitivity of the model behaves when trained under varying foreground definitions. Thereby, the underlying hypothesis is that looking for instance at the detection rate of primary tumors, the inclusion of remaining sub-groups into the foreground class increases the general incentive of the algorithm to predict any candidate as foreground and hence results in a higher number of true positive and false positive predictions for the primary tumor class, and vice versa considering the remaining sub-groups as background is expected to result in a higher number of true negative and false negative predictions. Specifically, we aim to quantify and visualize the extend of these effects and the associated

impact on overall detection performance.

The Dataset comprises 364 FDG-PET/CTs of patients with primary lung cancer acquired at the university hospital Basel. Scans were obtained one hour after intravenous injection of the radioactive tracer (see Section 2.1.2 for background information on Positron Emission Tomography). The CT component of the combined PET/CT examination was acquired with a slice thickness of 3 mm. The TNM classification scheme used during annotation distinguishes four main T-categories (T1-T4; depending on size and features like e.g. invasiveness), three N-categories (N1-N3, depending on location) and one M-category [202]. This scheme follows the official TNM classification, slightly simplified, leaving out the sub-sub categories (e.g. T1a, T1b, etc.). Annotation and 3D image segmentation with reference to the anonymized written PET/CT report was performed manually by a dual-board-certified radiologist and nuclear medicine physician as well as a supervised radiology resident. Each lesion was segmented as a 3D volume defined by multiple 2D regions of interest that were drawn on contiguous transverse sections of the CT component. Fusion with PET information was used in addition whenever the borders of a lesion were not clearly definable on CT.

The mean patient age was 67.1 years (SD: 10.3 years). 72.5% of the patients were male (n = 264), 27.5% were female (n = 100). The complete dataset was randomly split into a training dataset (50%), a validation dataset (25%) and testing dataset (25%). After the randomization process, we checked the distribution of the T-category of the main tumor of each case (T1-T4) to make sure that all categories were represented in all datasets. For the three datasets (training, validation, testing), there were no statistically significant differences in mean volumes of lung tumors ($\chi 2 = 4.81$, $p = 0.090$), malignant lymph nodes ($\chi 2 = 3.12$, $p = 0.211$) and distant metastasis ($\chi 2 = 0.86$, $p = 0.650$). The testing dataset contained n = 87 T-lesions, n = 216 N-lesions and n = 76 M-lesions. Further, there were no statistically significant differences in the data splits of training, validation and testing regarding age ($F = 1.13$, $p = 0.323$) or gender ($\chi 2 = 1.38$; $p = 0.501$). The whole dataset contained 576 T-lesions with a mean tumor volume of 39.9 cm^3 (SD: 100.7 cm^3), 1025 N-lesions with a mean volume of 4.8 cm^3 (SD: 12.4 cm^3) and 312 M-lesions with a mean volume of 6.2 cm^3 (SD: 15.0 cm^3).

Input images were resampled to the most common spacing of 0.97 in planar dimensions and 3.27 in the z-dimension. Intensity values were clipped at [-600, 1200], rescaled to [0,1] and z-score normalized. Images were cropped on the z-axis to slices containing lung tissue according to a lung segmentation based on intensity-thresholding and connected component analysis. We trained three foreground-background scenarios: (A) by defining only T-lesions as foreground, all other voxels as background (T-Model). (B) with T- and N-lesions defined as foreground, all other voxels as background (TN-Model). And (C), with

T-, N- and M-lesions defined as foreground, all other voxels as background (TNM-Mode). We did not differentiate subclasses according to TNM for training.

Toy Datasets

We created a series of three toy experiments to separately evaluate on sub-tasks commonly involved in object-categorisation on medical images (see Figure 6.9). More specifically, the aim is to investigate the importance of full segmentation supervision in the context of limited training data:

1. *Distinguishing object shapes:* Two classes of objects are to be detected and distinguished, circles and donuts (cut-out hole in the middle). Here, the corresponding segmentation mask's shape explicitly contains the discriminative feature (the cut-out hole), hence, full semantic supervision is expected to yield significant performance gains.
2. *Learning discriminative object patterns:* This task is identical to the previous one, except the central hole is not cut out from the segmentation masks of the donuts (class 2). This requires the model to pick up the discriminative pattern without explicitly receiving the respective training signal by means of the mask's shape. This setup could be considered more realistic in the context of medical images.
3. *Distinguishing object scales:* Circles of two different sizes (19 vs. 20 pixel diameter) are to be detected and distinguished. Here, class information is entirely encoded in object scales and hence in target box coordinates. No significant gain from semantic supervision is expected.

Each toy data set consists of 2500 artificially generated 2D images of size 320×320 (1000 train / 500 val / 1000 hold-out test). Images are zero-initialized and foreground objects imprinted by increasing intensity values by 0.2. Subsequently, uniform noise is added to all pixels.

6.4.2. Training & Evaluation Setup.

For comparability, experiments for all methods are run with identical training and inference schemes, as described below. In this study, we compare slice-wise 2D processing, while feeding the ± 3 neighbouring slices as additional input channels (2Dc) [203], against 3D convolutions. Oversampling of foreground regions is applied when training on patch crops. To account for the class-imbalance of object level classification losses, we stochastically mine the hardest negative object candidates according to softmax probability. Models are trained in a 5-fold cross validation (splits: train 60% / val 20% / test 20%) (except Lung Cancer Study on PET-CT which is trained on a single fold) with batch size 20 (8) in 2D (3D) using the Adam optimizer (with default settings) at a learning rate of

10^{-4}. Extensive data augmentation in 2D and 3D is applied to account for overfitting. To compensate for unstable statistics on small datasets, we report results on the aggregated inner loop test sets and ensemble by performing test-time mirroring as well as by testing on multiple models selected as the 5 highest scoring epochs according to validation metrics. Consolidation of box predictions from ensemble-members and overlapping tiles is done via clustering and weighted averaging of scores and coordinates (WBC, see Section 6.3.3). Since evaluation is performed entirely in 3D, an adaption of non-maximum suppression (NMS) is applied to consolidate box predictions from 2D networks to 3D cube predictions: Boxes of all slices are projected into one plane while retaining the slice-origin information. When applying NMS, only boxes with direct or indirect connection to the slice of the highest scoring box are considered as matches. The minimal and maximal slice numbers of all matches are assigned as z-coordinates to the resulting prediction cube. Experiments are evaluated using Mean Average Precision (mAP) [204] (see Appendix A). We determine mAP at a relatively low matching intersection over union (IoU) threshold of IoU = 0.1, which respects the clinical need for coarse localization and also exploits the non-overlapping nature of objects in 3D. Note that evaluation and matching is performed in 3D for all models and processing setups. Additionally, we report the AP of patient-scores, which are determined as the maximum of scores per class and patient (aggregating predictions as well as labels over potentially multiple lesions per patient).

6.5. Results

6.5.1. Detection and Classification of Lung Nodules and Breast Lesions

Results for the lesion detection and categorization tasks are shown in Table 6.1. Retina U-Net performs best on the 2Dc setups (50.2 mAP on and 33.4 mAP on Breast DWI). In 3D, Retina U-Net performs best on Breast DWI (35.8 mAP) and only slightly worse (49.8 mAP) than Retina U-Net$_{\text{2stage}}$ (50.5 mAP) on LIDC. Comparing these results to the remaining baselines shows the importance of semantic segmentation supervision. Mask-R-CNN for instance, which is trained with instance segmentation supervision instead, shows overall worse performance presumably due to the issues discussed in Section 6.2. U-Net performs worse with notable margins, seemingly suffering from high confidence false positive predictions caused by the necessary max-score aggregation (aggregating scores via mean in an alternative experiment did not improve performance). Evaluating on patient-level, Retina U-Net performs best on all tasks except 3D LIDC, where Mask R-CNN achieves the highest score. Example images with model predictions are shown in Figure 6.5.

6.5.2. Detection of Lung Cancer on Pet-CT

The results are shown in Figure 6.8 demonstrating increased sensitivity for primary cancer detection when assigning N-lesions or M-lesions to the foreground class. Surprisingly, the

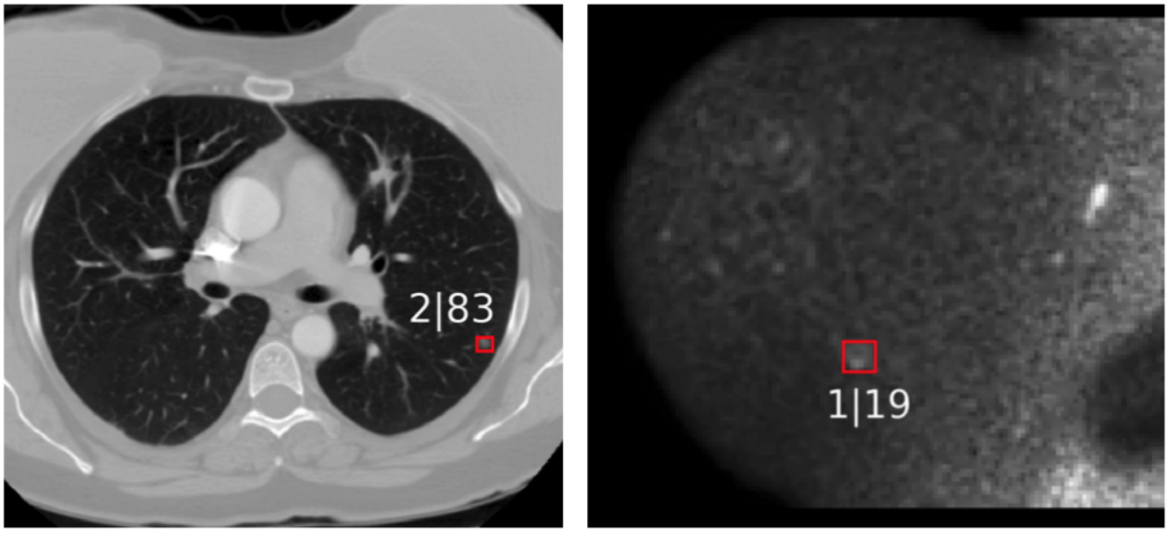

Figure 6.5.: **Example Detections on Medical Images.** Predictions of a malignant lesion in a CT scan of the lung (left) and a benign lesion on a Diffusion MRI of the breast (right). The numbers before and after the vertical bar denote the predicted class-id and the prediction confidence in percent, respectively.

Table 6.1.: **Test Set Results for Lung Lesion Detection on CT (LIDC) and Breast Lesion Detection on DWI.** Mean Average Precision (mAP)reported on the object-level (mAP_{10}) and patient-level (AP_{pat_m}) in [%].

		LIDC		Breast DWI	
Dim.	Model	mAP_{10}	AP_{pat_m}	mAP_{10}	AP_{pat_m}
2Dc	Retina U-Net (ours)	**50.2**	**73.9**	**33.4**	**86.9**
	Mask R-CNN	45.4	69.1	32.3	86.4
	Retina Net	48.2	71.5	33.2	84.4
	Retina U-Net$_{2\text{stage}}$ (ours)	49.1	71.6	33.2	84.7
	U-Net+Heuristics	41.1	66.1	25.8	81.6
3D	Retina U-Net (ours)	49.8	70.4	**35.8**	**88.0**
	Mask R-CNN	48.3	**71.8**	34.0	84.8
	Retina Net	45.9	68.8	31.9	86.4
	Retina U-Net$_{2\text{stage}}$ (ours)	**50.5**	70.7	35.1	86.5
	U-Net+Heuristics	36.6	62.8	26.9	85.1

number of false positives is not increased accordingly when adding M-lesions, indicating an overall improved detection performance of the TNM-model over the TN-model. Notably, drawing conclusions from false positive detections has to be taken with a grain of salt, since the foreground definition of different models itself might be a nuisance factor for this variable, e.g. the additional false positive predictions in the TN-model might be the M-lesions defined as foreground in the TNM-model. This predicament is the reason behind not applying standard detection metrics such as mAP or FROC and opting for a sensitivity study instead, and expected to be resolved in future work by performing false-positive matching, i.e. revealing the original class labels (or the background identity) of all false positive predictions. In summary, the results suggest potential in exploiting additional annotations in datasets if aiming for upstream integration into the diagnostic workflow in form of sensitive flagging of primary cancer ROIs. Qualitative examples of true positive detections are shown in Figure 6.6, while Figure 6.7 provides examples of false positive predictions.

6.5.3. Toy Datasets

Results for the toy experiments are shown in Figure 6.10. In the first task, where explicit class information is contained in segmentation annotations, models which optimally leverage those, i.e. Retina U-Net and Retina U-Net$_{2\text{stage}}$, perform best (again, the instance-based segmentation training of Mask R-CNN seems to yield inferior performance presumably due to issues discussed in Section 6.2. The resulting margin increases with decreasing amount of available training data. The second task, where class information is effectively removed from segmentation annotations, shows similar margins of Retina U-Net and Retina U-Net$_{2\text{stage}}$ to other models. This indicates the importance of full semantic segmentation supervision even in implicit setups and shows a particularly strong edge in the small dataset regime, where models that discard this supervision essentially collapse. In the third task, where class information is entirely contained in the target boxes, no gain from segmentation supervision is observed, at least for small training datasets. Surprisingly, two-shot detectors perform better at this task, which seems counter-intuitive given the scale-invariance enforced by the RoIAlign operation. We hypothesize, that discarding the spatial scale effectively forces the optimizer to encode this information into the feature maps previous to the RoIAlign operation, which seems to not hamper performance and possibly even cause a beneficial regularization effect. Comparing Mask R-CNN to Faster R-CNN+, the sub-optimal mask-supervision seems to yield no gains in detection performance when working with limited training data.

6.6. Medical Detection Toolkit

Medical Detection Toolkit is a comprehensive code framework based on the study presented in this chapter addressing object detection on medical images [38]. It is publicly

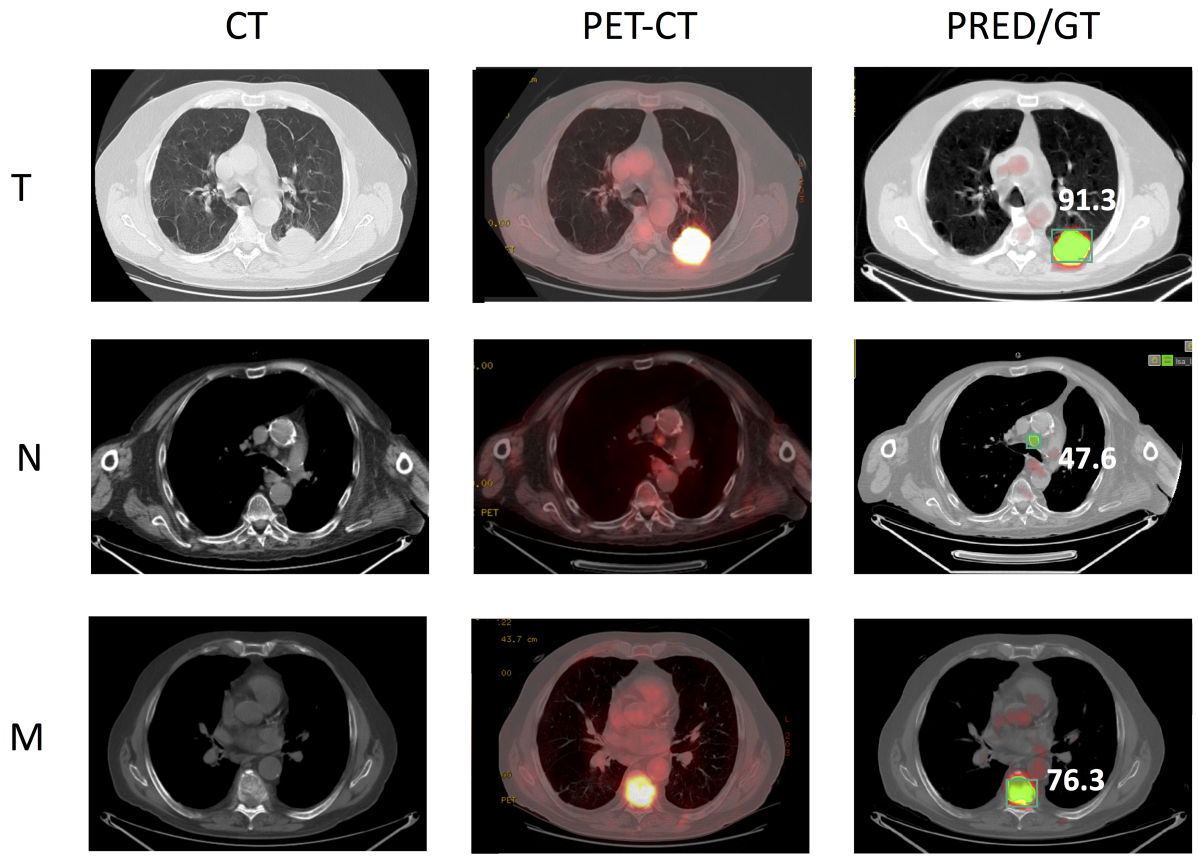

Figure 6.6.: **Qualitative Results for Lung Cancer Detection on PET-CT.** This figure visualizes true positive predictions of the TNM-model showing the raw CT input (left column), the CT input overlayed with PET (middle column) and the overlayed input with the model prediction (green box and confidence score in white) and the pixel-wise ground truth annotation (green mask). The top row shows a solitary lung tumor in the superior segment of the left lower lobe with invasion of the chest wall (T3). The middle row shows a metastatic lymph node (N) in the mediastinum (the central compartment of the thoracic cavity). The bottom row shows a bone metastasis in the 10th vertebra of the thoracic spine (M).

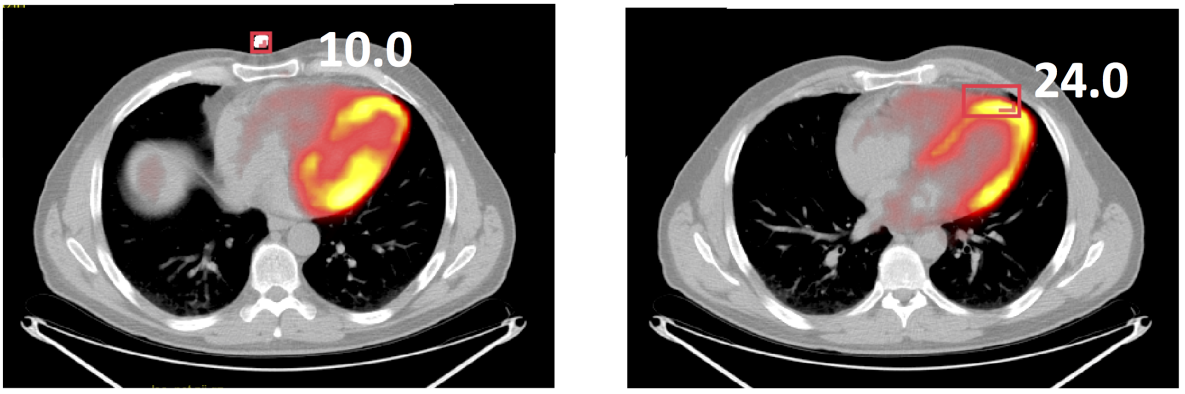

Figure 6.7.: **False Positive Predictions in Lung Cancer Detection on PET-CT.** This figure shows two false positive predictions with relatively low confidence threshold. The left figure shows a prediction on an extracorporal foreign body, while the right figure shows a physiological metabolism of the cardiac muscle.

available at `https://github.com/mic-dkfz/medicaldetectiontoolkit`, where it enjoys a fair amount of popularity due to multiple exclusive features with respect to other existing object detection repositories: To the best of our knowledge, the framework holds the first 3D implementations of prevalent object detection and instance segmentation models such as Retina Net [106] or Mask R-CNN [105]. Moreover, it is designed to handle large 3D images that exceed GPU memory, i.e. it features automated patching for training as well as tiling and stitching for test time inference including aggregation of box predictions over different overlapping patches or tiles (see Weighted Box Clustering in Section 6.3.3. Figure 6.11 visualizes the corresponding prediction pipeline. As for model evaluation, Medical Detection Toolkit integrates the official metric computation of the COCO object detection challenge [204].

6.7. Dicsussion

In this chapter, we addressed the task of end-to-end object detection and classification on medical images. Thereby, we identified a predicament between aligning the output structure of predictive models to the requirements of the intended clinical task and maintaining data efficient training on small datasets - a vivid example for the challenges faced when aiming for successful clinical application of deep learning systems. We saw, that the encountered predicament is currently worked around by either applying subpar post-processing heuristics or simply neglecting the classification scale of the clinical interest, which results in clinically irrelevant evaluation metrics and fuels the gap between research in medical image analysis and translation to real life applications. Thus, we vouch for considering clinical relevance an integral part of research project design including the choice

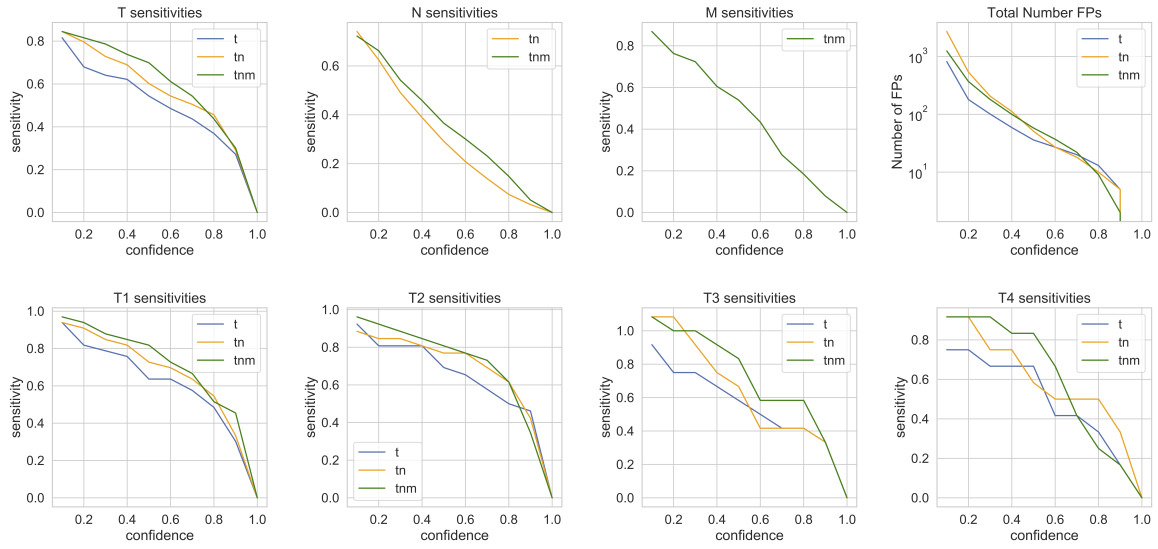

Figure 6.8.: **Sensitivity Study for Primary Lung Cancer Detection on PET-CT.** All plots show sensitivities when scanning over the confidence score (softmax probability of foreground class) of all test set predictions for different models, except the plot on the upper right corner, which shows the total number of false positives over confidence. The first three plots show sensitivities for the three tumor classes Primary Tumor (T), Tumorous Lymph Nodes (N), or Metastases (M), while the plot on the. The four plots on the bottom show sensitivities for the distinguished the subclasses of primary tumors T1-T4. The three different models evaluated are the T-model (only trained with primary tumors as foreground class), TN-model (trained with T and N as foreground class), and the TNM-model (trained with T, N, and M lesions as foreground class).

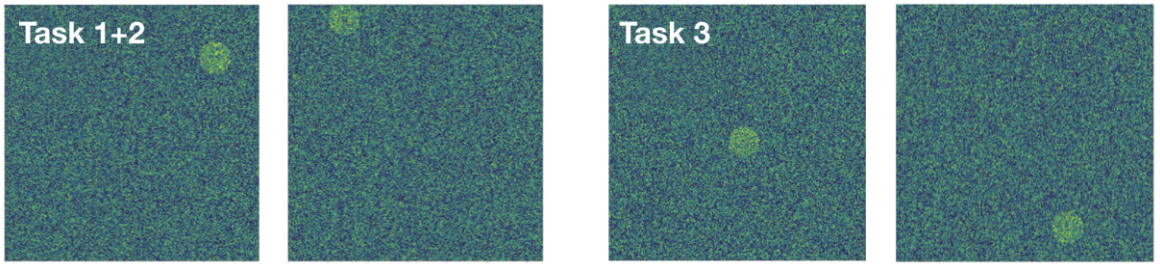

Figure 6.9.: **Example Images of the Toy Dataset for Medical Object Detection.** *left:* Example images for tasks 1 (distinguishing object shapes) and 2 (learning discriminative object patterns) of the toy experiment series. The left object is a filled circle, while the object on the right is a donut (cut-out hole in the middle). *right:* Example images for task 3 (distinguishing object scales), the circle on the left has a diameter of 19 pixels, as opposed to 20 pixels diameter of the circle on the right.

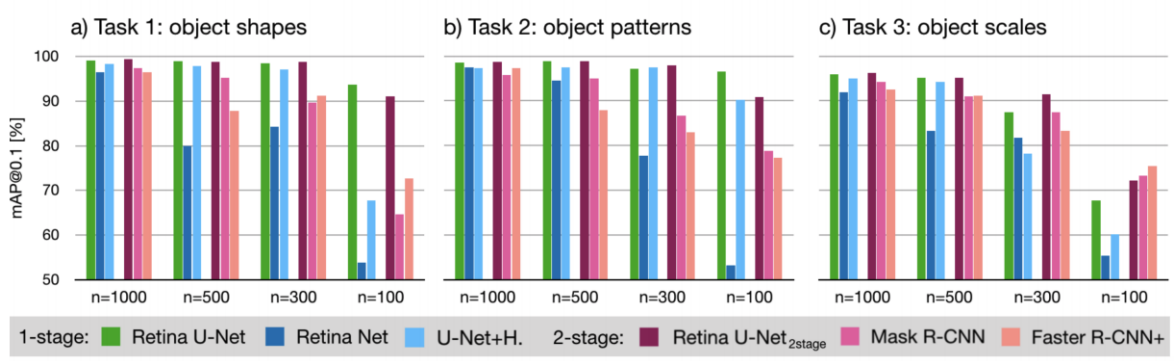

Figure 6.10.: **Results of the of toy experiment series.** The three tasks are displayed as (a) distinguishing objects of different shapes, (b) learning discriminative image patterns unrelated to an object's shape, and (c) distinguishing objects of different scales. Explored models are divided into two groups: One-stage methods have blue/green color, while two-stage methods are drawn in red. n denotes the number of training images.

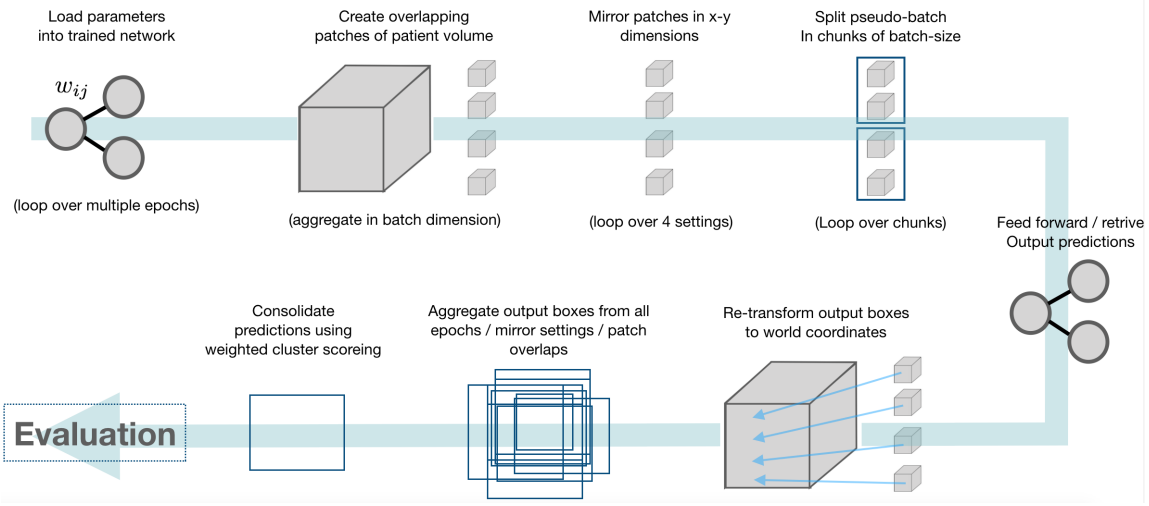

Figure 6.11.: **Prediction Pipeline of the Medical Detection Toolkit.** Images exceeding GPU memory are automatically tiled in overlapping 3D patches. Prediction instances of the same regions can amount to large numbers due to test time augmentations, ensembling strategies and patch overlaps. The resulting list of box predictions per image position are consolidated using the Weighted Box Clustering algorithm described in section 6.3.3.

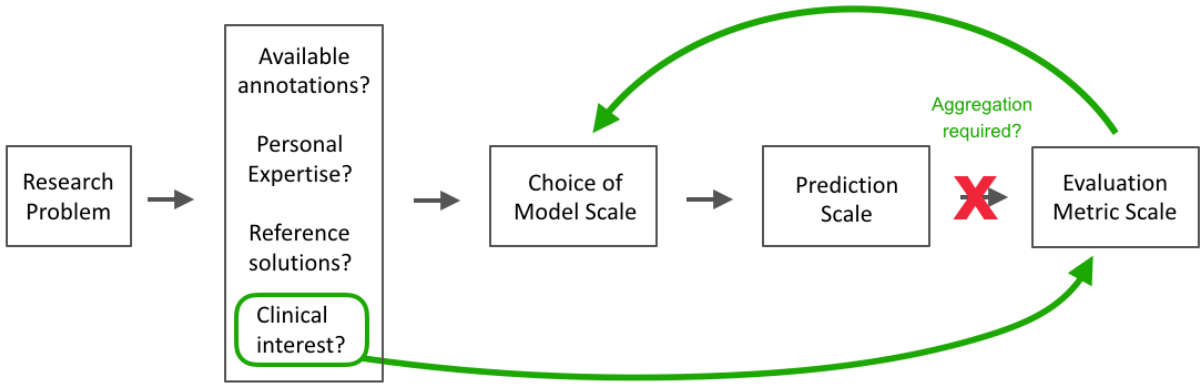

Figure 6.12.: **Stereotypical Research Workflow in Medical Image Analysis and Proposed Modifcations.** This Figure follows the scenario depicted in Section 6.1 and (somewhat cheekily) draws a stereotypical process of choosing the classification scale when designing a research model (black arrows): Faced with a clinical task, there are multiple factors influencing the choice of methodology, including the amount and spatial resolution of annotations in the training data, the personal expertise or experience with certain types of models, existing reference solutions. The classification scale of the clinical interest behind the task is often just one more factor to consider. The resulting methodology outputs predictions on the corresponding scale and the subsequent evaluation is often performed with metrics chosen to match this output scale, which results in diminished clinical relevance of the study results. However, in order to increase the overall clinical relevance of research studies, we consider it crucial to instead first define the scale of evaluation metrics according to the clinical interest and consider the pre-defined metrics an additional factor in the subsequent decision for model classification scale (red cross and green arrows). If the resulting model prediction scale happens to mismatch the scale of evaluation metrics it is necessary to aggregate the predictions accordingly before evaluation.

of methodology. Figure 6.12 visualizes the discrepancy between a stereotypical research workflow and the modifications deemed necessary towards crossing the AI chasm.

Following this narrative, we address the task of end-to-end lesion detection and classification under limited training data and propose Retina U-Net, a simple but effective method for leveraging segmentation supervision in object detection. We showed the importance of exploiting pixel-wise training signals at full resolution on multiple datasets, input dimensions and meticulously compared against the prevalent models in object detection, semantic segmentation and instance segmentation with a particular emphasis on the context of small datasets.

On the publicly available lung CT dataset as well as on the breast lesion DWI dataset,

Retina U-Net yields detection performance superior to models without full segmentation supervision. By means of a set of toy experiments we shed light on an important set of scenarios that can profit from the additional full supervision: Any such problem where there is discriminative power in features beyond mere scale can expect to pocket an edge in detection performance. In the preliminary sensitivity study on PET-CT, Retina U-Net showed robust detection of primary cancer and we observed indication that the associated performance increases when secondary cancer lesiosn are included into the foreground class during training.

A current limitation of Retina U-Net is that it does not allow for end-to-end patient level detection. Many scenarios e.g. in cancer screening utilize lesion detection on object-level as an explicit localization task to alleviate the data burden during training, but are ultimately interested in decisions on the patient-level (this problem in a sense repeats the general problem statement of this study, which however focused on on the finer-scale level of pixel-wise supervision for object detection tasks.). Currently, we required heuristics to aggregate object-level scores to patient-level scores as reported in Section 6.4.2. In theory, Retina U-Net could be extended to allow for end-to-end patient-level prediction while exploiting supervision on object-level and pixel-level in order to prevent overfitting due to sparse training signals from patient-level labels. This could be implemented by branching off a patient-prediction path in the bottleneck of the Feature Pyramid Network (FPN). In practice, however, this model clashes with a further nuisance factor affecting the choice of classification scales for clinical tasks: Hardware constraints in form of e.g. limited GPU memory. Since lowering the resolution of input images is expected to drastically hamper detection of small objects, models are trained on patches cropped out of the whole image at full resolution. This is an acceptable workaround, as long as it is possible to assign meaningful labels to patches, such as on pixel-level or object-level, but introduces incongruous gradients when patches are trained with patient-labels (e.g. a background patch might be associated with the label "cancer" during training). Further, we argue that the aggregation process of object scores to patient-level entails less noise than aggregating individual pixels [1] and deem our approach the current best practice for patient-level tasks with spatially resolved annotations.

Among other distinguishing characteristics, the domain of medical image analysis holds one prominent feature: scarcity of labeled data. Retina U-Net is designed to make the most of the given supervision signal which is a crucial aspect for robust model generalization on small datasets as high-lighted by our experiments. Taken together, Retina U-Net

[1] Beside the fact that individual pixels are subject to higher intensity fluctuations than aggregated object scores, "patient" as the target of aggregation constitutes one single instance of fixed extend as opposed to the need for defining instances of unknown number and extend when aggregating pixels to objects (such as visualized in Figure 6.2).

aims to enable end-to-end diagnostics by optimizing for model robustness under limited training data while ensuring clinical relevance of model predictions - two key factors in the strive towards closing the gap between research and real life application of end-to-end learning systems. On an anecdotal level, the relevance of addressing this challenge is highlighted by the fact that Turing award winner Yann LeCun mentions Retina U-net as the one application of segmentation in medical image analysis in a recent essay on the status and future of deep learning [205].

7. Increasing Robustness of End-to-end Medical Image Classification

"If biases and initial knowledge are at the heart of the ability to generalize beyond observed data, then efforts to study machine learning must focus on the ***combined*** *use of prior knowledge, biases, and observation in guiding the learning process. It would be wise to make the biases and their use in controlling learning just as explicit as past research has made the observations and their use."*

Tom Mitchell, 1980 [31]

The robustness of learning algorithms is the ability to generalize to unseen data regardless of varying data properties [31]. In the context of notoriously limited annotated training data in the medical domain model robustness is considered the key obstacle towards clinical application of learning algorithms [27]. Following Tom Mitchell's above quote in his report from 1980 and the discussion provided in Section 3.4, one way to increase robustness in learning algorithms under limited training data is to transform prior knowledge into informed inductive biases in order to constrain the space of possible solutions [31]. Components composing a model's inductive bias, i.e. possible injection points for prior knowledge in this context, are architecture, loss function, training data, and optimization [128]. From this perspective, the auxiliary segmentation loss deployed in Chapter 6 transforms task-relevant knowledge from pixel-wise annotations into additional constraints baked into the model during training thus increasing the model's ability to generalize to unseen data. In this chapter, we explore further leverage points in existing end-to-end learning systems, where domain knowledge can be condensed into informed constraints, this time harnessing all four components of the models' inductive bias. The main contributions in this chapter are:

- *Robustness against rater confusion:* We address the challenge of erroneous training annotations by substituting the classification component of end-to-end object detection for regression, which enables to train models directly on the continuous scale of underlying pathological processes [40].
- *Robustness against input variations:* When trained models are deployed across clinical sites they commonly face performance drops due to input domain shifts such as missing or altered modalities. We inject domain knowledge in form of a biophysical model that recaptures the original training domain from altered inputs and thus restores robust model generalization. [41].
- *Robust Hyperparameters:* We address the highly unsystematic, cumbersome and subjective trial-and-error process of finding a robust set of hyperparameters for a given task by condensing domain knowledge into a set of key design choices and systematic rules thus enabling automated and robust deep learning pipeline configuration on a large variety of medical datasets. [42].

The work regarding regression for mitigation of ambiguous annotations was presented at the UNSURE workshop at MICCAI 2019 as a spotlight talk [40] and the associated code is publicly available at `https://github.com/MIC-DKFZ/RegRCNN` . The work regarding domain adaptation towards robustness against input shifts was presented as a spotlight talk at the workshop on Image Analysis for Moving Organs, Breast and Thoracic Images at MICCAI 2018 and elected one of three best paper award nominees [41]. The work regarding automated model design is currently under review at Nature Methods with a preprint available on arXiv [42]. The code is publicized at `https://github.com/MIC-DKFZ/nnUNet` .

7.1. Robustness against Rater Confusion: End-to-End Regression

When adopting state-of-the-art object detectors for end-to-end lesion grading, there is one peculiarity in medical images that is currently unconsidered: The clinical grading of lesions denotes a subjective discretization of naturally continuous and ordered features (such as scale or intensity) to semantic categories with clinical meaning, e.g. the BI-RADS score [112], Gleason score [206], PI-RADS score [207], or TNM staging [126]. This is in contrast to typical tasks on natural images, where categories can be described as an unordered set (no natural ordinal relation exists between dogs and cars). Since state of the art object detection models are commonly developed based on natural images, they phrase the categorization as a classification task and are trained using cross-entropy loss, thus do not consider the continuous ordinal relation between classes.

In this study, we argue that the described underlying relation represents important prior knowledge that should be integrated into the training signal in order to enhance the biases inducted during optimization to the task at hand. Specifically, by considering ordinal class relations, we expect a gain in the models' robustness against ambiguity in training annotations: Medical images often contain highly ambiguous information that reflects in drastic variability of human annotations e.g. in the delineation of prostate cancer [208, 209], breast cancer [210], or lung cancer [211]. Under the assumption that rater confusions follow a distribution around the underlying ground truth grading, distance metrics used in regression such as the L1-distance are expected to be more tolerant to mild deviation from the target value as opposed to the categorical cross entropy which penalizes all off-target predictions in equal measure [212].

To this end, we propose Reg R-CNN, an end-to-end object detection and regression model that builds upon Mask R-CNN [105] by substituting the classification component with a regression component. Thus, training penalties for the categorization task follow a distance metric, i.e. directly operate on the underlying continuous scale. We show the superiority of Reg R-CNN in the task of lesion detection and grading on a public dataset with 1026 patients and a series of toy experiments.

7.1.1. Methods

Regression vs. Classification Training

In order to see why we expect the training of regression models to be more robust to label noise than classification models for the case when target classes lie on a continuous scale, let us first revisit the objective commonly minimized by classifiers. This objective is the cross entropy (CE), defined as

$$H(\mathbf{p}, \mathbf{q}; \mathbf{X}) = -\sum_{j} p_j(\mathbf{X}) \log q_j(\mathbf{X}) \tag{7.1}$$

between a target distribution $\mathbf{p}(\mathbf{X})$ over discrete labels $j \in C$ and the predicted distribution $\mathbf{q}(\mathbf{X})$ given data $\mathbf{X}$. For mutually exclusive classes, the target distribution is given by a delta distribution $\mathbf{p}(\mathbf{X}) = \{\delta_{ij}\}_{j \in C}$. To produce a prediction $\mathbf{q}(\mathbf{X})$, the network's logits $\mathbf{z}(\mathbf{X})$ are squashed by means of a softmax function:

$$\mathbf{q}(\mathbf{X}) = \frac{e^{\mathbf{z}(\mathbf{X})}}{\sum_{k \in C} e^{z_k(\mathbf{X})}}, \tag{7.2}$$

which, plugged into Eq. 7.1 and given the target class i, leads to the loss term

$$H = \mathcal{L}_{CE}(\mathbf{p} = \delta_{ij}, \mathbf{q}; \mathbf{X}) = -z_i + log \sum_k e^{z_k}. \tag{7.3}$$

From Eq. 7.3 it is apparent that the standard CE loss treats labels as an unordered bag of targets, where all off-target classes ($j \neq i$) are penalized in equal measure, regardless of their proximity to the target class i. Distance metrics on the other hand, as their name suggests, take into account the distance of a prediction to the target. This lets the loss scale in the deviation of prediction to target. Allowing to be more accepting of mild discrepancies, it better accommodates for noise from potentially conflicting labels in settings where the target labels lie on a continuum.

In the range of experiments below, we compare classification against regression setups, for which we employed the smooth L1 loss given by

$$\mathcal{L}_{reg}(p,t) = \begin{cases} \frac{1}{2}(t-p)^2, & |t-p| < 1 \\ |t-p| - \frac{1}{2}, & \text{otherwise} \end{cases} \tag{7.4}$$

for predicted value p and target value t. Other works have investigated adaptions to the CE loss to account for noisy labels in classification tasks, e.g. [213, 214]. Our approach is complementary to those works as it exploits label continua on medical images.

Reg R-CNN & Baseline

The proposed Reg R-CNN architecture is based on Mask R-CNN [105], a state-of-the-art two-stage detector (see Section 3.3). In Mask R-CNN, first, objects are discriminated from background irrespective of class, accompanied by bounding-box regression to generate region proposals of variable sizes. Second, proposals are re-sampled to a fixed-sized grid and fed through three head networks: A classifier for categorization, a second bounding-box regressor for refinement of coordinates, and a fully convolutional head producing output segmentations (the latter are not further used in this study except for the additional pixel-wise loss during training). Reg R-CNN (see Fig. 7.1) simply replaces the classification head by a regression head, which is trained with the smooth L1 loss instead of the cross-entropy loss.

For the final filtering of output predictions, non-maximum suppression (NMS) is performed based on detection-confidence scores. In Mask R-CNN, these are provided by the classification head. Since the regression head does not produce confidences, we use the objectness scores from the first stage instead.

In this study, we compare Reg R-CNN against Mask R-CNN as the classification counterpart of our approach. Only minor changes are made with respect to the original publication [105] regarding the adoption to operating on 3D images: The number of feature maps in the region proposal network is lowered to 64 to account for GPU memory constraints. The poolsize of 3D RoIAlign (a 3D re-implementation of the resampling method used to

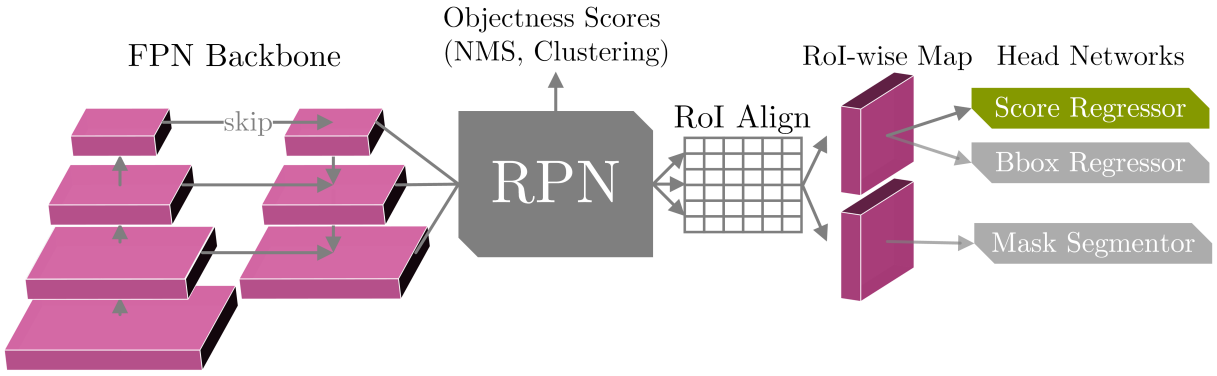

Figure 7.1.: **Reg R-CNN Model for Joint Detection and Continuous Grading of Objects.** The architecture is closely related to Mask R-CNN [105], where grading is done with a classification head instead of the displayed "Score Regressor" head network. FPN denotes Feature Pyramid Network [188], RPN denotes Region Proposal Network (RPN) and RoIAlign is the operation which re-samples object proposals to a fixed-sized grid before categorization.

create fixed-sized proposals) is set to (7, 7, 3) for the classification head and (14, 14, 5) for the mask head. The matching Intersection over Union (IoU) for positive proposals is lowered to 0.3. Objectness scores are used for the final NMS to reflect the desired disentanglement of detection and categorization tasks. All changes equally apply to Reg R-CNN, thus the only difference between the two compared models is the substitution of the classification task with a regression task.

Evaluation

Comparing the performance of regression to classification models requires taking into account additional considerations since both are trained along an upstream detection task. In order to evaluate continuous regression in comparison to discrete classification outputs, we bin the continuous regression output after training, such that bin centers match the discrete classification targets. What's more, the joint task of object detection and categorization is commonly evaluated using Mean Average Precision (mAP) [215] (see Appendix A). However, Mean Average Precision (mAP) requires per-category confidence scores, which are, as mentioned before, not provided by regression outputs. Instead, we borrow a metric commonly used in viewpoint estimation, the Average Viewpoint Precision (AVP) [216]. Based on AVP, we phrase the lesion scoring as an additional task on top of foreground vs. background object detection: In order for a box prediction to be considered a true positive, additionally to the requirement posed by the detection task to match the ground-truth box with an IoU > 0.1 (this relatively low matching threshold respects the clinical need for coarse localization and exploits the non-overlapping nature of objects in 3D images), in AVP the malignancy prediction score is required to lie in the correct

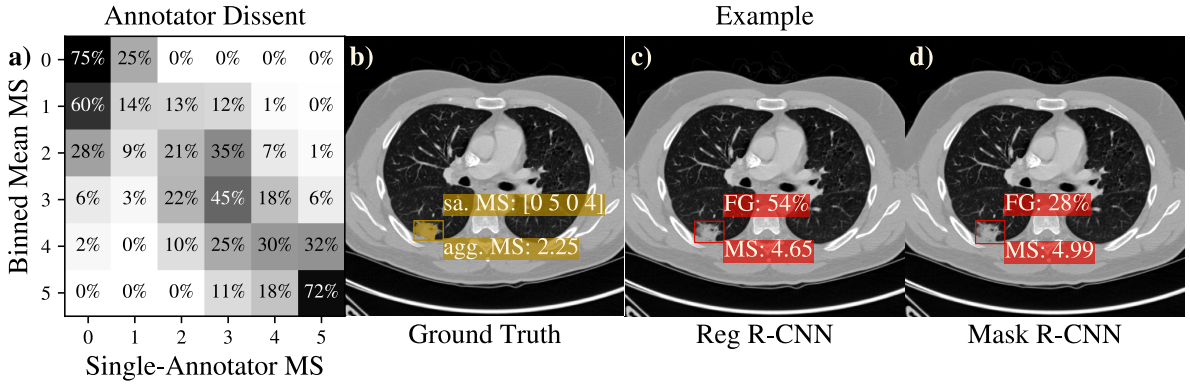

Figure 7.2.: **Visualization of Rater Dissent on the LIDC Dataset. a)** A confusion-matrix-style display of annotator dissent in the LIDC dataset. Rows represent the binned mean ratings of lesions (in place of the true class in a standard confusion matrix), columns the ratings of the corresponding single annotators. "MS" means malignancy score. Matrix is row-wisely normalized, hence cell values indicate distribution of lesion ratings within a bin. **b)-d)** Example slice from the LIDC dataset showing GT, Reg R-CNN, and Mask R-CNN prediction separately. GT note "sa. MS" shows the single-annotator grades (grade 0 means no finding), "agg. MS" their mean. In the predictions, "FG" means foreground confidence (objectness score), "MS" denotes the predicted malignancy score. Mask R-CNN MS can be non-integer due to Weighted Box Clustering 6.3.3. Color symbolizes bin.

category bin. This way, AVP condenses both the detection and malignancy-scoring performance of the model in one metric. We additionally disentangle the two tasks and evaluate separate metrics. To this end, we report the mAP to assess foreground vs. background detection (this poses an upper bound on AVP, since the additional requirement posed by AVP is missing) and the bin accuracy to assess categorization performance. The latter is determined by pre-selecting true positive predictions according to the detection metric and subsequently only consider this selection when counting malignancy-score matches with the target bin.

7.1.2. Experimental Setup

Utilized Datasets

Lung CT dataset The utilized LIDC-IDRI dataset is identical to the one described in Section 6.4.1 [195], except that the four rater assessments per lesions are not fused, but considered independent training cases and the corresponding malignancy scores are not mapped to binary values 0 and 1, but left on the original scale of 1-5. Having disposable multiple gradings from distinct annotators is a rare exception on medical images and allows to investigate the exhibited label noise [217].

Full agreement, which we define as all raters assigning the same malignancy label to all lesions (RoIs) in a patient, is observed on a mere 163 patients (this includes patients void of findings by all raters). This corresponds to a rater disagreement with respect to the malignancy scoring on 84% of the patients. On a lesion level (ROI-wise), the dataset comprises 2631 lesions when considering all lesions with a positive label by at least one rater. This number drops to 1834, 1333, or 821, when requiring 2, 3, or 4 positive labels, respectively. This shows that the labeling is both ambiguous with respect to whether or not a lesion is present as well as the prospective lesion's grading. The former has bearings on the detection head's performance, while the latter influences the network's classification or, respectively, regression head.

In order to evaluate the grading performance, the following malignancy statistics include only patients with at least one finding. Among those, we count 99 lesions (3.8% of all lesions) with full rater agreement, leaving disagreement on 2532 (or 96.2%). The standard deviation of the 4 graders averaged over all lesions amounts to 1.05 malignancy-score values (ms). In Fig. 7.2 a), we show how the single graders' malignancy ratings differ given the binned mean rating. The figure reveals significant label confusion across adjacent labels and even beyond. Figs. 7.2 b)-d) display example Reg and Mask R-CNN predictions next to the corresponding ground truth.

In order to investigate the models' performance under label noise, we randomly sample a rater assessment including pixel-wise annotation and malignancy score for a given lesion from the 4 given ratings at each training iteration. At test time, we however employ the lesions' ground truth label as the mean malignancy score, in order to approximate the "real" ground truth by reducing the noise of inter-rater variability.

Toy dataset Since the underlying ground truth on the LIDC dataset can only be approximated by computing the mean of all provided ratings, as no further clinical assessments such as biopsies are available, we design a toy dataset, where we are able to carefully control the ground truth as well as the associated distributions of rater confusions. The aim is to gain additional confirmation of the underlying hypotheses in this study without the noise of a real life clinical dataset. The artificial task is the joint detection and categorization of 3D cylinders, where five categories are distinguished as cylinders of five different radii. In order to simulate label confusion, Gaussian noise is added to the isotropic target radii during training, sampled with standard deviation $\sigma = r/6$ around object radius r, as depicted in Figs. 7.3 c) and d). This causes targets (especially of large-radius objects) to be shifted into wrong, yet mostly adjacent target bins. Fig. 7.3 a) portrays that these ambiguities are imprinted on the images as a belt of reduced intensity with width 2σ around the actual radius. At test time, model predictions are evaluated against the exact target radii without noise. The dataset consists of 1.5k randomly generated samples for

training and validation, as well as a hold-out test set of 1k images.

Training & Inference Setup

Both the LIDC and the toy dataset consist of volumetric images. In this study, we evaluate instances of the compared models that operate in 3D as well as in 2D (slice-wise processing). For the sake of comparability, all methods are implemented in a single framework and run with identical hyperparameters. Networks are trained on patch crops of sizes $160 \times 160 \times 96$ (LIDC) and $320 \times 320 \times 8$ (toy), oversampling of foreground regions is applied. Class imbalances in object-level classification losses are accounted for by stochastically mining the hardest negative object candidates according to softmax probability. On LIDC, models are trained for 130 epochs, each composed of 200 batches with size 8 (20) in 3D (2D) at a learning rate of 10^{-4}. Training is performed as a five-fold cross validation (splits: train 60% / val 20% / test 20%). At test time, we ensemble the four best performing models from all epochs according to validation metrics and generate four test-time views (three mirroring augmentations) in each fold. Aggregation of box predictions from ensemble members is done via clustering and weighted averaging of scores and coordinates (see Weighted Box Clustering (WBC) in Section 6.3.3). Predictions from 2D models are consolidated along the z-axis by means of an adaption of NMS and evaluated against the 3D ground truth.

7.1.3. Results

Results are shown in Table 7.1. In addition to the fold means of the metrics we report the corresponding standard deviations. On LIDC, Reg R-CNN outperforms Mask R-CNN on both input dimensions and all three considered metrics. On the toy dataset, Reg R-CNN shows superior performance in AVP_{10} and Bin Accuracy. AP_{10} reaches 100% in both models indicating that the detection task is solved entirely, i.e., the object grading task has been isolated successfully (hence, results for AVP_{10} converge towards the Bin Accuracy). All experiments demonstrate the superiority of distance losses in the supervision of models performing continuous and ordered grading under noisy labels. Interestingly, there is a marked increase in performance for both setups when running in 3D as opposed to 2D, suggesting that additional 3D context is generally beneficial for the task.

7.1.4. Discussion

Simultaneously detecting and grading objects is a common and clinically highly relevant task in medical image analysis. As opposed to natural images, where object categories are mostly well defined, the categorizations of interest for clinically relevant findings commonly leave room for interpretation. This ambiguity can bear on machine-learning models in the form of noisy labels, which may hamper the performance of classification models.

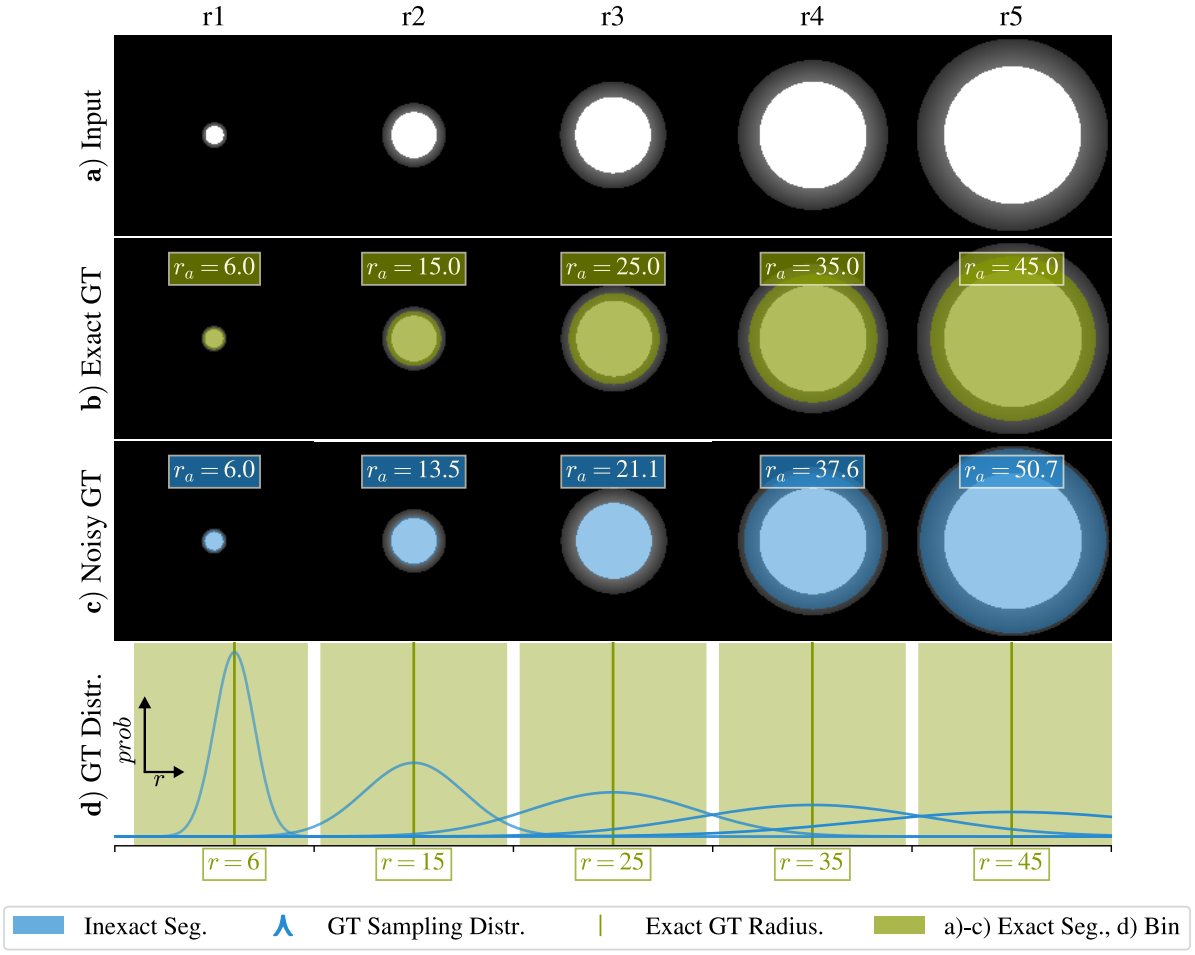

Figure 7.3.: **Visualization of the Toy Experiments for End-to-end Regression. a)** Cylinders (2D projections) of all five categories (r1-r5) in the toy experiment. **b)** Exact GT. **c)** Examples of a noisy GT for each category (r1-r5). r_a indicates the annotated radius (target regression value). **d)** Gaussian sampling distributions used to generate the noisy GT. Green vertical lines depict the exact ground-truth values, while blue lines are the corresponding label-noise distributions. Green rectangles are the bins (borders enlarged for illustration) used for training of the classifier as well as for evaluation of both methods. Note that distributions reach into neighboring bins leading to label confusions.

Table 7.1.: **Test Set Results of the Reg R-CNN study on LIDC and a Toy Dataset.** AVP_{10} measures joint detection and categorization performance, while AP_{10} measures the disentangled detection performance and Bin Accuracy shows categorization performance (conditioned on detection)

	Dim	Network Head	AVP_{10}	AP_{10}	Bin Accuracy
LIDC	3D	Reg R-CNN	**0.259±0.035**	**0.628±0.038**	**0.477±0.035**
		Mask R-CNN	0.235±0.027	0.622±0.029	0.411±0.026
	2D	Reg R-CNN	**0.148±0.046**	**0.414±0.052**	**0.468±0.057**
		Mask R-CNN	0.127±0.034	0.406±0.040	0.447±0.018
Toy	3D	Reg R-CNN	**0.881±0.014**	0.998±0.004	**0.887±0.014**
		Mask R-CNN	0.822±0.070	**1.000±.000**	0.826±0.069
	2D	Reg R-CNN	**0.859±0.021**	**1.000±0.000**	**0.860±0.021**
		Mask R-CNN	0.748±0.022	**1.000±0.000**	0.748±0.021

Clinical label categories however often reside on a continuous and ordered scale, suggesting that label confusions are likely more frequent between adjacent categories.

For this case, we show that both the performance of lesion detection and malignancy grading can be improved upon over a state-of-the-art detection model when simply trading its classification for a regression head and altering the loss accordingly. We document the success of the ensuing model Reg R-CNN on a large lung CT dataset and on a toy dataset that induces artificial ambiguity. We attribute the edge in performance to the loss formulation of the regression task, which naturally accounts for the continuous relation between labels and is therefore less prone to suffer from conflicting gradients from noisy labels. As Eq. 7.4 shows, we employ a metric approach to ordinal data. In general, this is not hazard-free as model performance may suffer from the imposed metric if the scale actually is non-metric [218]. In other words, our approach implicitly assumes the grading scale has sufficiently metric-like properties. To address this limitation, we plan to study alternative non-metric approaches in future work [219, 220]. In summary, we demonstrated how prior knowledge about the underlying task can be baked into the loss function in order to facilitate the learning process, i.e. increase the efficiency of extracting information from available training data, thus improving the generalization ability of the resulting model.

7.2. Robustness against Input Variations: Model-based Domain Adaptation

In Section 5 we demonstrated, how a Roi-to-end CNN (R2E) outperforms DKI-based approaches by integrating all components of the diagnosis pipeline into a CNN architecture and allowing for joint optimization with respect to the ultimate clinical task. However, a limitation of this approach is the intrinsic input dependence of CNNs [221], which in the studied clinical scenario is trained on specific diffusion-weighted images acquired at certain *b-values*, i.e. strengths and timings of gradient fields. This limitation is crucial for large-scale clinical application, since DWI scanning protocols deviate between sites and standardization is not expected in the near future [222]. Furthermore, due to limited training data, it is desirable to ship trained models across clinical sites for inference without re-training on unseen images acquired with arbitrary local protocols. This procedure implies heterogeneities between training data and local inference data, e.g. in the form of *shifted* or *missing* b-values.

Generative models such as generative adversarial networks [223, 224] and variational auto encoders [225, 226] have recently succeeded at domain transformations. Such models could potentially be used to transform images of altered test-time modalities to original images corresponding to the modalities originally seen during training, yet do not eliminate input dependencies: Similar to other domain adaptation methods such as transfer learning or learning common representations of varying inputs [227], all these workarounds themselves require explicit training on possible future input alteration modes, i.e. are not designed to generalize to arbitrary future input alterations. Since model-based approaches such as DKI come with an intrinsic robustness towards input variations (i.e. predictions based on inductive biases from previous training are substituted for an on-the-fly model fit on the test data), input independence could potentially be achieved by operating the CNN on the output coefficients of the model-fit instead of raw DWI inputs. However, the robustness of model-fits is proportional to the number of observed data points, which, as will be shown, is not sufficient in typical DWI acquisition setups.

To this end, we propose Model-based Domain Adaptation (MBDA), where the domain shift encountered by the CNN is minimized by reconstructing original training modalities using DKI on the varied inputs at test time. This method does not require re-training to specific input alterations and hence can be deployed in any clinical setting without prior assumptions about protocol deviations. MBDA significantly reduces input dependencies of CNNs while still exploiting the advantages of learning algorithms over entirely model-based approaches. To demonstrate the superiority of our approach, we compare against the intrinsic domain adaptation of DKI, i.e. against a network trained directly on output coefficients of DKI fits (fit-to-end, F2E).

7.2.1. Methods

The utilized dataset is described in Section 4.2 with the updated train test splits and extended patient cohort described in 6.4.1. The CNN for lesion classification ("Roi-to-end", R2E) as well as the ADC and AKC baselines are identical to the ones in Section 5.3. The dataset comprises measurements at b-values of 0, 100, 750 and 1500 $\mathrm{s\,mm^{-2}}$.

To overcome the dependency on specific b-values and enable clinical applicability of lesion classification regardless of scanning protocols, we propose to perform Model-based Domain Adaptation (MBDA): During inference, the DKI model is fit to the signal intensities of all available (potentially altered) b-values. In order to restore the original set of b-value modalities seen during CNN training, estimates of the original signal intensities $S(b)$ are interpolated or extrapolated per voxel from the fitted DKI model (see Equation 4.1). Subsequently, the restored set of inputs is fed into the trained CNN to perform test time inference (see Figure 7.4 top). To compare our approach against the intrinsic domain adaptation of DKI, we train a baseline method on DKI fit parameters ADC and AKC by feeding the parameter maps directly into the feature extraction and classification modules of the CNN (F2E) (see Figure 5.1). At test time, ADC and AKC are fit using the altered inputs (see Figure 7.4 bottom). For inference of subsets containing only two b-values, which causes the DKI model to be under-constrained, we set AKC = 0.

7.2.2. Experimental Setup

Two exemplary scenarios of inputs variations at test time were studied: *shifted scenario*, where one b-value seen by the CNN during training is provided at a different (shifted) value at test time, and *missing scenario*, where one training b-value is missing at test time. Both scenarios were imitated by training and testing on respective subsets of the four b-values provided by the utilized dataset. Note, that scenarios comprising alterations of multiple inputs were not studied due to the limited number of b-values provided. Furthermore, no alterations were applied to b0 as in practice all protocols include at least one b-value equal or close to zero [50, 228].

An upper bound performance for MBDA is given by training and testing on the same subset of b-values (*matched input*). A lower bound performance for MBDA is given by testing on the altered inputs without accounting for the discrepancy to the training data (*altered input*).

The network details and training setup are equal to the setup reported in 5.4. The networks are trained using 5-fold cross validation with with 60% training- , 20% validation- and 20% test data and the best epoch is selected based on the lowest validation error.

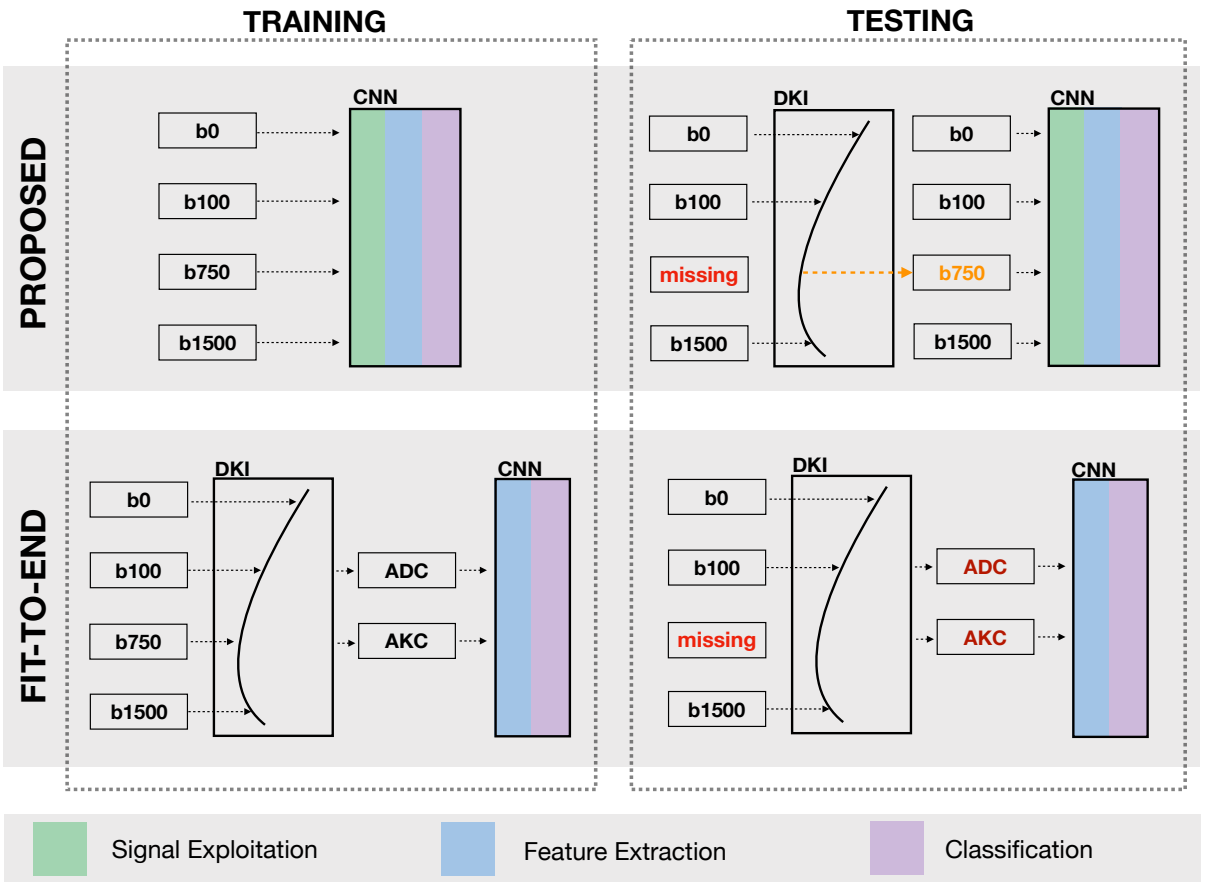

Figure 7.4.: **Conceptual Setup for Model-Based Domain Adaptation.** Visualization of the proposed MBDA for the scenario of missing input modalities at test time (*missing scenario*) (top). The missing b-value input channel is reconstructed voxel-wise from a DKI-model that has been fit to the remaining channels. The fit-to-end architecture trained on the DKI output coefficients ADC and AKC is used for comparison (bottom), where at test time ADC and AKC are re-fit under the altered inputs.

Evaluation is conducted by comparing the area under the receiver operator curves (AUC). Significance tests were performed using DeLong's method and corrected for multiple testing using the Holm-Bonferroni-Method (initial $\alpha = 0.05$).

7.2.3. Results

Results are shown in Table 7.2.

Matched Input These results represent the upper bound on the performance in this study. When decreasing the general number of b-value inputs (but identical during training an testing), the observed decrease of performance is moderate indicating a general redundancy of information across the four input images corresponding to four b-values in this dataset. For instance, subsets of three b-values seem to roughly contain the same information as the original four b-values with respect to overall performance.

Altered Input These results represent the lower bound on the MBDA performance in this study. Strong input dependence with an average performance drop of 19.2% is observed for R2E, when inputs differ between training and testing and no domain adaptation is applied. The baseline method F2E, The intrinsic domain adaptation of the baseline method F2E seems to work to some extend reducing the performance drop to an average of 10.6%.

Model-Based Domain Adaptation MBDA is able to significantly increase the lower bound performance in the shifted scenario by 12.4% and in the missing scenario by 16.8% (see Figure 7.5). Compared to F2E, MBDA considerably outperforms F2E by 5.3% for the shifted scenario and 12.4% for the missing scenario. Notably, extrapolation to large b-values is a poorly constrained problem, which causes performance drops across all explored methods. As expected, F2E only works when constraining the DKI model (setting AKC = 0) during CNN training.

7.2.4. Discussion

The results of this study suggest that Model-based Domain Adaptation (MBDA) is an effective approach to overcome input dependencies and avoid re-training at clinical sites during large-scale application of DWI-based lesion classification. Our approach significantly increases the performance for both missing and shifted input scenarios by combining optimal exploitation of input correlations of raw DWI with DKI-based signal estimation to restore information lost due to altered input. In other words, MBDA is a "minimal invasive" method, which leaves unaltered input untouched, while the implicit domain adaptation performed by training and testing on fit parameters generates entirely new fit parameters given partly altered input, discarding unaltered correspondences. The latter works in theory, given a sufficient number of b-value images, but suffers from fitting instabilities in a typical DWI setup. In addition, when relying on implicit domain adaptation

Table 7.2.: **Results of the Model-based Domain Adaptation Study.** Results comparing all explored methods. All numbers report AUC except for p-values. *x* marks the available b-values. *o* marks the derived b-value. * marks significance according to statistical testing.

a) Shifted Scenario.

Training b-values				R2E Matched Input	F2E Matched Input	Testing b-values				R2E Altered Input	F2E Altered Input	MBDA	p-value R2E;MBDA	p-value R2E;F2E
b0	b100	b750	b1500			b0	b100	b750	b1500					
x	x	x		0.893±0.04	0.819±0.05	x	x	o	x	0.741±0.06	0.768±0.05	**0.848**±0.05	0.0005*	0.011
						x	o	x	x	0.831±0.05	0.845±0.05	**0.893**±0.04	0.0052*	0.0622
x	x		x	0.882±0.04	0.855±0.05	x	x	x	o	0.799±0.06	**0.817**±0.06	0.751±0.07	0.1426	0.1132
						x	o	x	x	0.831±0.05	0.845±0.05	**0.880**±0.04	0.0019*	0.816
x		x	x	0.886±0.04	0.892±0.04	x	x	x	o	0.725±0.07	**0.845**±0.05	0.766±0.07	0.3199	0.0416
						x	x	o	x	0.737±0.07	0.844±0.05	**0.871**±0.05	6.96e-5*	0.422
x	x			0.777±0.06	0.674±0.072	x	o	x		0.680±0.07	0.679±0.07	**0.794**±0.06	0.00014*	0.0018*
						x	o		x	0.666±0.07	0.679±0.07	**0.791**±0.06	0.0002*	0.0015*
x		x		0.889±0.04	0.871±0.05	x	x	o		0.723±0.07	0.608±0.08	**0.796**±0.06	0.0467	4.08e-6*
						x		o	x	0.752±0.06	0.833±0.06	**0.869**±0.05	0.0009*	0.1426
x			x	0.882±0.04	0.877±0.05	x	x		o	0.729±0.07	0.589±0.08	**0.757**±0.06	0.4864	0.0002*
						x		x	o	0.817±0.06	0.825±0.06	**0.866**±0.05	0.0643	0.1485

b) Missing Scenario.

As for subsets of two available b-value images DKI is manually constrained by setting $AKC = 0$, performances for both training with and without the constraint are reported (DKI/ADC)

Training b-values				R2E Matched Input	F2E Matched Input (DKI/ADC)	Testing b-values				R2E Altered Input	F2E Altered Input	MBDA	p-value R2E;MBDA	p-value R2E;F2E (DKI/ADC)
b0	b100	b750	b1500			b0	b100	b750	b1500					
x	x	x	x	0.898±0.05	0.896±0.05	x	x	x	o	0.678±0.07	0.655±0.07	**0.745**±0.07	0.1463	0.0449*
						x	x	o	x	0.604±0.08	0.667±0.07	**0.882**±0.04	1.4e-12*	8.76e-8*
						x	o	x	x	0.823±0.53	0.678±0.07	**0.901**±0.04	0.00028*	1.04e-8*
x	x	x		0.893±0.04	0.819±0.05/ 0.859±0.05	x	x	o		0.513±0.08	0.522±0.08/ 0.617±0.07	**0.780**±0.06	2.1e-7*	1.18e-8*/ 0.00014*
						x	o	x		0.817±0.05	0.514±0.08/ 0.857±0.08	**0.891**±0.04	0.00026*	2.2e-16*/ 0.1041
x	x		x	0.882±0.04	0.855±0.05/ 0.860±0.05	x	x		o	0.512±0.08	0.612±0.08/ 0.652±0.074	**0.755**±0.06	6.92e-6*	0.00067*/ 0.0125*
						x	o		x	0.818±0.05	0.647±0.08/ 0.875±0.05	**0.879**±0.04	0.0003*	3.63e-9*/ 0.8804
x		x	x	0.886±0.04	0.892±0.04/ 0.860±0.05	x		x	o	0.657±0.07	0.646±0.07/ 0.836±0.05	**0.878**±0.04	5.14e-9*	8.72e-10*/ 0.1036
						x		o	x	0.649±0.07	0.699±0.07/ **0.868**±0.05	**0.868**±0.04	3.24e-7*	2.66e-6*/ 0.997

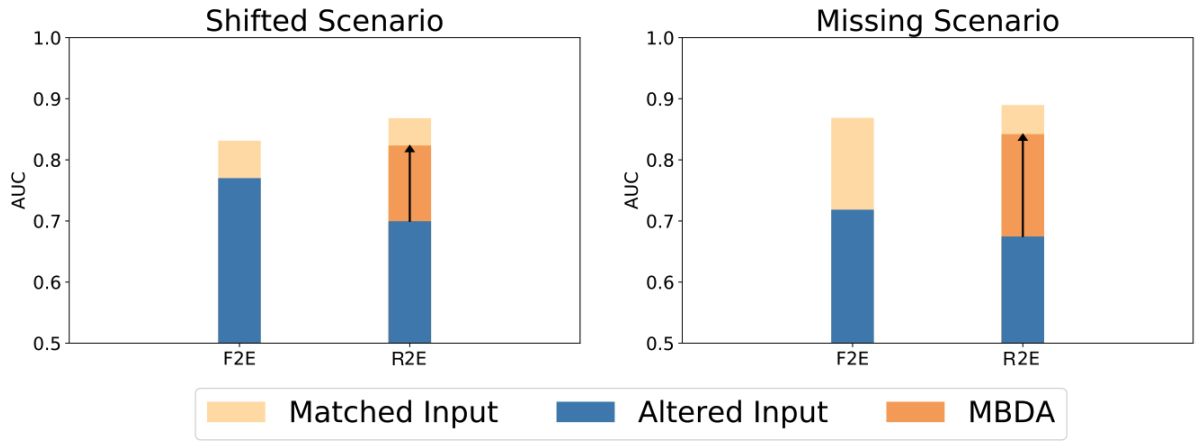

Figure 7.5.: **Visualization of Key Results from Input Variation Study.** Mean AUC derived from Figure 7.2. *Matched input* represents the upper bound with matching b-value subsets during training and inference. *Altered Input* represents the lower bound by testing on the altered subset without MBDA (intrinsic domain adaptation in baseline method F2E is present). The Roi-to-end CNN (R2E) with MBDA significantly improves the robustness towards heterogeneous inputs compared to the intrinsic domain adaptation of DKI present in F2E in both scenarios.

of DKI, strong assumptions have to be made on the amount of b-value images available during clinical inference prior to CNN training (as manually constraining the model by setting $AKC = 0$ might be required), which contradicts the desire for input independence. Future research includes studying multiple input alterations on datasets providing a larger number of b-values, application on whole image breast DWI including the task of lesion localization, investigating the generalization of deep learning models trained on large DWI datasets and exploring the applicability to further entities. In summary, this study shows how domain knowledge in form of a biophysical model can be exploited to increase the robustness of learning-based systems in real life clinical scenarios.

7.3. Robust Hyperparameters: Systematic and Automated Method Configuration

While often underconsidered in applied research such as medical image classification, the choice of a model's hyperparameters affects all components of inductive bias, i.e. architecture, training data, loss function, and optimization, thus finding a good configuration of hyperparameters for a given task is essential for the model's ability to generalize to unseen data. Currently, this process requires high levels of expertise and experience, with small errors leading to strong performance drops [30]. Especially in 3D medical imaging, where dataset properties like imaging modality, image size, (anisotropic) voxel spacing or class ratio vary drastically, the pipeline design can be cumbersome, because experience on

what constitutes a successful configuration may not translate to the dataset at hand. The numerous expert decisions involved in designing and training a neural network range from the exact network architecture to the training schedule and methods for data augmentation or post-processing. Each sub-component is controlled by essential hyperparameters like learning rate, batch size, or class sampling. An additional layer of complexity on the overall setup is posed by the hardware available for training and inference [205].

Attempts at alleviating this highly empirical task by algorithmic optimization range from large sweeps over possible configurations such as Grid Search or Random Search [229] to learning successful configurations directly from the data such as in AutoML [230, 231] or Neural Architecture Search [232]. However, algorithmic optimization of co-dependent design choices in this high dimensional space of hyperparameters amplifies both the number of required training cases as well as compute resources by orders of magnitude [232], vastly exceeding the common setup in medical imaging applications. As a consequence, the end-user is commonly left with an iterative trial and error process during method design that is mostly driven by their individual experience, only scarcely documented and hard to replicate, inevitably evoking suboptimal segmentation pipelines and methodological findings that do not generalize under shifts of data domains [30, 229].

Thus, following this chapter's introductory quote, in this study we condense domain knowledge into a set of key design choices and heuristic rules in order to pose constraints on the space of hyperparameters, shortcut the associated optimization process, and ultimately enable robust and systematic model configuration.

As a prove of concept, we turn to the field of 3D semantic segmentation on biomedical images (see Section 3.3), the most popular methodological sub-field in medical image analysis with 64% of contributions at MICCAI 2019 being related to it and 70% of international competitions in the biomedical sector being segmentation tasks [170]. This widespread interest has resulted in an unprecedented variety of public datasets providing a standardized environment for algorithm benchmarking, which is a key criterion for meaningful validation of a concept regarding automatic adaptation to arbitrary datasets.

The challenges faced when designing a method for a medical image segmentation task are further complicated by an unmanageable number of research papers that propose architecture variations and extensions for performance improvement. This bulk of studies is incomprehensible to the non-expert and difficult to evaluate even for experts [30]. Approximately 12000 studies cite the 2015 U-Net architecture on biomedical image segmentation [108] (see Section 3), many of which propose extensions and advances. We put forward the hypothesis that a basic U-Net is still hard to beat if the corresponding pipeline is designed adequately.

To this end, we propose nnU-Net ("no new net"), which automatically adapts to arbitrary datasets and enables out-of-the-box segmentation on account of two key contributions: 1. We formulate the hyperparameter optimization problem in terms of a *data fingerprint* (representing the key properties of a dataset) and a *pipeline fingerprint* (representing the key design choices of a segmentation algorithm); And 2. We make their relation explicit by condensing domain knowledge into a set of heuristic rules that robustly generate a high quality pipeline fingerprint from a corresponding data fingerprint while considering associated hardware constraints. In contrast to algorithmic method design that is formulated as a task-specific optimization problem, nnU-Net readily executes systematic rules to generate deep learning methods for previously unseen datasets. The superiority of this concept is demonstrated by setting a new state of the art in numerous international challenges by simply applying our algorithm without requiring additional user interaction. Thus, nnU-Net advances the field of medical image segmentation in the following ways:

- nnU-Net is an *open source tool* that enables state-of-the-art segmentation for users with little deep learning experience or researchers that require segmentations for their task without the need for manual intervention.
- nnU-Net is a *blueprint configuration* for high precision tasks such as clinical applications, i.e. may serve as a robust starting point for further task-specific optimization giving access superior configurations under limited compute and data budgets.
- nnU-Net is an *integrative framework* fostering the ambition and the ability of researchers to validate novel ideas on larger numbers of datasets.
- nnU-Net is the first *standardized baseline method* in biomedical image segmentation without the need for task-specific optimization, thus alleviating the noise currently present in literature and catalyzing scientific progress in the field of biomedical deep learning.

7.3.1. Analysis Current Practice and Formalizing the Process

We first provide a quantitative motivation for our contribution by shining light on the importance of method configuration for performance of segmentation models, in contrast to what often seems to be perceived as the crucial factor: the network architecture. To this end, we take an in-depth look at the 2019 Kidney and Kidney Tumor Segmentation (KiTS) semantic image segmentation challenge hosted by the MICCAI society [233]. The MICCAI society has consistently been hosting at least 50% of all annual biomedical image analysis challenges [170]. Challenges are international competitions that can be seen as the equivalent to clinical trials for algorithm benchmarking. Typically, they are hosted by individual researchers, institutes, or societies, aiming to assess the performance of multiple algorithms in a standardized environment [170]. With more than 100 competitors,

the KiTS challenge was the largest competition at MICCAI 2019.

Network architecture is no guarantee for a successful model A thorough analysis of the KiTS leaderboard (see 7.6) a reveals several insights on the current landscape of deep learning based segmentation methods: First, the top-15 methods were offspring of the (3D) U-Net architecture from 2016, confirming its impact on the field of biomedical image segmentation. Second, the figure demonstrates that contributions using the same type of network result in performances spread across the entire leaderboard. This observation is in line with Litjens et al., who, in their review from 2017, found that "many researchers use the exact same architectures [...] but have widely varying results" [30]. Third, when looking closer into the top-15, none of the commonly used architectural modifications (e.g. residual connections [234, 103], dense connections [235, 104], attention mechanisms [236] or dilated convolutions [109, 237]) represent a necessary condition for good performance on the KiTS task. By example this shows that many of the previously introduced algorithm modifications may not generally be superior to a properly tuned baseline method. Again, this finding agrees with an observation by Litjens et al., who concluded that "the exact architecture is not the most important determinant in getting a good solution" [30]. Our findings below confirm this observation across multiple international challenges.

Details in pipeline configuration make the difference Figure 7.6b details the results for the challenge-winning architecture, a 3D U-Net with residual connections. While one of the contributions using this architecture won the challenge, other contributions based on the same principle cover the entire range of metric scores and rankings. A selection of the key pipeline configuration parameters are shown for all non-cascaded residual U-Nets, illustrating the co-dependent design choices that each team made during pipeline design. The drastically varying configurations submitted by contestants indicate the underlying complexity of the high-dimensional optimization problem implicitly posed by designing a deep learning method for biomedical 3D image segmentation. This finding is again in agreement with Litjens et al., who noted that from a range of manually found sets of hyperparameters "disappointingly, no clear recipe can be given to obtain the best set of hyper-parameters as it is a highly empirical exercise" [30]. We refer to the entirety of choices being made during method design as the *pipeline fingerprint*.

Different datasets require different pipeline configurations We extract the key properties of 19 biomedical segmentation datasets including parameters like image sizes, voxel spacing information or class ratios. We refer to this condensed representation characterizing each dataset as the **data fingerprint**. The key parameters of these fingerprints (as visualized in Figure 7.7) document an exceptional dataset diversity in biomedical

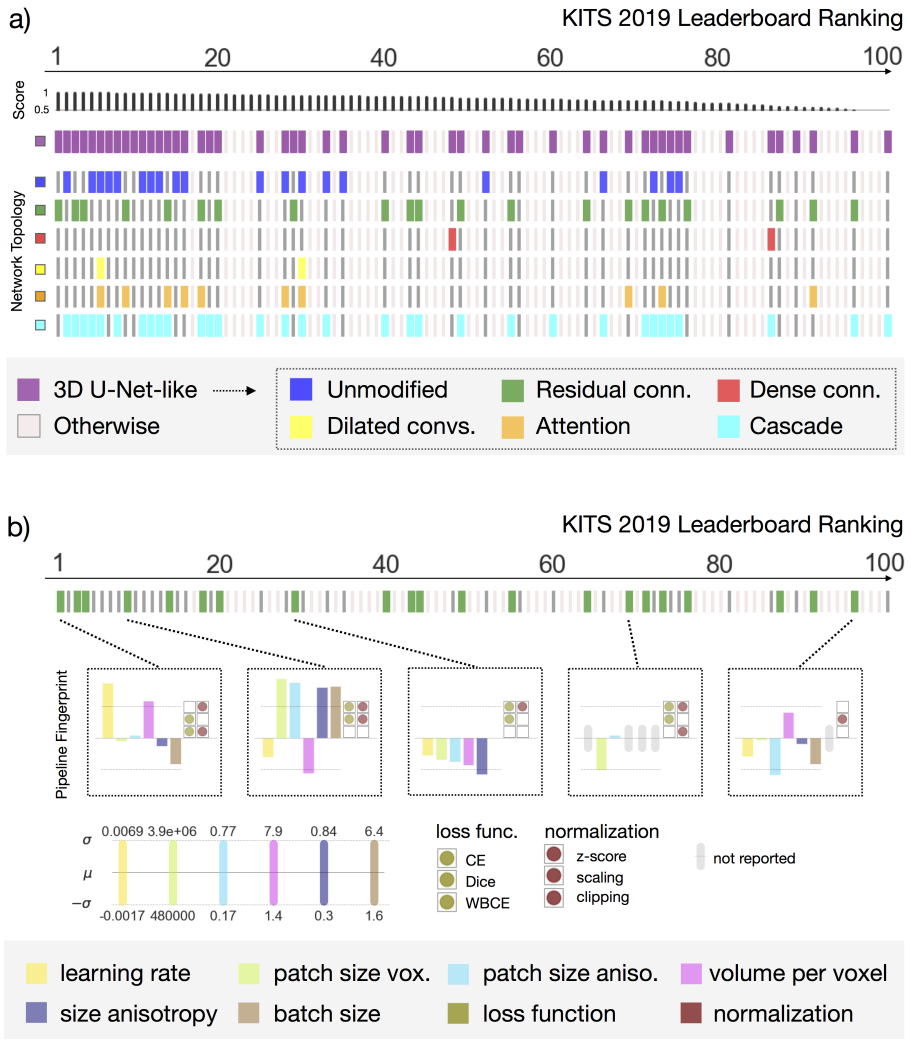

Figure 7.6.: **Pipeline Fingerprints from KITS 2019 Leaderboard Entries [233].** a) Coarse categorization of leaderboard entries by architecture variation. All top 15 contributions are encoder-decoder architectures with skip-connections, 3D convolutions and output stride 1 ("3D U-Net-like", purple). No clear pattern arises from further sub-groupings into different architectural variations. Also, none of the analyzed architectures guarantees good performance, indicating a large dependency of performance beyond architecture type. b) Finer-grained key parameters selected from the pipeline fingerprints of all non-cascade 3D-Unet-like architectures with residual connections (displayed on z-score normalized scale). The contributions vary drastically in their rankings as well as their fingerprints. Still, there is no evident or mono-parametric relation between single parameters and performance. Abbreviations: CE = Cross entropy loss function, Dice = Soft Dice loss function, WBCE = Weighted binary cross entropy loss function.

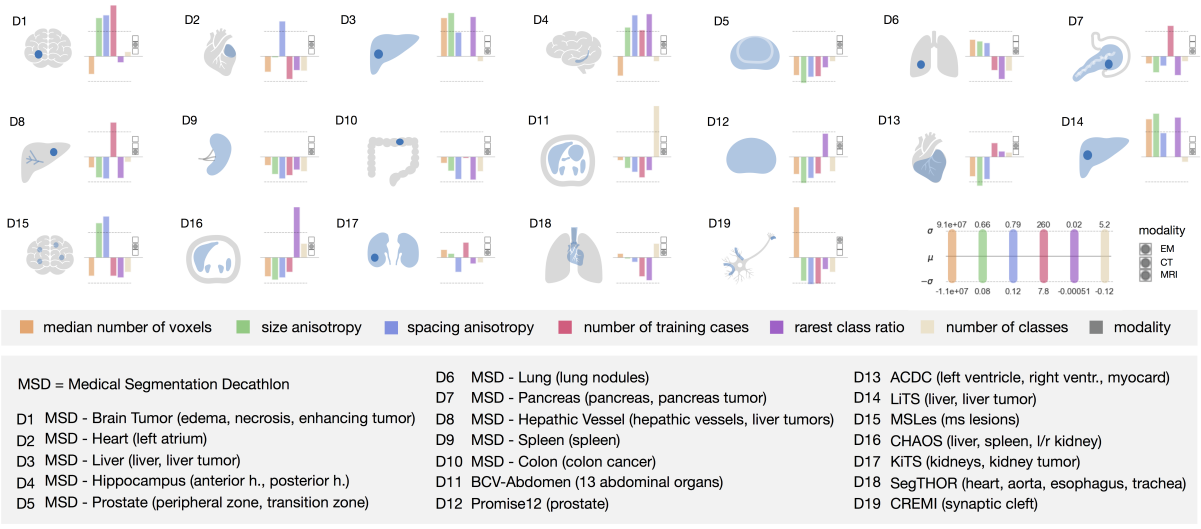

Figure 7.7.: **Data Fingerprints Across Different Challenge Datasets.** The data fingerprints show the key properties (displayed on z-score normalized scale) for the 19 datasets used in the nnU-Net experiments (see Appendix D.3 for detailed dataset properties and references). Datasets vary tremendously in their properties, requiring intense method adaptation to the individual dataset and underlining the need for evaluation of methodological claims on a larger number of datasets. Abbreviations: EM = Electron Microscopy, CT = Computed Tomography, MRI = Magnetic Resonance Imaging.

imaging. This diversity is the fundamental reason behind the lack of segmentation algorithms that generalize across datasets. For each individual dataset, methods need to be adapted and tuned for optimal performance. In this process, optimal pipeline settings either directly or indirectly depend on the data fingerprint, resulting in a complex high-dimensional optimization landscape of co-dependent parameters: To give one example, we note that the image size affects the size of patches the model sees during training, which in turn affects the required network topology (i.e. number of downsampling steps, size of convolution filters, etc.). The network topology itself again influences several other hyperparameters in the pipeline.

7.3.2. Methods

Figure 7.8a shows the current practice of adapting segmentation pipelines to a new dataset. This process is expert-driven and involves manual trial-and error experiments that are typically specific to the task at hand [30]. As shown in Figure 7.8b, nnU-Net addresses the adaptation process systematically. Hyperparameters are divided into three groups: blueprint, inferred and empirical parameters. The blueprint parameters represent fundamental design choices (such as using a plain U-Net-like architecture template) as well as hyperparameters for which a robust default value can simply be picked (for example loss function, training schedule and data augmentation). The inferred parameters encode the necessary adaptations to a new dataset and include modifications to the exact network topology, patch size, batch size and image preprocessing. The link between a data fingerprint and the inferred parameters is established via execution of a set of heuristic rules, without the need for expensive re-optimization when applied to unseen datasets. Note that many of these design choices are co-dependent: The target image spacing, for instance, affects image size, which in return determines the size of patches the model should see during training, which has to be counterbalanced by the size of training mini-batches in order to not exceed GPU memory limitations. nnU-Net strips the user of the burden to manually account for these co-dependencies. The empirical parameters are autonomously identified via cross-validation on the training cases. Per default, nnU-Net generates three different U-Net configurations: a 2D U-Net, a 3D U-Net that operates at full image resolution and a 3D U-Net cascade where the first U-Net operates on down-sampled images and the second is trained to refine the segmentation maps created by the former at full resolution. After cross validation nnU-Net empirically chooses the best performing configuration or ensemble. Finally, nnU-Net empirically opts for "non-largest component suppression" as a postprocessing step when performance gains are measured. The output of nnU-Net's automated adaptation and training process are fully trained U-Net models that can be deployed to make predictions on unseen images. For a more detailed description and implementation details please see Appendix D.1.

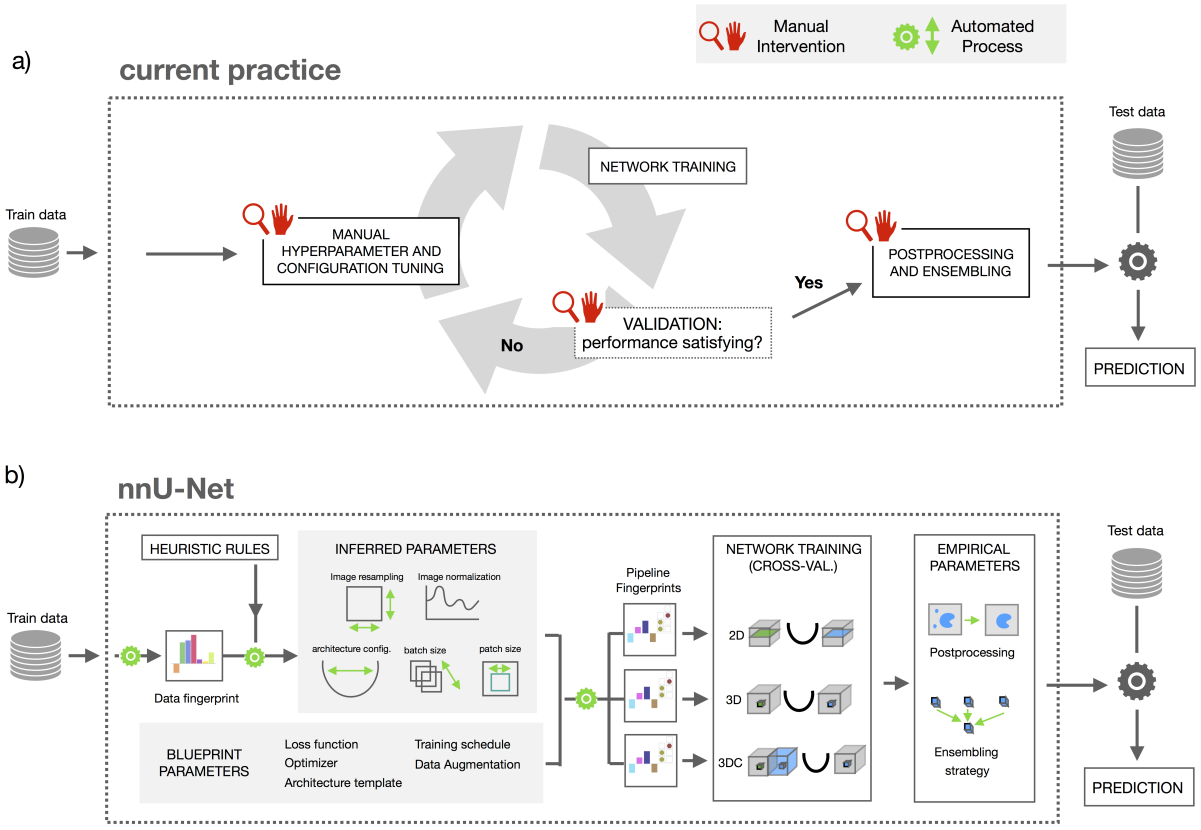

Figure 7.8.: **Manual and Proposed Automated Configuration of Deep Learning Methods.** a) Current practice of configuring a deep learning method for biomedical segmentation: An iterative trial and error process of manually choosing a set of hyperparameters and architecture configurations, training the model, and monitoring performance of the model on a validation set. b) Proposed automated configuration by nnU-Net: Dataset properties are summarized in a "dataset fingerprint". A set of heuristic rules operates on this fingerprint to infer the data-dependent hyperparameters of the pipeline. These are completed by blueprint parameters, the data-independent design choices to form "pipeline fingerprints". Three architectures are trained based on these pipeline fingerprints in a 5-fold cross validation. Finally, nnU-Net automatically opts for the optimal ensemble of these architectures and performs postprocessing if required.

7.3.3. Results

nnU-Net handles a wide variety of target structures and image properties

We tested nnU-Net by participating in various international biomedical image segmentation challenges that comprised 19 datasets and 49 segmentation tasks across a variety of organs, organ substructures, tumors, lesions and cellular structures in magnetic resonance imaging (MRI), computed tomography scans (CT) as well as electron microscopy (EM) images. In all tasks, nnU-Net was trained from scratch using only the provided challenge data. The nnU-Net blueprint parameter choices as well as the exact definition of heuristic rules for inferred parameter generation were optimized in an expert-driven process on basis of the 10 different Decathlon challenge training datasets. The remaining datasets and tasks were used for independent testing, i.e. all nnU-Net settings were frozen and simply applied without further optimization. Qualitatively, we observe that nnU-Net is able to handle a large disparity in dataset properties and diversity in target structures, i.e. generated pipeline configurations are in line with what human experts consider a reasonable or sensible setting. Examples for segmentation results generated by nnU-Net are presented in Appendix D.1.

nnU-Net outperforms specialized pipelines in a range of diverse tasks

Most international challenges use the Soerensen-Dice coefficient as a measure of overlap to quantify segmentation quality [233, 178, 238, 34]. Here, perfect agreement results in a Dice coefficient of 1, whereas no agreement results in a score of 0. Other metrics used by some of the challenges include the Normalized Surface Dice (higher is better) [121] and the Hausdorff Distance (lower is better), both quantifying the distance between the borders of two segmentations. Figure 7.9 provides an overview of the quantitative results retrieved by nnU-Net and the competing challenge teams across all 49 segmentation tasks. Despite its generic nature, nnU-Net outperforms most existing semantic segmentation solutions, even though the latter were specifically optimized towards the respective task. Overall, nnU-Net sets a new state of the art in 29 out of 49 target structures and otherwise shows performances on par with or close to the top leaderboard entries.

Multiple tasks enable robust design decisions

nnU-Net is a framework that enables benchmarking of new modifications or extensions of methods across multiple datasets without having to manually reconfigure the entire pipeline for each dataset. To demonstrate this, and also to support some of the core design choices made in nnU-Net, we systematically tested the performance of common pipeline variations in the nnU-Net blueprint parameters on 10 different datasets (Figure 7.10): the application of two alternative loss functions (Cross-entropy and TopK10 [239]),

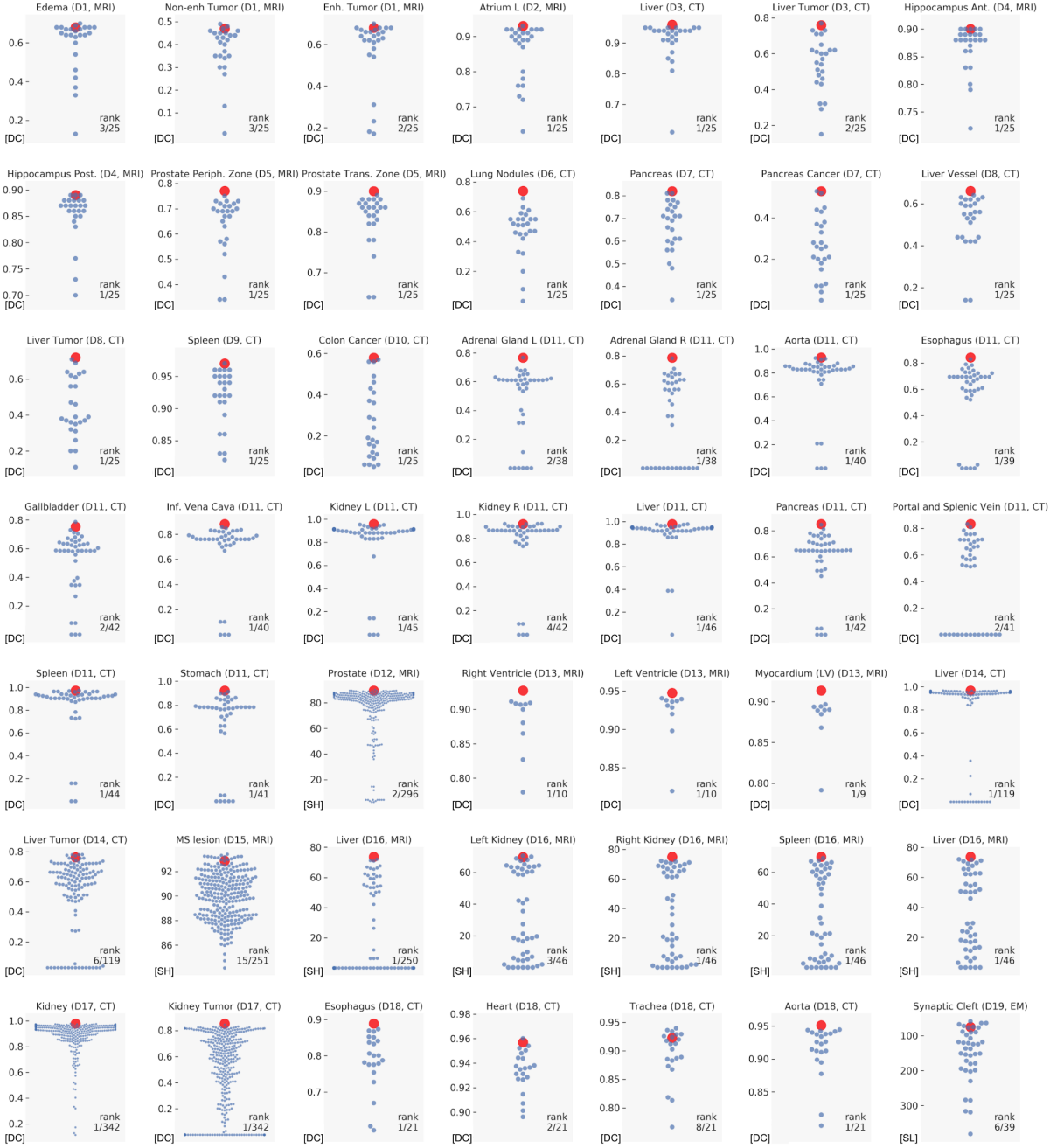

Figure 7.9.: **nnU-Net Outperforms Most Specialized Deep Learning Pipelines.** Quantitative results from all international challenges that nnU-Net competed in. For each segmentation task, results achieved by nnU-Net are highlighted in red, competing teams are shown in blue. For each segmentation task the respective rank is displayed in the bottom right corner as nnU-Net's rank / total number of submissions. Axis scales: [DC] Dice coefficient, [OH] other score (higher is better), [OL] other score (lower is better).

the introduction of residual connections in the encoder, using three convolutions per resolution instead of two (resulting in a deeper network architecture), two modifications of the optimizer (a reduced momentum term and an alternative optimizer (Adam [240])), batch norm instead of instance norm [111] and the omission of data augmentation. Ranking stability was estimated by bootstrapping as suggested by the challengeR tool [241].

The volatility of the ranking between datasets demonstrates how single hyperparameter choices can affect segmentation performance depending on the dataset. The results clearly show that caution is required when drawing methodological conclusions from evaluations that are based on an insufficient number of datasets. While five out of the nine variants achieved rank 1 in at least one of the datasets, neither of them exhibits consistent improvements across the ten tasks. The original nnU-Net configuration shows the best generalization and ranks first when aggregating results of all datasets.

In current research practice, evaluation is rarely done on more than two datasets and even then the datasets often have very similar properties (such as both being abdominal CT scans). As we show here, this evaluation practice is unsuitable for drawing general methodological conclusions. nnU-Net alleviates this shortcoming in two ways: the original nnU-Net serves as a plug-and-play, standardized and state-of-the-art baseline, and the presented underlying framework can be extended to enable effective evaluation of new concepts across multiple tasks.

7.3.4. Discussion

In this study, we demonstrated how the cumbersome trial-and-error process of deep learning model configuration can be automated without the excessive requirements on data and compute posed by algorithmic optimization such as in learning based-approaches [230, 229, 232, 231]. Instead, we condense years of expertise and domain knowledge into identifying robust key design choices and a set of systematic rules, thus posing constraints on the space of possible configurations and shortcutting the optimization process. Moreover, while algorithmic optimization approaches require task-specific re-optimization on the data at hand, our heuristics are developed on the basis of 10 different datasets of the Medical Segmentation Decathlon [173]. The diversity within these 10 datasets has proven sufficient to achieve robustness to the variability encountered in all the remaining challenge participations. This is quite remarkable given the underlying complexity of method design and strongly confirms the suitability of condensing the process in a few generally applicable rules that are simply executed when given a new dataset fingerprint and do not require any further task-specific actions. The formal definition and also publishing of these explicit rules is a step towards systematicity and interpretability in the task of hyperparameter selection, which has previously been considered a "highly empirical exercise" for which "no clear recipe can be given", as noted by Litjens et al. in their review

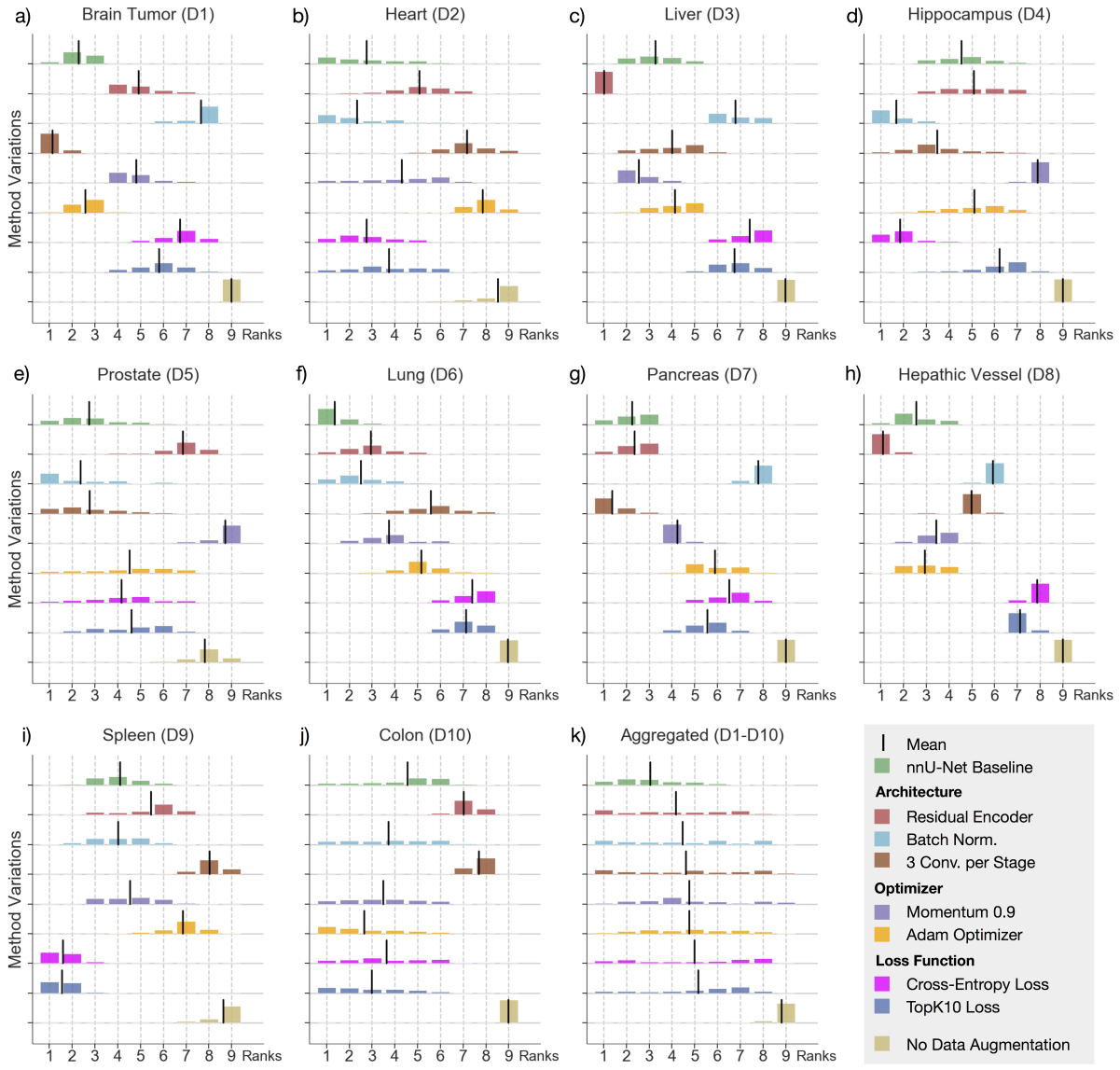

Figure 7.10.: **Evaluation of method design decisions across multiple tasks.** (a-j) Evaluation of exemplary model variations on ten datasets of the medical segmentation decathlon (D1-D10, see Figure 7.7 for dataset references). The analysis is done for every dataset by aggregating validation splits of the five-fold cross-validation into one large validation set. 1000 virtual validation sets are generated via bootstrapping (drawn with replacement). Algorithms are ranked on each virtual validation set, resulting in a distribution over rankings. The results indicate that evaluation of methodological variations on too few datasets is prone to result in a misleading level of generality, since most performance changes are not consistent over datasets. (k) The aggregation of rankings across datasets yields insights into what design decisions robustly generalize.

from 2017 [30].

Specifically, we presented nnU-Net, a deep learning framework for biomedical image analysis that automates model design for 3D semantic segmentation tasks. The method sets a new state of the art in the majority of tasks it was evaluated on, outperforming all respective specialized processing pipelines. The strong performance of nnU-Net is not achieved by a new network architecture, loss function or training scheme (hence the name nnU-Net - "no new net"), but by replacing the complex process of manual pipeline optimization with a systematic approach based on explicit and interpretable heuristic rules. Requiring zero user-intervention, nnU-Net is the first segmentation tool that can be applied out-of-the-box to any dataset. and is thus the ideal tool for users who require access to semantic segmentation methods and do not have the expertise, time, data resources or compute resources required to manually adapt existing solutions to their problem.

Our analysis on the KITS leaderboard as well as nn-UNet's performance across 19 datasets confirms our initial hypothesis that common architectural modifications proposed by the field during the last 5 years may not necessarily be required to achieve state-of-the-art segmentation performance. Instead, we observed that contributions using the same type of network result in performances spread across the entire leaderboard. This observation is again in line with Litjens et al., who found that "many researchers use the exact same architectures [...] but have widely varying results" [30]. There are several possible reasons for the fact that architectural extensions proposed by the literature may not hold beyond the dataset they were proposed on: Many of them are evaluated on a limited amount of datasets, often as low as a single one. In practice this largely limits their success on unseen datasets with varying properties, because the quality of the hyperparameter configuration often overshadows the effect of the evaluated architectural modification. This finding agrees with another observation by Litjens et al., who concluded that "the exact architecture is not the most important determinant in getting a good solution" [30]. Moreover, as shown above, it can be difficult to tune existing baselines to a given dataset. This obstacle can unknowingly, but nonetheless unduly, make a new approach look better than the baseline, resulting in biased literature.

In this work, we demonstrated that nnU-Net is able to alleviate this bottleneck of current research in biomedical image segmentation in two ways: On the one hand, nnU-Net serves as a framework for methodological modifications enabling simple evaluation on an arbitrary number of datasets. On the other hand, nnU-Net represents the first standardized method that does not require manual task-specific adaptation and as such can readily serve as a strong baseline on any new 3D segmentation task.

Despite its strong performance across 49 diverse tasks, there might be segmentation tasks

for which nnU-Net's automatic adaptation is suboptimal. For example, nnU-Net was developed with a focus on the Dice coefficient as performance metric. Some tasks, however, might require highly domain specific target metrics for performance evaluation, which could influence method design. Also, yet unconsidered dataset properties could exist which may cause suboptimal segmentation performance. One example is the synaptic cleft segmentation task of the CREMI challenge (`https://cremi.org`). While nnU-Net's performance is highly competitive (rank 6/39), manual adaptation of the loss function as well as electron microscopy-specific preprocessing may be necessary to surpass state-of-the-art performance [242]. In principle, there are two ways of handling cases that are not yet optimally covered by nnU-Net: For potentially re-occurring cases, nnU-Net's heuristics could be extended accordingly; for highly domain specific cases, nnU-Net should be seen as a good starting point for necessary modifications.

In summary, nnU-Net sets a new state of the art in various semantic segmentation challenges and as such displays strong generalization characteristics without need for any manual intervention, such as the tuning of hyper-parameters. As pointed out by Litjens et al. and quantitatively confirmed here, hyper-parameter optimization constitutes a major difficulty for past and current research in biomedical image segmentation. nnU-Net automates the otherwise often unsystematic and cumbersome procedure and may thus help alleviate this burden. We propose to leverage nnU-Net as an out-of-the box tool for state-of-the-art segmentation, a framework for large-scale evaluation of novel ideas without manual effort, as a standardized baseline method to compare ideas against, and as a starting point for further task-specific tuning in high performance tasks such as real life applications giving access to superior configurations under limited compute and data budgets.

While this study focused on semantic segmentation as a prove of concept effectively limiting the scope to end-to-end predictions on pixel-level, we currently work on adapting the presented method to generating predictions on object-level (such as in Chapter 6). Future work will further focus on leveraging cross-task information: The remarkable performance of nnU-Net compared to specialized solutions might be related to a regularizing effect stemming from cross-task information transfer. We hypothesize that nnU-Net only scratches the surface of the potential that lies withing exchanging knowledge and information between tasks, because currently the transfer happens on an abstract and implicit level of pipeline configuration, but models are still trained from scratch on each task independently of other tasks.

8. Conclusion

8.1. Summary

The work presented here aims to study the potential of end-to-end learning in clinical classification systems and contribute towards crossing the 'AI chasm' in medical image analysis, i.e. the gap between remarkable results of deep learning based diagnosis in research environments and their stuttering performance in clinical practice, by identify key leverage points where domain knowledge can be transformed into informed inductive biases to increase model robustness.

In Chapter 4 and 5 we studied the potential of learning algorithms for clinical diagnosis starting with the current clinical standard and gradually substituting single pipeline components by learning algorithms. In Chapter 4, we replaced current rule-based decision making in DWI-based breast cancer classification with large scale Radiomics feature extraction and machine learning-based classification and achieved human-level performance on an independent test set comprising images from a different clinical site acquired from a different scanner[32]. In clinical numbers, our approach is able to prevent 46 out of 66 unnecessary biopsies while detecting 60 out of 61 cancerous cases. We further showcase an application of this approach to cardiac diagnosis on cine-MRI and ranked second in an international competition [42, 34]. In Chapter 5 we continued to follow the dogma of end-to-end learning, i.e. the idea that enabling simultaneous optimization of all pipeline components with respect to the ultimate clinical target improves upon compound rule-based diagnosis pipelines. To this end, we proposed a CNN architecture designed to integrate the biophysical model for image normalization utilized in DWI, handcrafted (Radiomics) feature extraction as well as clinical categorization, enabling ROI-based classification of breast lesions on DWI [35]. In an in-depth analysis, we revealed potential hidden in DWI by demonstrating the benefits of learned image normalizations as com-

pared to the biophysical model currently deployed in clinical research and provide results indicating a complementary value of representations learned in the CNN with respect to handcrafted feature extraction [36]. The fact, that the CNN is not showing results considerably superior to the radiomics approach based on handcrafted features indicates that the utilized dataset is not large enough to reveal the advantage of unbounded model complexity intrinsic to end-to-end learning. While Chapters 4 and 5 studied the benefits of learning algorithms in clinical diagnosis systems, one important part of the pipeline was not taken into account: The localization of Regions of Interest (ROIs) in the image. The Machine Learning Classifier based on Radiomics features presented in Chapter 4 as well as the CNN presented in Chapter 5 relied on previous manual annotation of lesions (or previous automated annotation of cardiac structures in the case of cardiac diagnosis, see Section 4.1.2). This scenario, however, fails to significantly alleviate the current workload related to human readout of medical images, which is to a large extend assigned to location of ROIs [169]. When including the task of localizing ROIs into the learning process to enable true end-to-end diagnosis starting at the raw images in Chapter 6, there are three deep learning methodologies to be considered that attend to the problem at three different granularity levels of image classification, which translate to specific model evaluation metrics and in return answer to different clinical questions. This relation is in practice distorted by several nuisance factors inducing discrepancies between model evaluation in research environments and actual clinical requirements, ultimately leading to incongruous evaluation metrics and in return contributing to preventing end-to-end diagnosis systems from clinical application [27, 129, 130]. Specifically, we revealed a largely neglected predicament between the strive for crossing the AI chasm by evaluating models at clinically relevant scales on one side, and optimizing for efficient training under the burden of data scarcity on the other side. To address this challenge, we proposed Retina U-Net, a deep learning architecture that fuses state-of-the-art object detection with pixel-wise training signals from semantic segmentation. We demonstrated, how the additional segmentation supervision poses an additional constraint on the training process, which increases data efficiency and thus enables end-to-end object detection and classification on medical images, which corresponds to aligning the model output to the clinically relevant scale [37]. We showcased the superiority of this concept by means of an in-depth comparison to prevalent models from object detection, semantic segmentation and instance segmentation operating in 2D as well as 3D on Breast DWI, CT of the Lung and a series of toy experiments. The corresponding code was open sourced as the Medical Detection Toolkit, the first comprehensive framework for object detection on medical images including e.g. modular implementations of all explored models operating in 2D and 3D [38]. We finally applied our approach to the task of lung cancer staging on PET-CT and performed a sensitivity study under varying clinical training scenarios, indicating the complementary value of secondary tumor annotations for the overall sensitivity of primary tumors. This study was performed in preparation for prospective clinical trials of

our algorithm to be performed in the university hospital Basel [39]. While Chapters 4 and 5 reveal large potential of learning algorithms for diagnosis systems compared to the current clinical standard, Chapter 7 continued on the path of Chapter 6 in identifying further leverage points in existing end-to-end learning systems and presented three examples, where expertise specific to the domain of medical image analysis is condensed into inductive biases aiming to increase model robustness. First, we addressed the challenge of erroneous training annotations by substituting the classification component of end-to-end object detection for regression, which enables to train models directly on the continuous scale of underlying pathological processes, thus elevating the models' robustness against rater confusions [40]. Secondly, we addressed the challenge of input heterogeneities faced by trained models when deployed across clinical sites by proposing model-based domain adaptation, which recaptures the original training domain given altered inputs and thus restores robust generalization [41]. And finally, we addressed the highly unsystematic, elaborate and subjective trial-and-error process of finding a robust set of hyperparameters for a given task by introducing a method that condenses domain knowledge in a set of key design choices and systematic rules and thus enables automated and robust deep learning pipeline configuration on a large variety of medical datasets [42].

8.2. Outlook

Chapter 1 outlined the rise of deep learning in medical image analysis and the tremendous hopes and expectations regarding its impact on the future of health care. As summarized in Section 8.1, the work presented in this thesis experimentally confirms the vast potential of end-to-end learning in diagnosis systems and contributes towards solutions of some of the key obstacles that currently prevent widespread clinical application such as data scarcity, discrepancy between evaluation metrics and clinical relevance, ambiguity in training annotations, or the high manual effort of method adaptation to unseen datasets. In this Section, we shine light on how the findings and contributions presented here tie into the bigger picture of the digital transformation of health care in the 21st century and discuss implications for the projected future of three key players: Researchers, clinicians, and patients.

Quo Vadis Medical Image Analysis?

The success of deep learning has had severe impact on the core identity of applied research fields such as medical image analysis: The shift of research focus away from the engineering of task-specific systems and towards the successful adaptation of existing deep learning methods drastically clashes with research's intrinsic strive for methodological novelty. The consequential re-definition of what constitutes a "novel contribution" in such research areas, where large parts of domain specific knowledge have become seemingly unhelpful or even bothering overnight, is an ongoing process. While deep learning has

brought the long-existing objective of automated medical image classification into reach, it also projected a drastic dependency on training data onto the field. As a consequence, the current endeavors in both medical image analysis and deep learning research align to large extents in addressing the challenge of sample efficiency and model generalization.

One possible way for researchers from within the applied domain to move towards the overarching goal of widespread clinical application of deep learning systems has been demonstrated in this thesis as identifying key leverage points, where domain knowledge can be condensed into inductive biases that reduce the space of possible solutions during generalization and thus increase model robustness. Other promising lines of research from within the medical domain include the attempt to tackle data scarcity at the source in form of federated learning algorithms that enable decentralized training across multiple clinical sites while circumventing patient privacy issues that often prevent large-scale centralized data platforms [243, 244, 245], or the deployment of probabilistic models, which resolve and quantify ambiguities in medical images thus enhancing systematic decision making under limited evidence [217, 246, 247].

Due to the close alignment of methodological challenges, a look at current endeavors in artificial intelligence research might amount to a complementary perspective on the future of research in the medical domain. We observe that current considerations in this field nicely tie in with one of the driving principles of this thesis: To scrutinize end-to-end learning and insert prior knowledge into models in order to increase sample efficiency. In the AI community, the paradigm of end-to-end learning is increasingly rivaled by the preceding paradigm of "Symbolic AI", which, while object to varying interpretations, is described as increasing the prior knowledge in models to enforce symbolic entities as well as logical reasoning among them [248]. There is a strong current believe that true machine intelligence can only be accomplished by a combination of the two paradigms [249, 250]. Turing award winner and AI pioneer Geoffrey Hinton, for instance, devoted large extends of his research in recent years to this idea, specifically to increasing the sample efficiency of CNNs by complementing their inherent translational invariance towards objects in images by additional prior knowledge in form of capsule structures that enable to exploit rotational invariance as well as scale invariance. [251, 252, 253]. Moreover, his fellow Turing award recipient and AI pioneer Yoshua Bengio recently urged for scrutinizing end-to-end learning and drawing inspiration from "high level cognition" in human learning, instead of simply striving for larger datasets, larger models and larger computers [254, 255]. In contrast to pure end-to-end learning, the narrative of combining the two paradigms is backed up by strong evidence from neurology, where high level reasoning on top of observation learning in the human brain seems to drastically increase sample efficiency and enable robust generalization [256, 257]. On the other hand, there does not appear to be a global objective in the brain that is optimized by backpropagating error signals [258, 259].

Instead, the biological brain is highly modular and learns predominantly based on local information [260]. Taken together, there is an increasing call for approaches to artificial intelligence beyond pure learning from observations, which results in a growing number of publications in this direction, one of which bearing the vivid title "Putting an End to End-to-End" [261].

Will Artificial Intelligence Replace Radiologists?

As elaborated on in Chapter 1, hopes are high regarding the impact of deep learning based diagnosis systems on health care, mainly due to an increasing number of models reaching human-level performance on specific tasks [15, 16, 17, 18]. While challenges such as insufficient model robustness under data shift or incongruous evaluation metrics in research environments currently hamper clinical translations, there exist numerous promising approaches to overcoming this AI chasm, many of which have been demonstrated or discussed in the scope of this thesis. Following increased generalization abilities and sound technical validation, a growing number of proposed models is expected to enter the clinical validation stage, i.e. the evaluation of algorithms in prospective studies and randomized clinical trials [262, 26, 27, 263]. In this context, widespread clinical application of deep learning based diagnosis systems is arguably not a matter of if, but when. The questions remain, will this innovation result in a significant change of day-to-day clinical workflows? And will technology eventually replace radiologists?

Geoffrey Hinton stated in late 2016 it is "quite obvious that we should stop training radiologists" as image perception algorithms are very soon going to be demonstrably better than humans. Radiologists are, he said, "the coyote already over the edge of the cliff who hasn't yet looked down" [264]. This statement is to be seen in contemporary context of the peaking AI-hype in 2016 inspiring a considerable number of reputable figures from science [264], economy [265] and health care [266] to predict a looming redundancy of radiologists. Many of these voices, including Geoffry Hinton, have since revised their thinking [267] by courtesy of two key reasons that make automated decision making in medical diagnostics appear more like a distant dream than an imminent revolution: Coverage of rare diseases and ethical concerns.

As Curtis Langlotz, Radiology Professor at Stanford, states in his editorial from 2019, the ability to identify a single disease, as commonly pursued in current attempts at deep learning based diagnosis systems, is a drastic oversimplification of what radiologists do [268]: A comprehensive catalog of radiology diagnoses lists nearly 20000 terms for disorders and imaging observations and over 50000 causal relations [269]. According to Langlotz, "an AI algorithm that diagnoses common chest conditions [...] is a major step forward, an incredible asset to underserved regions, and could serve as a valued assistant for a subspecialty radiologist. But human radiologists are also trained to detect uncom-

mon diseases in the long tail of the distribution including rheumatoid arthritis, sickle cell disease, and posttransplantation lymphoproliferative disorder".

Further, it is unsurprising that the idea of algorithmic decisions regarding human health raises more than one red flag among ethics committees, legal scholars and regulators [270], whose concerns mainly revolve around three topics: fairness, accountability, and transparency [263]: Learning algorithms can be shaped unwittingly by biases in the training data such as under-representation of certain sub-groups of the population e.g. concerning age, gender, ethnicity or social status, thus potentially amplify existing fairness issues in health care systems around the globe. There is a further risk of "automation bias," meaning that humans start to rely entirely on the work of a machine, instead of applying their own critical judgment and scrutiny [271]. Such biases highlight the role of radiologists to guide AI developers, users, and regulators and to hold them accountable to ensure algorithms are safe and unbiased [263]. Moreover, it is unclear how courts will assign liability for patient harm resulting from an algorithm's decision, because moral responsibility is, at least in the developed world, widely accepted as an intrinsically human property that cannot be allocated or shifted to algorithms or machines, however sophisticated they may be [272]. Since AI systems exhibit 'autonomy' to some degree, the European Group on Ethics requires "meaningful human control" being maintained and that humans ultimately remain in control of the decision-making process [272]. This postulate is commonly referred to as 'human in the loop'. An exception could be made in poor-resource areas such as in the developing world, where more radical setups might be required, because a large number of patients are currently treated by nurses or paramedical health workers, who could be trained to receive diagnostic decisions from automated tools to compensate for a lack of doctors [273, 274].

Taken together, the clinical role for automated diagnosis systems in current projections is not one of fully autonomous and unsupervised decision making, but one of decision support for specific tasks. This innovation, however, signifies an imperative component in future health care enabling clinicians to cope with aging populations and increasing demand of medical imaging [6, 4]. Specifically, automation is expected to take on the burden of repetitive perception duties such as image annotation or disease detection and at the same time enable radiologists to focus on other important tasks such as embedding of imaging into clinical workflows and patient history, image acquisition, high level decision making and reporting, research and innovation, or patient interaction [275]. Eric Topol describes this transformation in his book "Deep Medicine: How Artificial Intelligence can make Health Care Human Again" as follows: "What we need in medicine is more inter-human contact and bonding [...]. [Radiologists] could come out of the basement, which is where they live in the dark" [276]. And Curtis Langlotz concisely summarizes the current perspectives on the future of radiology as: "AI won't replace radiologists, but radiologists

who use AI will replace radiologists who don't" [268].

The Patient of the Future

Finally the question remains: How is the AI-driven automation of health care in general, and of medical image classification in particular, going to affect the lives of patients in the future? At present, nearly every patient will experience a diagnostic error in their lifetime, sometimes with devastating consequences [277]. In the united states, medical errors amount to the third leading cause of death [278]. The promise of AI-driven automation in health care is to diminish erroneous decision making by enabling personalized care on the basis of all known information about a patient in real time while incorporating lessons from a collective experience based on vast amounts of data [279]. In radiological diagnostics, 60–80% of errors happen during image perception [280], implying an urgent need for automated medical image classification. Further, automation is expected to bring down costs and to reduce sources of human error such as fatigue or misinterpretation [276], ultimately driving a transformation towards preventive, personalized and ubiquitous health care [27, 276]. In his recent book "The patient will see you know: The future of medicine is in your hands" [281], Eric Topol ties this process into the bigger picture of global digitalization and predicts a "Gutenberg moment" for medicine, that, similar to how the printing press took learning out of the priestly class, AI-driven automation and the mobile internet will give unprecedented control to patients over their own health care, and eventually lead to a democratization of medicine.

List of Own Publications

*equal contribution

Jaeger, P. F., Kohl, S. A. A., Bickelhaupt, S., Isensee, F., Kuder, T. A., Schlemmer, H.-P., & Maier-Hein, K. H. (2020). Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection. Proceedings of Machine Learning Research 116, p.171-183, Machine Learning for Health Workshop at Neurips

Jaeger, P. F., Bickelhaupt, S., Laun, F. B., Lederer, W., Heidi, D., Kuder, T. A., ... Maier-Hein, K. H. (2017). Revealing Hidden Potentials of the q-Space Signal in Breast Cancer. Medical Image Computing and Computer Assisted Intervention - MICCAI 2017 (pp. 664–671). Cham: Springer International Publishing. `https://doi.org/10.1007/978-3-319-66182-7_76`

Isensee, F.*, **Jaeger, P. F.***, Full, P. M., Wolf, I., Engelhardt, S., & Maier-Hein, K. H. (2018). Automatic Cardiac Disease Assessment on cine-MRI via Time-Series Segmentation and Domain Specific Features. Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges (pp. 120–129). Cham: Springer International Publishing. preprint: `https://doi.org/10.1007/978-3-319-75541-0_13`

Isensee, F.*, **Jaeger, P. F.***,Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2020). Automated Design of Deep Learning Methods for Biomedical Image Segmentation. *Under Review at Nature Methods*. preprint: `http://arxiv.org/abs/1904.08128`

Bickelhaupt, S.*, **Jaeger, P. F.***, Laun, F. B., Lederer, W., Daniel, H., Kuder, T. A., ... Maier-Hein, K. H. (2018). Radiomics Based on Adapted Diffusion Kurtosis Imaging Helps to Clarify Most Mammographic Findings Suspicious for Cancer. Radiology, 287(3), 761–770. `https://doi.org/10.1148/radiol.2017170273`

Weikert, T.*, **Jaeger, P. F.***, Bremerich, J., Sommer, G., Stieltjes, B., Yang, S., ... Sauter, A. (2020). Evaluation of different training approaches for the automated detection of lung cancer of all stages on FDG-PET/CT using a Retina U-Net algorithm. *In Submission.*

Kamphenkel, J.*, **Jaeger, P. F.***, Bickelhaupt, S., Laun, F. B., Lederer, W., Daniel, H., ... Maier-Hein, K. H. (2018). Domain Adaptation for Deviating Acquisition Protocols in

CNN-Based Lesion Classification on Diffusion-Weighted MR Images. Image Analysis for Moving Organ, Breast, and Thoracic Images (pp. 73–80). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-00946-5_8

Ramien, G. N., **Jaeger, P. F.**, Kohl, S. A. A., & Maier-Hein, K. H. (2019). Reg R-CNN: Lesion Detection and Grading Under Noisy Labels. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures (pp. 33–41). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-32689-0_4

Petersen, J., **Jaeger, P. F.**, Isensee, F., Kohl, S. A. A., Neuberger, U., Wick, W., ... Maier-Hein, K. H. (2019). Deep Probabilistic Modeling of Glioma Growth. Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 (pp. 806-814). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-32245-8_89

Bernard, O., Lalande, A., Zotti, C., Cervenansky, ..., **Jaeger P.F.**, ..., 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(11), 2514–2525. https://doi.org/10.1109/TMI.2018.2837502

Jaeger, P. F., Bickelhaupt, S., Laun, F. B., Lederer, W., Daniel, H., Kuder, T. A., ... Maier-Hein, K. H. (2018). Complementary value of End-to-end Deep Learning and Radiomics in Breast Cancer Classification on Diffusion-Weighted MRI. Presented at the International Society of Magnetic Resonance in Medicine Annual Meeting. http://archive.ismrm.org/2018/4336.html

Jaeger, P. F., Ramien, G., & Maier-Hein, K. (2020). Medical Detection Toolkit. Zenodo. https://doi.org/10.5281/zenodo.3668656

Appendices

A. Evaluation Metrics for Classification Tasks

Classification models assign events (i.e. images, objects, or pixels according to the granularity level of the model) to different classes. Evaluation is commonly performed on a per-class basis, meaning each possible class is evaluated individually by checking whether the model assigned events correctly or incorrectly to this class according to some ground truth value. This results in the evaluation scheme depicted in Figure A.1, where for each class, events are grouped into the categories True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN). Based on this event categorization, several standard metrics can be computed:

$$\text{Accuracy} = (\text{TP} + \text{TN}) \;/\; (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{A.1}$$

$$\text{Sensitivity} = \text{TP} \;/\; (\text{TP} + \text{FN}) \tag{A.2}$$

$$\text{Specificity} = \text{TN} \;/\; (\text{TN} + \text{FP}) \tag{A.3}$$

$$\text{Precision} = \text{TP} \;/\; (\text{TP} + \text{FP}) \tag{A.4}$$

$$\text{F1} - \text{Score} = 2\text{TP} \;/\; (2\text{TP} + \text{FN} + \text{FP}) \tag{A.5}$$

While Accuracy might be the most intuitive measurement by simply representing the ratio of correctly classified events, it leads to deceived interpretation on imbalanced datasets. Consider for instance a binary task, where 99.9% of events belong to class 1. By simply predicting class 1 for all events, the model would achieve an accuracy of 99.9%. Thus, in

research, models are often evaluated by reporting a trade-off between sensitivity an either precision or specificity, such as the F1-Score.

In practice, classification models often do not predict discrete class assignments of events, but rather continuous scores that may sometimes be interpreted as the confidence or probability of an event belonging to a certain class. The common evaluation strategy for such continuous predictions is to apply discretization operations and subsequently proceed with standard counting metrics as described above. In multi-class setups, *argmax* is typically applied for discretization, where events are assigned to the class with the highest prediction score. In binary classification problems (or class-individual evaluation of multi-task setups) on the other hand, discretization is achieved by selecting a cut-off value and separating events by whether their associated prediction score lies above or below this threshold. This process of discretization, however, entails loss of fine-grained information regarding the model's predictions and thus performance. In binary classification setups, there are attempts at recapturing this information by reporting metrics for a whole range of cut-off values. The most prominent example are Receiver Operating Characetristic (ROC) curves, where sensitivity and specificity are reported for all possible cut-off values in a 2-dimensional plot. ROC curves can be reduced to a single score for simplified reporting by integrating the area under the ROC curve. The resulting score is referred to as the Area Under the Receiver Operating Characetristic Curve (AUC). Constraining the AUC to certain subranges of sensitivity in order to enable task-specific focus of evaluation results in the partial AUC (pAUC). Since ROC includes the computation of specificity and hence depends on the amount of true negatives, it has recently been reported to exhibit brittle behavior e.g. in binary tasks, where a large part of events can easily be assigned to the background class resulting in high amounts of true negatives without requiring separation power by the model [282]. It has thus been suggested to report the PRC on such imbalanced datasets, which substitutes specificity by precision to discard the dependency on true negatives. Equivalently to AUC the integral under the PRC is reported as the Area under the Precision Recall Curve (AUPRC). The latter is in practice often reported as Average Precision (AP), with the small difference that instead of numerical integration (i.e. trapezoidal interpolation) precision values at all sensitivity values are simply averaged. Multi-class setups are dealt with by computing AP over all individual classes and averaging the results, which is referred to as mAP.

Evaluating Image Classification at Different Levels of Granularity Image classification models generate output predictions at different levels of granularity. This is important, because the meaning of associated model evaluation changes across levels and in return affects which initial question about the data is a model answers to (such as different clinical questions in medical applications). While the differing strategies for evaluation were introduced above by generically describing the classification of "events", the defini-

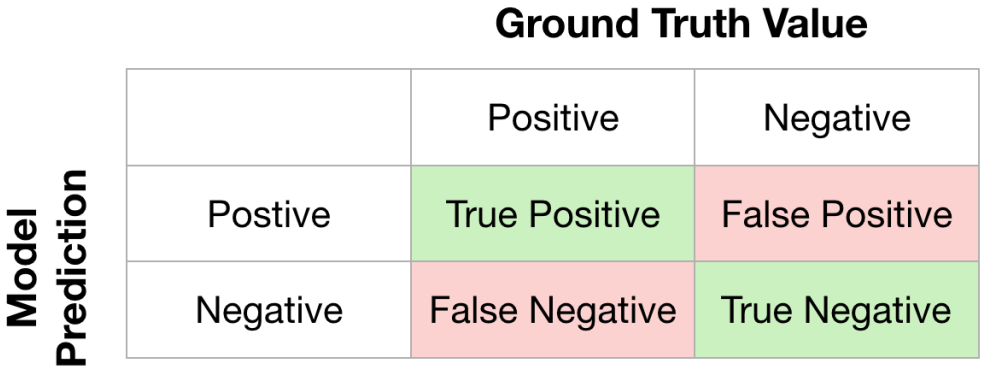

Figure A.1.: **Evaluation Scheme of a Classification Model.**

tion of a event changes across granularity levels, i.e. evaluation on patient-level essentially counts correctly vs. incorrectly classified patients, etc. . Evaluation on Patient-level and pixel-level tasks is straightforward, often reporting AUC for binary problems and F1-score for multi-class setups (in segmentation tasks on pixel-level, F1-score is often called *Soerensen-Dice-Coefficient.*), while evaluation on object-level needs some additional considerations. Here, the set of classified events is not predefined such as the number of images or pixels in a dataset, but depends on an upstream detection task. This task is implemented as an additional localization requirement on predictions in the form of Intersection over Union (IoU) matching, i.e. box coordinates associated with predictions need to exhibit minimum overlap with the coordinates of the ground truth object according to some IoU-threshold. Predictions not satisfying this criterion are filtered out, i.e assigned to either false-positives or true-negatives according to the associated confidence score an the current classification-threshold. As a consequence, true negative-based metrics such as AUC) suffer from skewed datasets during evaluation of the downstream classification. Consider for instance a detection algorithm that predicts many objects on background with low confidence score. These low-confidence objects would result in high amounts of true negative predictions leading to over-optimistic AUC values. As a consequence, mAP, which does not consider true negatives by substituting specificity for precision, is often reported in object-level evaluation such as performed in object detection tasks. Alternatively, object-level evaluation in the medical domain often reports the Free-Response Receiver Operating Characteristic (FROC) curve, where sensitivity is plotted over the number of false positives per image (or patient). This curve is sometimes condensed into a FROC-score by averaging the sensitivity over predefined values of false positives per image.

B. Breast DWI Dataset Extended Information

Table B.1.: **Histopathology of Lesions in the Breast DWI Dataset.**

Histopathology of the Lesions

Histopathologic Result	No. of Patients
Independent test set	127
Benign	66
Sclerosing fibroadenoma	12 (18.2)
Fibrosis	13 (19.7)
Fibrocystic changes	12 (18.2)
Cylindric cell hyperplasia	1 (1.5)
Radial scar	1 (1.5)
Compacted breast tissue	6 (9.1)
Fat tissue necrosis	1 (1.5)
Lobular atrophy	1 (1.5)
Sclerosing adenosis	7 (10.6)
Papilloma	4 (6.1)
Chronic inflammation	2 (3.0)
Apocrine metaplasia	1 (1.5)
Cyst	4 (6.1)
Ductal hyperplasia	1 (1.5)
Malignant	61
Ductal carcinoma in situ	14 (22.9)
Invasive ductal carcinoma	41 (67.2)
Invasive lobular carcinoma	4 (6.6)
Non-Hodgkin lymphoma	1 (1.6)
Papillary carcinoma	1 (1.6)
Training set	95
Benign	34
Fibrosis	8 (23.5)
Atypical ductal hyperplasia	3 (8.8)
Papilloma	3 (8.8)
Granulomatous inflammation	1 (2.9)
Compacted breast tissue	4 (11.8)
Sclerosing fibroadenoma	8 (23.5)
Sclerosing adenosis	1 (2.9)
Fibrocystic changes	3 (8.8)
Cyst	3 (8.8)
Malignant	61
Invasive ductal carcinoma	49 (80.3)
Invasive lobular carcinoma	7 (11.5)
Ductal carcinoma in situ	4 (6.5)
Tubular carcinoma	1 (1.6)

Note.—Data in parentheses are percentages.

Table B.2.: **Lesions Size Statistics of the Breast DWI Dataset.**

Type of Lesion and Measure	Training Set	Test Set	*P* Value
Benign			
Median maximum three-dimensional diameter (mm)	12.61 (5.20–29.45)	11.72 (3.76–61.30)	.527*
Median no. of voxels	14 (3–68)	10 (2–2662)	.700*
Malignant			
Median maximum three-dimensional diameter (mm)	13.18 (7.79–33.90)	14.31 (3.76–51.69)	.214†
Median no. of voxels	25 (5–376)	22 (1–151)	.688†
Nonvisible lesions: median maximum two-dimensional diameter‡	11 (4–50)	11 (4–50)	

Note.—Data in parentheses are the range.

* *t* test, training set vs test set: *P* = (maximum three-dimensional diameter), *P* = (no. of voxels).

† *t* test, malignant vs benign lesions.

‡ As indicated at x-ray mammography.

C. List of Radiomics Features

We calculated a total number of 359 imaging features, divided in four groups as follows: (1) 21 first-order features (FO) (2) 17 volume and shape features (VSF) and (3) 321 texture features (TF). Briefly, FO were calculated from the histogram of voxel intensities using first-order statistics. VSF include diametral, volumetric and surface measurements, however also shape features, e.g. compactness or sphericity. TF are able to characterize the topography of intensity distribution and periodicity in the tumor volume and include co-occurrence, run-length, size-zone, as well as neighbourhood gray level based features. Wavelet decompositions of the original images evaluate the radiomics features at varying spatial frequencies and resolutions with pronounced focus on edge information, along the three spatial directions.

Group 1: First order statistics

First order statistics describe the distribution of gray values within an image. Let $\boldsymbol{X}$ and $\boldsymbol{X_{all}}$ denote the intensity values of all voxels within the corresponding Region of Interest (ROI) with N voxels and the whole image, respectively. The mean value of the gray values within the ROI is $\overline{X}$**.**

The probability vector of the first order histogram with N_l discrete bins is denoted with $\boldsymbol{P}$ and the center gray values with $\boldsymbol{B}$.

1. **Covered Image intensity Rage**:
$$intensity\ range := \frac{\max \boldsymbol{X} - \min \boldsymbol{X}}{\max \boldsymbol{X_{all}} - \min \boldsymbol{X_{all}}}$$
2. **Energy**:
$$energy := \sum_{i}^{N_l} \left(N * \boldsymbol{P}(i)\right)^2$$
3. **Entropy**:
$$entropy := \sum_{i}^{N_l} \boldsymbol{P}(i) * \log_2 \boldsymbol{P}(i)$$
4. **Kurtosis:**
$$kurtosis := \frac{\sum_{i}^{N_l} \boldsymbol{P}(i) * \left(\boldsymbol{B}(i) - \overline{X}\right)^4}{\left(\sqrt{\sum_{i}^{N_l} \boldsymbol{P}(i) * \left(\boldsymbol{B}(i) - \overline{X}\right)}\right)^4}$$
5. **Maximum**:
The maximum intensity value in $\boldsymbol{X}$, i.e. $\max \boldsymbol{X}$
6. **10%-Percentile**
The 10%-Percentile is the lowest intensity value $\boldsymbol{X}_{10}$ which is higher than the intensity of at least 10% of the observation.
7. **90%-Percentile**
The 90%-Percentile is the lowest intensity value $\boldsymbol{X}_{90}$ which is higher than the intensity of at least 90% of the observation.
8. **Interquantile Range**

The interquantile range is defined as the difference between 75% and the 25% quantile:

$$Interquantile\ Range := \boldsymbol{X_{75} - X_{25}}$$

9. **Mean**:

$$mean\ = \overline{X} := \frac{1}{N}\sum_{i}^{N} X(i)$$

10. **Mean absolute deviation:**

$$mean\ absolute\ deviation := \sum_{i}^{N_l} \boldsymbol{P}(i) * (\boldsymbol{B}(i) - \overline{X})$$

11. **Robust Mean Absolute Deviation**
    The robust mean absolute deviation is the 'mean absolute deviation' for all observations between the 10% and 90% Percentile.
12. **Median**:
    The median value of all intensity values in $\boldsymbol{X}$, i.e. the gray value x_{Median} for which
    $$|\{x \mid x \in \boldsymbol{X}\ and\ x < x_{Median}\}| = |\{x \mid x \in \boldsymbol{X}\ and\ x_{Median} < x\}|$$
13. **Minimum**:
    The minimum intensity value in $\boldsymbol{X}$, i.e. $\min \boldsymbol{X}$
14. **No. of Voxels:**
    The number of voxels in $\boldsymbol{X}$**,** i.e. $no.of\ voxels := |\boldsymbol{X}|$
15. **Range:**
    $$range := \max \boldsymbol{X} - \min \boldsymbol{X}$$
16. **Root Means Square (RMS)**:
    $$RMS := \sum_{i}^{N_l} \boldsymbol{P}(i) * \boldsymbol{B}(i)^2$$
17. **Skewness**:
    $$kurtosis := \frac{\sum_{i}^{N_l} \boldsymbol{P}(i) * \left(\boldsymbol{B}(i) - \overline{X}\right)^3}{\left(\sqrt{\sum_{i}^{N_l} \boldsymbol{P}(i) * \left(\boldsymbol{B}(i) - \overline{X}\right)}\right)^3}$$
18. **Standard deviation**:
    $$standard\ deviation := \sqrt{\frac{1}{N-1}\sum_{i}^{N}\left(\boldsymbol{X}(i) - \overline{X}\right)^2}$$
19. **Sum of intensities**:
    $$sum\ of\ intensities := \sum_{i}^{N} \boldsymbol{X}(i)$$
20. **Uniformity**:
    $$uniformity := \sum_{i}^{N_l} \boldsymbol{P}(i)^2$$
21. **Variance**:
    $$standard\ deviation := \frac{1}{N-1}\sum_{i}^{N}\left(\boldsymbol{X}(i) - \overline{X}\right)^2$$

Group 2: Volume and shape based features

The features that are described within this group assess the shape and size of a given Region of Interest. A mesh is generated from the voxel-based annotation and used to calculate the surface area A. The corresponding volume V is estimated based on voxel number.

22. Compactness 1:

$$compactness\ 1 := \frac{V}{\sqrt{\pi} * A^{\frac{2}{3}}}$$

23. Compactness 2:

$$compactness\ 2 := 36\pi \frac{A^2}{V^3}$$

24. Compactness 3:

$$compactness\ 3 := \frac{V}{\sqrt{\pi} * A^{\frac{3}{2}}}$$

25. Maximum 3D diameter:
The maximum 3D diameter is defined as the largest Euclidean distance between any two voxels on the surface of the ROI area.

26. Spherical disproportion:

$$spherical\ disproportion := \frac{A}{4\pi * R^2} = \frac{A}{\left(6\sqrt{\pi} * V\right)^{\frac{2}{3}}}$$

with R being the radius of a sphere with volume V

27. Sphericity:

$$sphericity := \frac{(6\pi^2 V)^{\frac{2}{3}}}{A}$$

28. Asphericity:

$$asphericity := \left(\frac{1}{36} * \frac{A^3}{V^2}\right)^{\frac{1}{3}} - 1$$

29. Surface area:
The surface area is defined as the area of the ROI. It is calculated by summing the area of all triangles of the mesh generated from the ROI annotation.

30. Surface to volume ratio:

$$surface\ to\ volume\ ratio := \frac{A}{V}$$

31. Center of mass shift:
The center of mass shift specifies the differences between the geometric center of mass if each voxel is weighted equally to the geometric center of mass if each voxel is weighted according to its intensity.

32. Volume (mesh based):
The mesh based volume estimation V_{mesh} is calculated by summing the volume of the segmented voxels.

33. Volume (Voxel based):
The volume (V) is estimated as the number of the voxel in the ROI times the voxel spacing, i.e.

$$V \coloneqq N * r_{lat} * r_{cor} * r_{ax}$$

34. PCA Major Axis:

A Principal Component Analysis (PCA) of the locations of the annotated voxels leads to three orthogonal eigenvectors and three eigenvalues$\lambda_1, \lambda_2, \lambda_3$. These eigenvalues can be ordered so that $\lambda_{major} \geq \lambda_{minor} \geq \lambda_{least}$. The PCA Major Axis is then defined as:

$$PCA\ major\ axis \coloneqq 4\sqrt{\lambda_{major}}$$

35. PCA Minor Axis:

$$PCA\ minor\ axis \coloneqq 4\sqrt{\lambda_{minor}}$$

36. PCA Least Axis:

$$PCA\ least\ axis \coloneqq 4\sqrt{\lambda_{least}}$$

37. PCA Elongation:

$$PCA\ elongation \coloneqq \sqrt{\frac{\lambda_{minor}}{\lambda_{major}}}$$

38. PCA Flatness:

$$PCA\ flatness \coloneqq \sqrt{\frac{\lambda_{least}}{\lambda_{major}}}$$

Group 3: Texture features

While first order statistics describe the distribution of the gray values within an image, they do not contain information about the texture of a given region. This can be done by using Gray-Level-co-occurences or Run-length based features.

Gray-level co-occurrence based texture features

Gray-level co-occurrence based texture features provides information about the textural presentation within a given ROI. The gray values are binned into $N_g$bins. Based on these binned gray values a gray-level co-occurrence matrix (GLCM) $\boldsymbol{P}$ with the size $N_g \times N_g$. The (i, j)th element of the matrix is defined as the number of times a voxels is binned into bin i and the voxel in a distance δ and direction α is binned in bin j. Following is two-dimensional examples image I and the corresponding GLCM for $\delta = 1, 2, 3$ and horizontal direction:

I=

2	4	1	1	2
1	3	5	5	1
2	3	4	2	3
1	2	5	2	1
5	3	1	3	5

GLCM=

1	2	2	0	0
1	0	2	1	1
1	0	0	1	2
1	1	0	0	0
1	1	1	0	1

We calculated the GLCM-based features in 3D, reporting the mean and standard deviation of all features calculated on all 13 possible directions and a pixel distance of either $\delta = 1, 2, 3$.

Definitions used during the definition of the features:

- N_g: Number of discrete intensity levels in the image
- $\boldsymbol{P}_{\alpha,\delta}(i,j) = \boldsymbol{P}(i,j)$: Probability of co-occurrence matrix for the pair i,j.
- μ be the mean of $\boldsymbol{P}(i,j)$
- σ be the standard deviation of $\boldsymbol{P}(i,i)$
- $\boldsymbol{P}_{\boldsymbol{x}}(i) = \sum_j^{N_g} \boldsymbol{P}(i,j)$: marginal row probabilities
- μ_x be mean of $\boldsymbol{P}_{\boldsymbol{x}}(i)$
- σ_x be the standard deviation of $\boldsymbol{P}_{\boldsymbol{x}}(i)$
- $\boldsymbol{P}_{\boldsymbol{x+y}}(k) \coloneqq \sum_i^{N_g} \sum_j^{N_g} P(i,j)\ ,\ i+j=k$
- $\boldsymbol{P}_{\boldsymbol{x-y}}(k) \coloneqq \sum_i^{N_g} \sum_j^{N_g} P(i,j)\ ,\ |\, i-j \,| = k$
- $\boldsymbol{HXY} \coloneqq -\sum_i^{N_g} \sum_j^{N_g} \boldsymbol{P}(i,j) \log_2(\boldsymbol{P}(i,j))$
- $\boldsymbol{HX} \coloneqq -\sum_{i=1}^{N_g} \boldsymbol{P}_{\boldsymbol{x}}(i) \log_2 \boldsymbol{P}_{\boldsymbol{x}}(i)$
- $\boldsymbol{HXY_1} \coloneqq -\sum_i^{N_g} \sum_j^{N_g} \boldsymbol{P}(i,j) \log_2(\boldsymbol{P}_x(i)\boldsymbol{P}_{\boldsymbol{y}}(j))$
- $\boldsymbol{HXY_2} \coloneqq -\sum_i^{N_g} \sum_j^{N_g} \boldsymbol{P}_x(i)\boldsymbol{P}_{\boldsymbol{y}}(j) \log_2(\boldsymbol{P}_x(i)\boldsymbol{P}_{\boldsymbol{y}}(j))$

39. – 44. Autocorrelation (mean and std.dev. for $= (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$autocorrelation \coloneqq \sum_i^{N_g} \sum_j^{N_g} i * j * \boldsymbol{P}(i,j)$$

45. – 50. Cluster Prominence (mean and std.dev. for $= (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$cluster\ prominence \coloneqq \sum_i^{N_g} \sum_j^{N_g} (i+j-2\mu)^4 \boldsymbol{P}(i,j)$$

51. – 56. Cluster Shade (mean and std.dev. for $= (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$cluster\ shade \coloneqq \sum_i^{N_g} \sum_j^{N_g} (i+j-2\mu)^3 \boldsymbol{P}(i,j)$$

57. – 62. Cluster Tendency (mean and std.dev. for $= (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$cluster\ tendency \coloneqq \sum_i^{N_g} \sum_j^{N_g} (i+j-2\mu)^2 \boldsymbol{P}(i,j)$$

63. – 68. Contrast / Inertia (mean and std.dev. for $= (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$contrast \coloneqq \sum_i^{N_g} \sum_j^{N_g} (i-j)^2 \boldsymbol{P}(i,j)$$

69. – 74. Correlation (mean and std.dev. for $= (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$correlation: = \frac{1}{\sigma}\sum_{i}^{N_g}\sum_{j}^{N_g}(i-\mu)(j-\mu)\boldsymbol{P}(i,j)$$

75. – 80. Difference Average (mean and std.dev. for $= (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$difference\ average := \sum_{i}^{N_g} i * \boldsymbol{P_{x-y}}(i)$$

81. – 86. Difference Entropy (mean and std.dev. for $= (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$difference\ entropy := \sum_{i}^{N_g} \boldsymbol{P_{x-y}}(i) * \log_2(\boldsymbol{P_{x-y}}(i))$$

87. – 92. Difference Variance (mean and std.dev. for $= (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$difference\ variance := \sum_{i}^{N_g}\left(i - \overline{P_{x-y}}\right)^2 * \boldsymbol{P_{x-y}}(i)$$

93. – 98. Dissimilarity (mean and std.dev. for $= (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$dissimilarity := \sum_{i}^{N_g}\sum_{j}^{N_g}|i-j|\boldsymbol{P}(i,j)$$

99. – 104. Energy (mean and std.dev. for $= (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$energy: = \sum_{i}^{N_g}\sum_{j}^{N_g}\boldsymbol{P}(i,j)^2$$

105. – 110. Entropy (mean and std.dev. for $= (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$entropy := \sum_{i}^{N_g}\sum_{j}^{N_g}\boldsymbol{P}(i,j)\log_2[\boldsymbol{P}(i,j)]$$

111. – 116. Harralick Correlation (mean and std.dev. for $= (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$harralick\ correlation := \frac{1}{\sigma_x}\left[\sum_{i}^{N_g}\sum_{j}^{N_g} i * j * \boldsymbol{P}(i,j) - \mu_x\right]$$

117. – 122. Inverse Difference (Homogeneity 1) (mean and std.dev. for $= (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$inverse\ difference := \sum_{i}^{N_g}\sum_{j}^{N_g}\frac{\boldsymbol{P}(i,j)}{1+\ |i-j|}$$

123. – 128. Inverse Difference Moment (Homogeneity 2, IDM) (mean and std.dev. for $\boldsymbol{\delta} = (\mathbf{1},\mathbf{2},\mathbf{3})$)

$$IDM := \sum_{i}^{N_g} \sum_{j}^{N_g} \frac{\boldsymbol{P}(i,j)}{1 + (i-j)^2}$$

129. – 134. Inverse Difference Moment Normalized (IDMN) (mean and std.dev. for $= (\mathbf{1}, \mathbf{2}, \mathbf{3})$)

$$IDMN := \frac{1}{N^2} \sum_{i}^{N_g} \sum_{j}^{N_g} \frac{\boldsymbol{P}(i,j)}{1 + (i-j)^2}$$

135. – 140. Inverse Difference Normalized (IDN) (mean and std.dev. for $= (\mathbf{1}, \mathbf{2}, \mathbf{3})$)

$$IDN := \frac{1}{N} \sum_{i}^{N_g} \sum_{j}^{N_g} \frac{\boldsymbol{P}(i,j)}{1 + |i-j|}$$

141. – 146. Inverse Variance (mean and std.dev. for $= (\mathbf{1}, \mathbf{2}, \mathbf{3})$)

$$inverse\ variance := \sum_{i}^{N_g} \sum_{j}^{N_g} \frac{\boldsymbol{P}(i,j)}{(i-j)^2}\ ; i \neq j$$

147. – 152. Maximum Probability (mean and std.dev. for $= (\mathbf{1}, \mathbf{2}, \mathbf{3})$)

$$maximum\ probability := \max\{\boldsymbol{P}(i,j)\}$$

153. – 158. Sum Average (mean and std.dev. for $= (\mathbf{1}, \mathbf{2}, \mathbf{3})$)

$$sum\ average := \sum_{i}^{2N_g} i * \boldsymbol{P}_{\boldsymbol{x+y}}(i)$$

159. – 164 Sum Entropy (mean and std.dev. for $= (\mathbf{1}, \mathbf{2}, \mathbf{3})$)

$$sum\ entropy := \sum_{i}^{2N_g} \boldsymbol{P}_{\boldsymbol{x+y}}(i) * \log_2(\boldsymbol{P}_{\boldsymbol{x+y}}(i))$$

165. – 170. Sum Variance (mean and std.dev. for $= (\mathbf{1}, \mathbf{2}, \mathbf{3})$)

$$sum\ variance := \sum_{i}^{2N_g} \left(i - \overline{P_{x+y}}\right)^2 * \boldsymbol{P}_{\boldsymbol{x+y}}(i)$$

171. – 176. Variance (mean and std.dev. for $= (\mathbf{1}, \mathbf{2}, \mathbf{3})$)

$$variance := \sum_{i}^{N_g} \sum_{j}^{N_g} (i-\mu)^2 \boldsymbol{P}(i,j)$$

177. – 182. Joint Average (mean and std.dev. for = (1,2,3))

$$joint\ average := \sum_{i}^{N_g}\sum_{j}^{N_g} i\boldsymbol{P}(i,j)$$

183. – 188. First measure of Information Correlation (mean and std.dev. for = (1,2,3))

$$first\ measure\ of\ information\ correlation := \frac{(HXY - HXY_1)}{HX}$$

189. – 194. Second measure of Information Correlation (mean and std.dev. for = (1,2,3))

$$second\ measure\ of\ information\ correlation := \sqrt{\left(1 - \exp\left(-2(HXY_2 - HXY)\right)\right)}$$

Gray-level run-length based texture features

A gray level run length matrix (GLRLM) $\boldsymbol{P}$ can be used to assess the structure of an image. Each element (k, l) of such a matrix describes the number of runs with l consecutively voxels within a gray level bin k in a given direction θ. The gray values are binned into $N_g$different bins to avoid short bins due to noise. Following is two-dimensional examples image I and the corresponding GLRLM in horizontal direction:

I=

2	4	1	1	2
1	1	5	5	5
2	3	4	2	3
1	2	2	2	1
5	3	1	3	5

GLRLM =

3	2	0	0	0
4	0	1	0	0
4	0	0	0	0
2	0	0	0	0
2	0	1	0	0

We calculated the GLRLM -based features in 3D, reporting the mean and standard deviation of the features based on all possible 13 directions.

Necessary definitions:

- $\boldsymbol{P}_{\boldsymbol{\theta}}(k,l) = \boldsymbol{P}(k,l)$: Number of runs with gray value k and length l in direction θ
- N_g: Number of discrete gray values
- N_r: Number of discrete run lengths
- N_{run}: Number of different runs
- N_p: Number of voxels in ROI

195. – 200. Gray level nonuniformity (GLN) (mean and std.dev. , Ng = (64,128,256))

$$GLN := \frac{1}{N_{run}}\sum_{k}^{N_g}\left[\sum_{l}^{N_r}\boldsymbol{P}(k,l)\right]^2$$

201. – 206. Gray level nonuniformity normalized (GLNN) (mean and std.dev. , Ng = (64,128,256))

$$GLNN \coloneqq \frac{1}{N_{run}^2}\sum_{k}^{N_g}\left[\sum_{l}^{N_r}\boldsymbol{P}(k,l)\right]^2$$

207. – 212. High gray level run emphasis (HGLRE) (mean and std.dev. , Ng = (64,128,256))

$$HGLRE \coloneqq \frac{1}{N_{run}}\sum_{k}^{N_g}\sum_{l}^{N_r}k^2 * \boldsymbol{P}(k,l)$$

213. – 218. Long run emphasis (LRE) (mean and std.dev. , Ng = (64,128,256))

$$LRE \coloneqq \frac{1}{N_{run}}\sum_{k}^{N_g}\sum_{l}^{N_r}l^2 * \boldsymbol{P}(k,l)$$

219. – 224. Long run high gray level emphasis (LRHGLE) (mean and std.dev. , Ng = (64,128,256))

$$LRHGLE \coloneqq \frac{1}{N_{run}}\sum_{k}^{N_g}\sum_{l}^{N_r}k^2l^2 * \boldsymbol{P}(k,l)$$

225. – 230. Long run low gray level emphasis (LRLGLE) (mean and std.dev. , Ng = (64,128,256))

$$LRLGLE \coloneqq \frac{1}{N_{run}}\sum_{k}^{N_g}\sum_{l}^{N_r}\frac{l^2}{k^2}\boldsymbol{P}(k,l)$$

231. – 236. Low gray level run emphasis (LGLRE) (mean and std.dev. , Ng = (64,128,256))

$$LGLRE \coloneqq \frac{1}{N_{run}}\sum_{k}^{N_g}\sum_{l}^{N_r}\frac{1}{k^2}\boldsymbol{P}(k,l)$$

237. – 242. Number of runs (mean and std.dev. , Ng = (64,128,256))

$$number\ of\ runs = N_{runs} \coloneqq \sum_{k}^{N_g}\sum_{l}^{N_r}\boldsymbol{P}(k,l)$$

243. – 248. Run length nonuniformity (RLN) (mean and std.dev. , Ng = (64,128,256))

$$RLN \coloneqq \frac{1}{N_{run}}\sum_{l}^{N_r}\left[\sum_{k}^{N_g}\boldsymbol{P}(k,l)\right]^2$$

249. – 254. Run length nonuniformity normalized (RLNN) (mean and std.dev. , Ng = (64,128,256))

$$RLNN := \frac{1}{N_{run}^2}\sum_{l}^{N_r}\left[\sum_{k}^{N_g}\boldsymbol{P}(k,l)\right]^2$$

255. – 260. Run percentage (RP) (mean and std.dev. , Ng = (64,128,256))

$$RP := \frac{N_{run}}{N_p}$$

261. – 266. Short run emphasis (SRE) (mean and std.dev. , Ng = (64,128,256))

$$SRE := \frac{1}{N_{run}}\sum_{k}^{N_g}\sum_{l}^{N_r}\frac{1}{l^2} * \boldsymbol{P}(k,l)$$

267. – 272. Short run high gray level emphasis (SRHGLE) (mean and std.dev. , Ng = (64,128,256))

$$SRHGLE := \frac{1}{N_{run}}\sum_{k}^{N_g}\sum_{l}^{N_r}\frac{k^2}{l^2}\boldsymbol{P}(k,l)$$

273. – 278. Short run low gray level emphasis (SRLGLE) (mean and std.dev. , Ng = (64,128,256))

$$SRLGLE := \frac{1}{N_{run}}\sum_{k}^{N_g}\sum_{l}^{N_r}\frac{1}{k^2l^2}\boldsymbol{P}(k,l)$$

279. – 284. Gray Level Variance (GLV) (mean and std.dev. , Ng = (64,128,256))

With

$$\mu = \frac{1}{N_S}\sum_{k}^{N_g}\sum_{l}^{N_r} i\boldsymbol{P}(k,l)$$

GLV is defined as:

$$GLV := \frac{1}{N_S}\sum_{k}^{N_g}\sum_{l}^{N_r}(i-\mu)^2\boldsymbol{P}(k,l)$$

285. – 290. Run Length Variance (RLV) (mean and std.dev. , Ng = (64,128,256))

With

$$\mu = \frac{1}{N_S}\sum_{k}^{N_g}\sum_{l}^{N_r} j\boldsymbol{P}(k,l)$$

RLV is defined as:

$$RLV \coloneqq \frac{1}{N_s}\sum_{k}^{N_g}\sum_{l}^{N_r}(j-\mu)^2\boldsymbol{P}(k,l)$$

291. – 296. Run Length Entropy (RLE) (mean and std.dev. , Ng = (64,128,256))

$$RLV \coloneqq -\sum_{k}^{N_g}\sum_{l}^{N_r}\frac{\boldsymbol{P}(k,l)}{N_S}\log_2\frac{\boldsymbol{P}(k,l)}{N_S}$$

Gray-level size zone based texture features

A gray level size zone matrix (GLSZM) P can be used to assess the structure of an image. Each element (k,l) of such a matrix gives the number of connected areas with the discretized intensity k and l connected voxels. A 26 neighborhood is used to determine the connectivity between two voxels. The gray values are binned into $\mathrm{N_g}$different bins to avoid short bins due to noise. Following is two-dimensional examples image I and the corresponding GLSZM:

I=

2	4	1	1	2
1	1	5	5	5
2	3	4	2	3
1	2	2	2	1
5	3	1	3	5

GLSZM =

4	1	0	0	0
2	0	0	0	5
4	0	0	0	0
2	0	0	0	0
2	0	1	0	0

We calculated the GLRLM -based features in 3D.

Necessary definitions:

- $\boldsymbol{P}(k,l)$: Number of areas with gray value k and voxel count l
- N_g: Number of discrete gray values
- N_z: Largest voxel count
- N_v: Number of voxels
- $N_S \coloneqq \sum_{i=1}^{N_G}\sum_{j=1}^{N_Z}\boldsymbol{P}(i,j)$: Number of zones
- $\boldsymbol{P}(k,\cdot) \coloneqq \sum_{j=1}^{N_z}\boldsymbol{P}(k,j)$: Number of areas with gray value k
- $\boldsymbol{P}(\cdot,l) \coloneqq \sum_{i=1}^{N_G}\boldsymbol{P}(i,l)$: Number of areas with voxel count l

297. – 299. Small Zone Emphasis (SZE) (Ng = (64,128,256))

$$SZE \coloneqq \frac{1}{N_s}\sum_{j}^{N_Z}\frac{\boldsymbol{P}(\cdot,j)}{j^2}$$

300. – 302. Large Zone Emphasis (LZE) (Ng = (64,128,256))

$$LZE \coloneqq \frac{1}{N_s}\sum_{j}^{N_Z} \boldsymbol{P}(\cdot, j) \cdot j^2$$

303. – 305. Low Gray Level Zone Emphasis (LGLZE) (Ng = (64,128,256))

$$LGLZE \coloneqq \frac{1}{N_s}\sum_{i}^{N_G} \frac{\boldsymbol{P}(i,\cdot)}{i^2}$$

306. – 308. High Gray Level Zone Emphasis (HGLZE) (Ng = (64,128,256))

$$HGLZE \coloneqq \frac{1}{N_s}\sum_{i}^{N_G} \boldsymbol{P}(i,\cdot) \cdot i^2$$

309. – 311. Small Zone Low Gray Level Emphasis (SZLGLE) (Ng = (64,128,256))

$$SZLGLE \coloneqq \frac{1}{N_s}\sum_{i}^{N_G}\sum_{j}^{N_Z} \frac{\boldsymbol{P}(i,j)}{i^2 j^2}$$

312. – 314. Small Zone High Gray Level Emphasis (SZHGLE) (Ng = (64,128,256))

$$SZHGLE \coloneqq \frac{1}{N_s}\sum_{i}^{N_G}\sum_{j}^{N_Z} \frac{i^2 \boldsymbol{P}(i,j)}{j^2}$$

315. – 317. Large Zone High Gray Level Emphasis (LZHGLE) (Ng = (64,128,256))

$$LZHGLE \coloneqq \frac{1}{N_s}\sum_{i}^{N_G}\sum_{j}^{N_Z} i^2 j^2 \boldsymbol{P}(i,j)$$

318. – 320. Large Zone Low Gray Level Emphasis (LZLGLE) (Ng = (64,128,256))

$$LZLGLE \coloneqq \frac{1}{N_s}\sum_{i}^{N_G}\sum_{j}^{N_Z}\frac{j^2\boldsymbol{P}(i,j)}{i^2}$$

321. – 323. Gray Level Non-Uniformity (GLNU) (Ng = (64,128,256))

$$GLNU \coloneqq \frac{1}{N_s}\sum_{i}^{N_G}\boldsymbol{P^2}(i,\cdot)$$

324. – 326. Gray Level Non-Uniformity Normalized (GLNUN) (Ng = (64,128,256))

$$GLNUN \coloneqq \frac{1}{N_s^2}\sum_{i}^{N_G}\boldsymbol{P^2}(i,\cdot)$$

327. – 329. Zone Size Non-Uniformity (ZSNU) (Ng = (64,128,256))

$$ZSNU \coloneqq \frac{1}{N_s}\sum_{j}^{N_Z}\boldsymbol{P^2}(\cdot,j)$$

330. – 332. Zone Size Non-Uniformity Normalized (ZSNU) (Ng = (64,128,256))

$$ZSNUN \coloneqq \frac{1}{N_s^2}\sum_{j}^{N_Z}\boldsymbol{P^2}(\cdot,j)$$

333. – 335. Zone Percentage (ZP) (Ng = (64,128,256))

$$ZP \coloneqq \frac{N_S}{N_V}$$

336. – 338. Gray Level Variance (GLV) (Ng = (64,128,256))

With

$$\mu \coloneqq \frac{1}{N_S}\sum_{i}^{N_G}\sum_{j}^{N_Z} i\ \boldsymbol{P}(i,j)$$

the Gray Level Variance is defined as:

$$GLV \coloneqq \frac{1}{N_S}\sum_{i}^{N_G}\sum_{j}^{N_Z}(i-\mu)\ \boldsymbol{P}(i,j)$$

339. – 341. Zone Size Variance (ZSV) (Ng = (64,128,256))

With

$$\mu := \frac{1}{N_S} \sum_{i}^{N_G} \sum_{j}^{N_Z} j \; \boldsymbol{P}(i,j)$$

the Zone Size Variance is defined as:

$$ZSV := \frac{1}{N_S} \sum_{i}^{N_G} \sum_{j}^{N_Z} (j-\mu) \; \boldsymbol{P}(i,j)$$

342. – 344. Zone Size Entropy (ZSE) (Ng = (64,128,256))

$$ZSE := \frac{1}{N_S} \sum_{i}^{N_G} \sum_{j}^{N_Z} \boldsymbol{P}(i,j) \; \log_2\left(\frac{\boldsymbol{P}(i,j)}{N_s}\right)$$

Neighborhood Gray Level Difference based texture features
Different statistical measures are calculated for each intensity value i to calculate Neighborhood Gray Level Difference based texture features:

- N_v Number of all annotated voxels
- N_G: Number of discrete gray values
- N_{GP}: Number of discrete gray values that are actually present in the region
- n_i: Number of voxels with intensity i
- $p_i = \frac{n_i}{N_V}$: Probability for the occurrence of intensity i
- a_x : Average Intensity of all voxels in a defined neighborhood around voxel x, excluding x itself.
- $\overline{A}_\iota = \frac{1}{|x \,\forall\, I(x)=i|} \sum_{x \,\forall\, I(x)=i} a_x$: The mean value of all a_x that correspond to voxels with the intensity i
- $s_i = \begin{cases} |\overline{A}_\iota - i\,| & \text{for } n_i > 0 \\ 0 & \text{otherwise} \end{cases}$

An example of some if these features is given below. Note that only the dark area is used as calculation input, the border area is necessary to be able to calculate the 1-size neighborhood.

I=

1	2	2	3

Features

i	n_i	p_i	s_i
1	0	0	0

1	2	3	3
4	2	4	1
4	1	2	3

2	2	0.5	1
3	1	0.25	0.625
4	1	0.25	1.825

345. – 347. Coarsness (Ng = (64,128,256))

$$coarsness := \frac{1}{\sum_{i=1}^{N_G} p_i s_i}$$

348. – 350. Contrast (Ng = (64,128,256))

$$contrast := \left(\frac{1}{N_{GP}(N_{GP}-1)}\sum_{i}^{N_G}\sum_{j}^{N_G} p_i p_j (i-j)^2\right)\left(\frac{1}{N_v}\sum_{i}^{N_G} s_i\right)$$

351. – 353. Busyness (Ng = (64,128,256))

$$busyness := \frac{\sum_i^{N_G} p_i s_i}{\sum_i^{N_G}\sum_j^{N_G}|ip_i - jp_j|}, \qquad p_i \neq 0, p_j \neq 0$$

354. – 356. Complexity (Ng = (64,128,256))

$$complexity := \frac{1}{N_V}\sum_{i}^{N_G}\sum_{j}^{N_G}|i-j|\frac{p_i s_i + p_j s_j}{p_i + p_j} \qquad p_i \neq 0, p_j \neq 0$$

357. – 359. Strength (Ng = (64,128,256))

$$strength := \frac{\sum_i^{N_G}\sum_j^{N_G}(p_i + p_j)(i-j)^2}{\sum_i^{N_G} s_i} \qquad p_i \neq 0, p_j \neq 0$$

D. nnU-Net Extended Information

D.1. Details of nnU-Net implementation

Dataset fingerprints

nnU-Net creates a dataset fingerprint that captures all relevant parameters and properties from the provided training data: image sizes (i.e. number of voxels per spatial dimension), image spacings (i.e. the physical size of the voxels), modalities (read from metadata) and number of classes for all images as well as the total number of training cases. Furthermore, the fingerprint includes the mean, standard deviation as well as the 0.5 and 99.5 percentiles of the intensity values in the foreground regions, i.e. the voxels belonging to any of the class labels, computed over all training cases. As a first processing step that precedes fingerprint extraction, nnU-Net crops all images to their nonzero region. While this had no effect on most datasets in our experiments, it reduced the image size of brain datasets such as D1 (Brain Tumor) and D15 (MSLes) substantially and thus improved computational efficiency. In order to enable the dataset fingerprint extraction, nnU-Net expects its input data in a specific format (see `https://github.com/MIC-DKFZ/nnUNet` for more detail).

Pipeline fingerprints

nnU-Net automizes the design of deep learning methods for biomedical image segmentation by generating a so-called pipeline fingerprint that contains all relevant information. Importantly, nnU-Net reduces the design choices to the really essential ones and automatically infers these choices using a set of heuristic rules. These rules condense the domain knowledge and operate on the above-described data fingerprint and the project-specific hardware constraints. These inferred parameters are completed by blueprint parameters, which are data-independent, and empirical parameters, which are optimized during training.

Blueprint parameters

Architecture template: All U-Net architectures configured by nnU-Net originate from the same template. This template closely follows the original U-Net [108] and its 3D counterpart [283]. According to our hypothesis that a well-configured plain U-Net is still hard to beat, none of our U-Net configurations make use of recently proposed architectural variations such as residual connections [103, 284], dense connections [104, 235], attention mechanisms [236], squeeze and excitation [285] or dilated convolutions [109]. Minor changes with respect to the original architecture were made: To enable large patch sizes, the batch size of the networks in nnU-Net is small. In fact, most 3D U-Net configurations were trained with a batch size of only 2 (see Supplementary Material S1a). Batch normalization [110], which is often used to speed up or stabilize the training, does not perform well with small batch sizes [286, 287]. We therefore use instance normalization [111] for all U-Net models. Furthermore, we replace ReLU with leaky ReLUs [288] (negative slope 0.01). Networks are trained with deep supervision: additional auxiliary losses are added in the decoder to all but the two lowest resolutions, allowing gradients to be injected deeper into the network and facilitating the training of all layers in the network. All U-Nets employ the very common configuration of two blocks per resolution step in both encoder and decoder, with each block consisting of a convolution, followed by instance normalization and a leaky ReLU nonlinearity. Downsampling is implemented as strided convolution (motivated by representational bottleneck, see [289]) and upsampling as convolution transposed. As a tradeoff between performance and memory consumption, the initial number of feature maps is set to 32 and doubled (halved) with each downsampling (upsampling) operation. To limit the final model size, the number of feature maps is additionally capped at 320 and 512 for 3D and 2D U-Nets, respectively.

Training schedule: Based on experience and as a tradeoff between runtime and reward, all networks are trained for 1000 epochs with one epoch being defined as iteration over 250 minibatches. Stochastic gradient descent with nesterov momentum ($\mu = 0.99$) and an initial learning rate of 0.01 is used for learning network weights. The learning rate is decayed throughout the training following the 'poly' learning rate policy [109]: $(1 - \text{epoch}/\text{epoch}_{\text{max}})^{0.9}$. The loss function is the sum of cross-entropy and Dice loss [290]. For each deep supervision output, a corresponding downsampled ground truth segmentation mask is used for loss computation. The training objective is the sum of the losses at all resolutions: $L = w_1 \cdot L_1 + w_2 \cdot L_2 + \dots$. Hereby, the weights halve with each decrease in resolution, resulting in $w_2 = 1/2 \cdot w_1; w_3 = 1/4 \cdot w_1$, etc. and are normalized to sum to 1. Samples for the mini batches are chosen from random training cases. Oversampling is implemented to ensure robust handling of class imbalances: 66.7% of samples are from random locations within the selected training case while 33.3% of patches are guaranteed to contain one of the foreground classes that are present in the selected training sample

(randomly selected). The number of foreground patches is rounded with a forced minimum of 1 (resulting in 1 random and 1 foreground patch with batch size 2). A variety of data augmentation techniques are applied on the fly during training: rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma and mirroring. Details are provided in Supplementary information ZZ.

Inference: Images are predicted with a sliding window approach, where the window size equals the patch size used during training. Adjacent predictions overlap by half the size of a patch. The accuracy of segmentation decreases towards the borders of the window. To suppress stitching artifacts and reduce the influence of positions close to the borders, a Gaussian importance weighting is applied, increasing the weight of the center voxels in the softmax aggregation. Test time augmentation by mirroring along all axes is applied.

Inferred Parameters

Intensity normalization: There are two different image intensity normalization schemes supported by nnU-Net. The default setting for all modalities except CT images is z-scoring. For this option, during training and inference, each image is normalized independently by subtracting its mean, followed by division with its standard deviation. If cropping resulted in an average size decrease of 25% or more, a mask for central non-zero voxels is created and the normalization is applied within that mask only, ignoring the surrounding zero voxels. For computed tomography (CT) images, nnU-Net employs a different scheme, as intensity values are quantitative and reflect physical properties of the tissue. It can therefore be beneficial to retain this information by using a global normalization scheme that is applied to all images. To this end, nnU-Net uses the 0.5 and 99.5 percentiles of the foreground voxels for clipping as well as the global foreground mean and standard deviation for normalization on all images.

Resampling: In some datasets, particularly in the medical domain, the voxel spacing (the physical space the voxels represent) is heterogeneous. Convolutional neural networks operate on voxel grids and ignore this information. To cope with this heterogeneity, nnU-Net resamples all images to the same target spacing (see paragraph below) using either third order spline, linear or nearest neighbor interpolation. The default setting for image data is third order spline interpolation. For anisotropic images (maximum axis spacing / minimum axis spacing ¿ 3), in-plane resampling is done with third order spline whereas out of plane interpolation is done with nearest neighbor. Segmentation maps are resampled by converting them to one hot encodings. Each channel is then interpolated with linear interpolation and the segmentation mask is retrieved by an argmax operation. Again, anisotropic cases are interpolated using "nearest neighbor" on the low resolution axis.

Target spacing: The selected target spacing is a crucial parameter. Larger spacings result

in smaller images and thus a loss of details whereas smaller spacings result in larger images preventing the network from accumulating sufficient contextual information under a given patch size. Although this tradeoff is in part addressed by the 3D U-Net cascade (see below), a sensible target spacing for low and full resolution is still required. For the 3D full resolution U-Net, nnU-Net uses the median value of the spacings found in the training cases computed independently for each axis as default target spacing. For anisotropic datasets, this default can result in severe interpolation artifacts or in a substantial loss of information due to large variances in resolution across the training data. Therefore, the target spacing of the lowest resolution axis is selected to be the 10th percentile of the spacings found in the training cases if both voxel and spacing anisotropy (i.e. the ratio of lowest spacing axis to highest spacing axis) are larger than 3. For the 2D U-Net, nnU-Net generally operates on the two axes with the highest resolution. If all three axes are isotropic, the two trailing axes are utilized for slice extraction. The target spacing is the median spacing of the training cases (computed independently for each axis). For slice-based processing, no resampling along the out-of-plane axis is required.

Adaptation of network topology, patch size and batch size: Finding an appropriate U-Net architecture configuration is crucial for good segmentation performance. nnU-Net prioritizes large patch sizes while remaining within a predefined GPU memory budget. Larger patch sizes allow for more contextual information to be aggregated and thus typically increase segmentation performance. They come, however, at the cost of a decreased batch size which results in noisier gradients during backpropagation. To improve the stability of the training, we require a minimum batch size of 2 and choose a large momentum term for network training (see blueprint parameters). Image spacing is also considered in the adaptation process: Downsampling operations may operate only on specific axes and convolutional kernels in the 3D U-Nets can operate on certain image planes only (pseudo-2D). The network topology for all U-Net configurations is chosen on basis of the median image size after resampling as well as the target spacing the images were resampled to. A flow chart for the adaptation process is presented in the Supplements inFigure S1a. The adaptation of the architecture template, which is described in more detail in the following, is computationally inexpensive. Due to the GPU memory consumption estimate being based on feature map sizes, no GPU is required to run the adaptation process.
Initialization: The patch size is initialized as the median image shape after resampling. If the patch size is not divisible by 2^{n_d} for each axis, where n_d is the number of downsampling operations, it is padded accordingly. Architecture topology: The architecture is configured by determining the number of downsampling operations along each axis depending on the patch size and voxel spacing. Downsampling is performed until further downsampling would reduce the feature map size to smaller than 4 voxels or the feature map spacings become anisotropic. The downsampling strategy is determined by the voxel spacing: high resolution axes are downsampled separately until their resolution is within

factor 2 of the lower resolution axis. Subsequently, all axes are downsampled simultaneously. Downsampling is terminated for each axis individually, once the respective feature map constraint is triggered. The default kernel size for convolutions is $3 \times 3 \times 3$ and 3×3 for 3D U-Net and 2D U-Net, respectively. If there is an initial resolution discrepancy between axes (defined as a spacing ratio larger than 2), the kernel size for the out-of-plane axis is set to 1 until the resolutions are within a factor of 2. Note that the convolutional kernel size then remains at 3 for all axes.

Adaptation to GPU memory budget: The largest possible patch size during configuration is limited by the amount of GPU memory. Since the patch size is initialized to the median image shape after resampling, it is initially too large to fit into the GPU for most datasets. nnU-Net estimates the memory consumption of a given architecture based on the size of the feature maps in the network, comparing it to reference values of known memory consumption. The patch size is then reduced in an iterative process while updating the architecture configuration accordingly in each step until the required budget is reached (see Figure S1 in the Supplements). The reduction of the patch size is always applied to the largest axis relative to the median image shape of the data. The reduction in one step amounts to 2^{n_d} voxels of that axis, where n_d is the number of downsampling operations.

Batch size: As a final step, the batch size is configured. If a reduction of patch size was performed the batch size is set to 2. Otherwise, the remaining GPU memory headroom is utilized to increase the batch size until the GPU is fully utilized. To prevent overfitting, the batch size is capped such that the total number of voxels in the minibatch do not exceed 5% of the total number of voxels of all training cases. Examples for generated U-Net architectures are presented in Supplementary Information ZZ.

Configuration of 3D U-Net cascade: Running a segmentation model on downsampled data increases the size of patches in relation to the image and thus enables the network to accumulate more contextual information. This comes at the cost of a reduction in details in the generated segmentations and may also cause errors if the segmentation target is very small or characterized by its texture. In a hypothetical scenario with unlimited GPU memory, it is thus generally favored to train models at full resolution with a patch size that covers the entire image. The 3D U-Net cascade approximates this approach by first running a 3D U-Net on downsampled images and then training a second, full resolution 3D U-Net to refine the segmentation maps of the former. This way, the "global", low resolution network utilizes maximal contextual information to generate its segmentation output, which then serves as an additional input channel that guides the second, "local" U-Net. The cascade is triggered only for datasets where the patch size of the 3d full resolution U-Net covers less than 12.5% of the median image shape. If this is the case,

the target spacing for the downsampled data and the architecture of the associated 3D low resolution U-Net are configured jointly in an iterative process. The target spacing is initialized as the target spacing of the full resolution data. In order for the patch size to cover a large proportion of the input image, the target spacing is then increased stepwise by 1% while updating the architecture configuration accordingly in each step until the patch size of the resulting network topology surpasses 25% of the current median image shape. If the current spacing is anisotropic (factor 2 difference between lowest and highest resolution axis), only the spacing of the higher resolution axes is increased. The configuration of the second 3D U-Net of the cascade is identical to the standalone 3D U-Net for which the configuration process is described above (except that the upsampled segmentation maps of the first U-Net are concatenated to its input). Figure S1b in the Supplements provides an overview of this optimization process.

Empirical parameters

Ensembling and selection of U-Net configuration(s): nnU-Net automatically determines which (ensemble of) configuration(s) to use for inference based on the average foreground Dice coefficient computed via cross-validation on the training data. The selected model(s) can be either a single U-Net (2D, 3D full resolution, 3D low resolution or the full resolution U-Net of the cascade) or an ensemble of any two of these configurations. Models are ensembled by averaging softmax probabilities.

Postprocessing: Connected component-based postprocessing is commonly used in medical image segmentation [178, 233]. Especially in organ segmentation it often helps to remove spurious false positive detections by removing all but the largest connected component. nnU-Net follows this assumption and automatically benchmarks the effect of suppressing smaller components on the cross-validation results. First, all foreground classes are treated as one component. If suppression of all but the largest region improves the average foreground Dice coefficient and does not reduce the Dice coefficient for any of the classes, this procedure is selected as the first postprocessing step. Finally, nnU-Net builds on the outcome of this step and decides whether the same procedure should be performed for individual classes.

Implementation details

nnU-Net is implemented in Python utilizing the PyTorch [291] framework. The Batchgenerators library [292] is used for data augmentation. For reduction of computational burden and GPU memory footprint, mixed precision training is implemented with Nvidia Apex/Amp (`https://github.com/NVIDIA/apex`). For use as a framework, the source code is available on GitHub (`https://github.com/MIC-DKFZ/nnUNet`). Users who seek to use nnU-Net as a standardized benchmark or to run inference with our pretrained

models can install nnU-Net via PyPi. For a full description of how to use nnU-Net, please refer to the online documentation available on the GitHub page.

D.2. Qualitative Results of nnU-Net

Figure D.1 shows segmentation generated by nnU-Net on a variety of datasets and modalities. Test set target structures from different international segmentation challenges are shown in 2D projected onto the raw data (left) and in 3D together with a volume rendering of the raw data (right). All visualizations are created with the MITK Workbench [148].

D.3. Details of Datasets utilized for nnU-Net evaluation

Table D.1 provides an overview of the datasets used in this manuscript. The numeric values presented here are computed based on the training cases for each of these datasets. They are the basis of the dataset fingerprints presented in Figure 7.7.

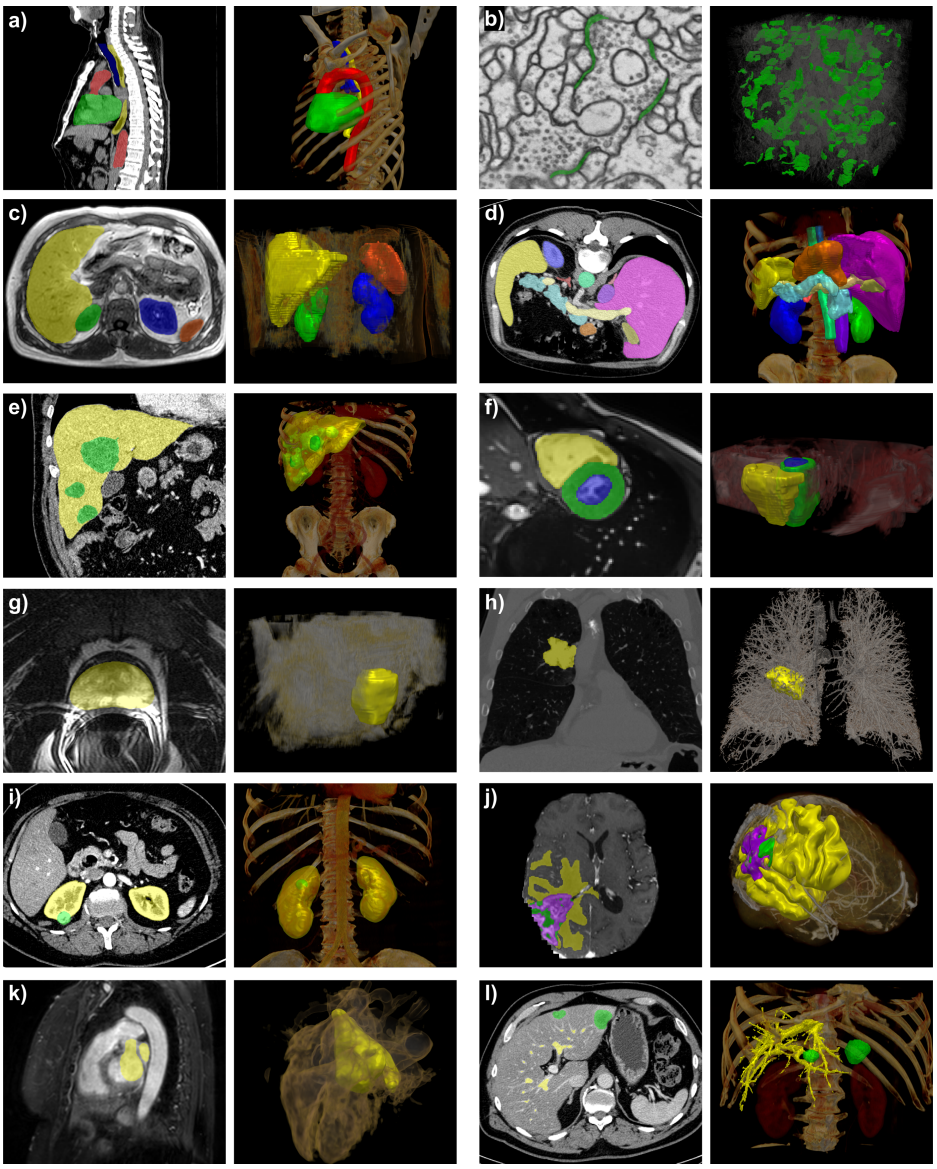

Figure D.1.: **nnU-Net Handles a Broad Variety of Datasets and Target Image Properties.** Dataset references point to Table D.1. a: heart (green), aorta (red), trachea (blue) and esophagus (yellow) in CT images (D18). b: synaptic clefts (green) in electron microscopy scans (D19). c: liver (yellow), spleen (orange), left/right kidney (blue/green) in T1 in-phase MRI (D16). d: thirteen abdominal organs in CT images (D11). e: liver (yellow) and liver tumors (green) in CT images (D14). f: right ventricle (yellow), left ventricular cavity (blue) myocardium of left ventricle (green) in cine MRI (D13). g: prostate (yellow) in T2 MRI (D12). h: lung nodules (yellow) in CT images (D6). i: kidneys (yellow) and kidney tumors (green) in CT images (D17). j: edema (yellow), enhancing tumor (purple), necrosis (green) in MRI (T1, T1 with contrast agent, T2, FLAIR) (D1). k: left ventricle (yellow) in MRI (D2). l: hepatic vessels (yellow) and liver tumors (green) in CT (D8).

Table D.1.: **Overview over the challenge datasets nnU-Net was evaluated on.**

ID	Dataset Name	Associated Challenges	Modalities	Median Shape (Spacing [mm])	N Classes	Rarest Class Ratio	N Training Cases	Segmentation Tasks
D1	Brain Tumour	[173], [238]	MRI (T1, T1c, T2, FLAIR)	138x169x138 (1, 1, 1)	3	7.310^{-3}	484	edema, active tumor, necrosis
D2	Heart	[173]	MRI	115x320x232 (1.37, 1.25, 1.25)	1	4.010^{-3}	20	left ventricle
D3	Liver	[173], [178]	CT	432x512x512 (1, 0.77, 0.77)	2	2.610^{-2}	131	liver, liver tumors
D4	Hippocampus	[173]	MRI	36x50x35 (1, 1, 1)	2	2.710^{-2}	260	anterior and posterior hippocampus
D5	Prostate	[173]	MRI (T2, ADC)	20x320x319 (3.6, 0.62, 0.62)	2	5.410^{-3}	32	peripheral and transition zone
D6	Lung	[173]	CT	252x512x512 (1.24, 0.79, 0.79)	1	3.910^{-4}	63	lung nodules
D7	Pancreas	[173]	CT	93x512x512 (2.5, 0.80, 0.80)	2	2.010^{-3}	282	pancreas, pancreas cancer
D8	HepaticVessel	[173]	CT	49x512x512 (5, 0.80, 0.80)	2	1.110^{-3}	303	hepatic vessels, tumors
D9	Spleen	[173]	CT	90x512x512 (5, 0.79, 0.79)	1	4.710^{-3}	41	spleen
D10	Colon	[173]	CT	95x512x512 (5, 0.78, 0.78)	1	5.610^{-4}	126	colon cancer
D11	AbdOrgSeg	[293]	CT	128x512x512 (3, 0.76, 0.76)	13	4.410^{-3}	30	13 abdominal organs
D12	Promise	[294]	MRI	24x320x320 (3.6, 0.61, 0.61)	1	2.010^{-2}	50	prostate
D13	ACDC	[34]	cine MRI	9x256x216 (10, 1.56, 1.56)	3	1.210^{-2}	200 (100x2) *	left ventricle, right ventricle, myocardium
D14	LiTS **	[178]	CT	432x512x512 (1, 0.77, 0.77)	2	2.610^{-2}	131	liver, liver tumors
D15	MSLesion	[295]	MRI (FLAIR, MPRAGE, PD, T2)	137x180x137 (1, 1, 1)	1	1.710^{-3}	42 (21x2) *	multiple sclerosis lesions
D16	CHAOS	[296]	MRI	30x204x256 (9, 1.66, 1.66)	4	3.310^{-2}	60 (20 + 20x2) *	liver, spleen, left and right kidney
D17	KiTS	[233]	CT	107x512x512 (3, 0.78, 0.78)	2	7.510^{-3}	206	kidney, kidney tumor
D18	SegTHOR	[297]	CT	178x512x512 (2.5, 0.98, 0.98)	4	4.610^{-4}	40	heart, aorta, esophagus, trachea
D19	CREMI	[242]	Electron Microscopy	125x1250x1250 (40, 4, 4)	1	5.210^{-3}	3	synaptic clefts

* multiple annotated examples per training case
** almost identical to Decathlon Liver; Decathlon changed the training cases and test set slightly

Bibliography

[1] G. S. Lodwick, C. L. Haun, W. E. Smith, R. F. Keller, and E. D. Robertson, "Computer Diagnosis of Primary Bone Tumors," *Radiology*, vol. 80, no. 2, pp. 273–275, Feb. 1963.

[2] P. H. Meyers, C. M. Nice, H. C. Becker, W. J. Nettleton, J. W. Sweeney, and G. R. Meckstroth, "AUTOMATED COMPUTER ANALYSIS OF RADIOGRAPHIC IMAGES," *Radiology*, vol. 83, pp. 1029–1034, Dec. 1964.

[3] F. Winsberg, M. Elkin, J. Macy, V. Bordaz, and W. Weymouth, "Detection of Radiographic Abnormalities in Mammograms by Means of Optical Scanning and Computer Analysis," *Radiology*, vol. 89, no. 2, pp. 211–215, Aug. 1967.

[4] Dell and EMC, "Next-generation infrastructure with no-compromise storage and data protection for today and tomorrow," Tech. Rep., 2019.

[5] "Clinical radiology UK workforce census report 2018 — The Royal College of Radiologists," https://www.rcr.ac.uk/publication/clinical-radiology-uk-workforce-census-report-2018.

[6] E. Sokolovskaya, T. Shinde, R. B. Ruchman, A. J. Kwak, S. Lu, Y. K. Shariff, E. F. Wiggins, and L. Talangbayan, "The Effect of Faster Reporting Speed for Imaging Studies on the Number of Misses and Interpretation Errors: A Pilot Study," *Journal of the American College of Radiology*, vol. 12, no. 7, pp. 683–688, Jul. 2015.

[7] B. I. Reiner, N. Knight, and E. L. Siegel, "Radiology reporting, past, present, and future: The radiologist's perspective," *Journal of the American College of Radiology: JACR*, vol. 4, no. 5, pp. 313–319, May 2007.

[8] N. R. C. U. C. o. A. F. f. D. a. N. T. of Disease, *Toward Precision Medicine*. National Academies Press (US), 2011.

[9] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using

a quantitative radiomics approach," *Nature Communications*, vol. 5, p. 4006, Jun. 2014.

[10] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data," *Radiology*, vol. 278, no. 2, pp. 563–577, Nov. 2015.

[11] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. de Jong, J. van Timmeren, S. Sanduleanu, R. T. H. M. Larue, A. J. G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F. M. Mottaghy, J. E. Wildberger, and S. Walsh, "Radiomics: The bridge between medical imaging and personalized medicine," *Nature Reviews Clinical Oncology*, vol. 14, no. 12, pp. 749–762, Dec. 2017.

[12] K. Doi, "Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential," *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, vol. 31, no. 4-5, pp. 198–211, 2007.

[13] J. Taylor and J. Fenner, "Clinical Adoption of CAD: Exploration of the Barriers to Translation through an Example Application," *Procedia Computer Science*, vol. 90, pp. 93–98, Jan. 2016.

[14] A. Kohli and S. Jha, "Why CAD Failed in Mammography," *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 535–537, Mar. 2018.

[15] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F. J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C. J. Kelly, D. King, J. R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J. J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K. C. Young, J. De Fauw, and S. Shetty, "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, Jan. 2020.

[16] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[17] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, no. 1, pp. 65–69, Jan. 2019.

[18] A. Majkowska, S. Mittal, D. F. Steiner, J. J. Reicher, S. M. McKinney, G. E. Duggan, K. Eswaran, P.-H. Cameron Chen, Y. Liu, S. R. Kalidindi, A. Ding, G. S. Corrado,

D. Tse, and S. Shetty, "Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation," *Radiology*, vol. 294, no. 2, pp. 421–431, Dec. 2019.

[19] E. A. McGlynn, K. M. McDonald, and C. K. Cassel, "Measurement Is Essential for Improving Diagnosis and Reducing Diagnostic Error: A Report From the Institute of Medicine," *JAMA*, vol. 314, no. 23, pp. 2501–2502, Dec. 2015.

[20] A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn, and D. Koller, "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Science Translational Medicine*, vol. 3, no. 108, p. 108ra113, Nov. 2011.

[21] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature Biomedical Engineering*, vol. 2, no. 3, pp. 158–164, Mar. 2018.

[22] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, Jan. 2019.

[23] "IBM's Watson recommended 'unsafe and incorrect' cancer treatments," *https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/*, Jul. 2018.

[24] P. A. Keane and E. J. Topol, "With an eye to AI and autonomous diagnosis," *npj Digital Medicine*, vol. 1, no. 1, pp. 1–3, Aug. 2018.

[25] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, Jan. 2019.

[26] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. A. Ioannidis, G. S. Collins, and M. Maruthappu, "Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies," *BMJ*, p. m689, Mar. 2020.

[27] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, p. 195, Oct. 2019.

[28] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis, "A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, Apr. 2020, pp. 1–12.

[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, ser. Adaptive Computation and Machine Learning. Cambridge, Massachusetts: The MIT Press, 2016.

[30] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017.

[31] T. M. Mitchell, "The Need for Biases in Learning Generalizations," *Computer Scoence Deparment, Rutgers University, CBM-TR-117*, p. 3, 1980.

[32] S. Bickelhaupt*, P. F. Jaeger*, F. B. Laun, W. Lederer, H. Daniel, T. A. Kuder, L. Wuesthof, D. Paech, D. Bonekamp, A. Radbruch, S. Delorme, H.-P. Schlemmer, F. H. Steudle, and K. H. Maier-Hein, "Radiomics Based on Adapted Diffusion Kurtosis Imaging Helps to Clarify Most Mammographic Findings Suspicious for Cancer," *Radiology*, vol. 287, no. 3, pp. 761–770, Jun. 2018.

[33] F. Isensee*, P. F. Jaeger*, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic Cardiac Disease Assessment on cine-MRI via Time-Series Segmentation and Domain Specific Features," in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, ser. Lecture Notes in Computer Science, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, Eds. Cham: Springer International Publishing, 2018, pp. 120–129. *equal contribution.

[34] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M.-M. Rohe, X. Pennec, M. Sermesant, F. Isensee, P. Jager, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Isgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P.-M. Jodoin, "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

[35] P. F. Jaeger, S. Bickelhaupt, F. B. Laun, W. Lederer, D. Heidi, T. A. Kuder, D. Paech, D. Bonekamp, A. Radbruch, S. Delorme, H.-P. Schlemmer, F. Steudle, and K. H. Maier-Hein, "Revealing Hidden Potentials of the q-Space Signal in Breast Cancer," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, ser. Lecture Notes in Computer Science, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham: Springer International Publishing, 2017, pp. 664–671.

[36] P. F. Jaeger, S. Bickelhaupt, F. B. Laun, W. Lederer, H. Daniel, T. A. Kuder, S. Delorme, H.-P. Schlemmer, F. Steudle, and K. H. Maier-Hein, "Complementary

value of End-to-end Deep Learning and Radiomics in Breast Cancer Classification on Diffusion-Weighted MRI," in *International Society of Magnetic Resonance in Medicine Annual Meeting*, 2018.

[37] P. F. Jaeger, S. A. A. Kohl, S. Bickelhaupt, F. Isensee, T. A. Kuder, H.-P. Schlemmer, and K. H. Maier-Hein, "Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection," in *Machine Learning for Health Workshop*, Apr. 2020, pp. 171–183.

[38] P. Jaeger, G. Ramien, and K. Maier-Hein, "Medical Detection Toolkit," Zenodo, Jan. 2020.

[39] T. Weikert*, P. F. Jaeger*, J. Bremerich, G. Sommer, B. Stieltjes, S. Yang, K. H. Maier-Hein, and A. Sauter, "Evaluation of different training approaches for the automated detection of lung cancer of all stages on FDG-PET/CT using a Retina U-Net algorithm," *In Submission*, 2020.

[40] G. N. Ramien, P. F. Jaeger, S. A. A. Kohl, and K. H. Maier-Hein, "Reg R-CNN: Lesion Detection and Grading Under Noisy Labels," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures*, ser. Lecture Notes in Computer Science, H. Greenspan, R. Tanno, M. Erdt, T. Arbel, C. Baumgartner, A. Dalca, C. H. Sudre, W. M. Wells, K. Drechsler, M. G. Linguraru, C. Oyarzun Laura, R. Shekhar, S. Wesarg, and M. Á. González Ballester, Eds. Cham: Springer International Publishing, 2019, pp. 33–41.

[41] J. Kamphenkel*, P. F. Jäger*, S. Bickelhaupt, F. B. Laun, W. Lederer, H. Daniel, T. A. Kuder, S. Delorme, H.-P. Schlemmer, F. König, and K. H. Maier-Hein, "Domain Adaptation for Deviating Acquisition Protocols in CNN-Based Lesion Classification on Diffusion-Weighted MR Images," in *Image Analysis for Moving Organ, Breast, and Thoracic Images*, ser. Lecture Notes in Computer Science, D. Stoyanov, Z. Taylor, B. Kainz, G. Maicas, R. R. Beichel, A. Martel, L. Maier-Hein, K. Bhatia, T. Vercauteren, O. Oktay, G. Carneiro, A. P. Bradley, J. Nascimento, H. Min, M. S. Brown, C. Jacobs, B. Lassen-Schmidt, K. Mori, J. Petersen, R. San José Estépar, A. Schmidt-Richberg, and C. Veiga, Eds. Cham: Springer International Publishing, 2018, pp. 73–80. *equal contribution.

[42] F. Isensee*, P. F. Jäger*, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Automated Design of Deep Learning Methods for Biomedical Image Segmentation," *Under Review at Nature Methods*, Apr. 2020.

[43] W. Schlegel, C. P. Karger, and O. Jäkel, Eds., *Medizinische Physik: Grundlagen – Bildgebung – Therapie – Technik*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2018.

[44] A. Radbruch, L. D. Weberling, P. J. Kieslich, O. Eidel, S. Burth, P. Kickingereder, S. Heiland, W. Wick, H.-P. Schlemmer, and M. Bendszus, "Gadolinium Retention in the Dentate Nucleus and Globus Pallidus Is Dependent on the Class of Contrast Agent," *Radiology*, vol. 275, no. 3, pp. 783–791, Apr. 2015.

[45] T. Kanda, K. Ishii, H. Kawaguchi, K. Kitajima, and D. Takenaka, "High Signal Intensity in the Dentate Nucleus and Globus Pallidus on Unenhanced T1-weighted MR Images: Relationship with Increasing Cumulative Dose of a Gadolinium-based Contrast Material," *Radiology*, vol. 270, no. 3, pp. 834–841, Dec. 2013.

[46] Y. Errante, V. Cirimele, C. A. Mallio, V. Di Lazzaro, B. B. Zobel, and C. C. Quattrocchi, "Progressive Increase of T1 Signal Intensity of the Dentate Nucleus on Unenhanced Magnetic Resonance Images Is Associated With Cumulative Doses of Intravenously Administered Gadodiamide in Patients With Normal Renal Function, Suggesting Dechelation:," *Investigative Radiology*, vol. 49, no. 10, pp. 685–690, Oct. 2014.

[47] R. J. McDonald, J. S. McDonald, D. F. Kallmes, M. E. Jentoft, D. L. Murray, K. R. Thielen, E. E. Williamson, and L. J. Eckel, "Intracranial Gadolinium Deposition after Contrast-enhanced MR Imaging," *Radiology*, vol. 275, no. 3, pp. 772–782, Mar. 2015.

[48] E. O. Stejskal and J. E. Tanner, "Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient," *The Journal of Chemical Physics*, vol. 42, no. 1, pp. 288–292, Jan. 1965.

[49] D. Le Bihan and E. Breton, "Imagerie de diffusion in-vivo par résonance magnétique nucléaire," *Comptes-Rendus de l'Académie des Sciences*, vol. 93, no. 5, pp. 27–34, Dec. 1985.

[50] J. H. Jensen, J. A. Helpern, A. Ramani, H. Lu, and K. Kaczynski, "Diffusional kurtosis imaging: The quantification of non-gaussian water diffusion by means of magnetic resonance imaging," *Magnetic Resonance in Medicine*, vol. 53, no. 6, pp. 1432–1440, 2005.

[51] J. H. Jensen and J. A. Helpern, "MRI quantification of non-Gaussian water diffusion by kurtosis analysis," *NMR in biomedicine*, vol. 23, no. 7, pp. 698–710, Aug. 2010.

[52] P. Raab, E. Hattingen, K. Franz, F. E. Zanella, and H. Lanfermann, "Cerebral gliomas: Diffusional kurtosis imaging analysis of microstructural differences," *Radiology*, vol. 254, no. 3, pp. 876–881, Mar. 2010.

[53] S. Van Cauter, J. Veraart, J. Sijbers, R. R. Peeters, U. Himmelreich, F. De Keyzer, S. W. Van Gool, F. Van Calenbergh, S. De Vleeschouwer, W. Van Hecke, and S. Sunaert, "Gliomas: Diffusion kurtosis MR imaging in grading," *Radiology*, vol. 263, no. 2, pp. 492–501, May 2012.

[54] A. B. Rosenkrantz, E. E. Sigmund, A. Winnick, B. E. Niver, B. Spieler, G. R. Morgan, and C. H. Hajdu, "Assessment of hepatocellular carcinoma using apparent diffusion coefficient and diffusion kurtosis indices: Preliminary experience in fresh liver explants," *Magnetic Resonance Imaging*, vol. 30, no. 10, pp. 1534–1540, Dec. 2012.

[55] A. B. Rosenkrantz, E. E. Sigmund, G. Johnson, J. S. Babb, T. C. Mussi, J. Melamed, S. S. Taneja, V. S. Lee, and J. H. Jensen, "Prostate cancer: Feasibility and preliminary experience of a diffusional kurtosis model for detection and assessment of aggressiveness of peripheral zone cancer," *Radiology*, vol. 264, no. 1, pp. 126–135, Jul. 2012.

[56] S. Suo, X. Chen, L. Wu, X. Zhang, Q. Yao, Y. Fan, H. Wang, and J. Xu, "Non-Gaussian water diffusion kurtosis imaging of prostate cancer," *Magnetic Resonance Imaging*, vol. 32, no. 5, pp. 421–427, Jun. 2014.

[57] L. Nogueira, S. Brandão, E. Matos, R. G. Nunes, J. Loureiro, I. Ramos, and H. A. Ferreira, "Application of the diffusion kurtosis model for the study of breast lesions," *European Radiology*, vol. 24, no. 6, pp. 1197–1203, Jun. 2014.

[58] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[59] R. Etzioni, N. Urban, S. Ramsey, M. McIntosh, S. Schwartz, B. Reid, J. Radich, G. Anderson, and L. Hartwell, "The case for early detection," *Nature Reviews. Cancer*, vol. 3, no. 4, pp. 243–252, Apr. 2003.

[60] C. K. Kuhl, "Abbreviated Magnetic Resonance Imaging (MRI) for Breast Cancer Screening: Rationale, Concept, and Transfer to Clinical Practice," *Annual Review of Medicine*, vol. 70, no. 1, pp. 501–519, 2019.

[61] A. C. Society, "Cancer Facts & Figures 2020," 2020.

[62] N. F. Boyd, J. M. Rommens, K. Vogt, V. Lee, J. L. Hopper, M. J. Yaffe, and A. D. Paterson, "Mammographic breast density as an intermediate phenotype for breast cancer," *The Lancet Oncology*, vol. 6, no. 10, pp. 798–808, Oct. 2005.

[63] https://drandrewkiu.com.au/breast-surgery-adelaide/.

[64] https://www.bebrcaware.com/breast-cancer-the-brca-link/what-is-breast-cancer.html.

[65] American College of Radiology and others, "Breast imaging reporting and data system," *BI-RADS*, 2003.

[66] A. N. A. Tosteson, D. G. Fryback, C. S. Hammond, L. G. Hanna, M. R. Grove, M. Brown, Q. Wang, K. Lindfors, and E. D. Pisano, "Consequences of false-positive screening mammograms," *JAMA internal medicine*, vol. 174, no. 6, pp. 954–961, Jun. 2014.

[67] L. Tabár, B. Vitak, T. H.-H. Chen, A. M.-F. Yen, A. Cohen, T. Tot, S. Y.-H. Chiu, S. L.-S. Chen, J. C.-Y. Fann, J. Rosell, H. Fohlin, R. A. Smith, and S. W. Duffy, "Swedish two-county trial: Impact of mammographic screening on breast cancer mortality during 3 decades," *Radiology*, vol. 260, no. 3, pp. 658–663, Sep. 2011.

[68] M. Tonelli, S. Connor Gorber, M. Joffres, J. Dickinson, H. Singh, G. Lewin, R. Birtwhistle, D. Fitzpatrick-Lewis, N. Hodgson, D. Ciliska, M. Gauld, Y. Y. Liu, and Canadian Task Force on Preventive Health Care, "Recommendations on screening for breast cancer in average-risk women aged 40-74 years," *CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne*, vol. 183, no. 17, pp. 1991–2001, Nov. 2011.

[69] R. A. Smith, K. S. Andrews, D. Brooks, S. A. Fedewa, D. Manassaram-Baptiste, D. Saslow, O. W. Brawley, and R. C. Wender, "Cancer screening in the United States, 2018: A review of current American Cancer Society guidelines and current issues in cancer screening," *CA: a cancer journal for clinicians*, vol. 68, no. 4, pp. 297–316, Jul. 2018.

[70] M. G. Marmot, D. G. Altman, D. A. Cameron, J. A. Dewar, S. G. Thompson, and M. Wilcox, "The benefits and harms of breast cancer screening: An independent review," *British Journal of Cancer*, vol. 108, no. 11, pp. 2205–2240, Jun. 2013.

[71] E. R. Myers, P. Moorman, J. M. Gierisch, L. J. Havrilesky, L. J. Grimm, S. Ghate, B. Davidson, R. C. Mongtomery, M. J. Crowley, D. C. McCrory, A. Kendrick, and G. D. Sanders, "Benefits and Harms of Breast Cancer Screening: A Systematic Review," *JAMA*, vol. 314, no. 15, pp. 1615–1634, Oct. 2015.

[72] C. D. Lehman, R. F. Arao, B. L. Sprague, J. M. Lee, D. S. M. Buist, K. Kerlikowske, L. M. Henderson, T. Onega, A. N. A. Tosteson, G. H. Rauscher, and D. L. Miglioretti, "National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium," *Radiology*, vol. 283, no. 1, pp. 49–58, Apr. 2017.

[73] R. A. Smith, S. W. Duffy, R. Gabe, L. Tabar, A. M. F. Yen, and T. H. H. Chen, "The randomized trials of breast cancer screening: What have we learned?" *Radiologic Clinics*, vol. 42, no. 5, pp. 793–806, Sep. 2004.

[74] B. CMgKM, "Results of the MammographyScreening-Program in Germany," Berlin, Tech. Rep., 2010.

[75] P. Autier, M. Boniol, A. Gavin, and L. J. Vatten, "Breast cancer mortality in neighbouring European countries with different levels of screening but similar access to treatment: Trend analysis of WHO mortality database," *BMJ (Clinical research ed.)*, vol. 343, p. d4411, Jul. 2011.

[76] N. K. Stout, S. J. Lee, C. B. Schechter, K. Kerlikowske, O. Alagoz, D. Berry, D. S. M. Buist, M. Cevik, G. Chisholm, H. J. de Koning, H. Huang, R. A. Hubbard, D. L. Miglioretti, M. F. Munsell, A. Trentham-Dietz, N. T. van Ravesteyn, A. N. A. Tosteson, and J. S. Mandelblatt, "Benefits, harms, and costs for breast cancer screening after US implementation of digital mammography," *Journal of the National Cancer Institute*, vol. 106, no. 6, p. dju092, Jun. 2014.

[77] R. J. Brenner and E. A. Sickles, "Acceptability of periodic follow-up as an alternative to biopsy for mammographically detected lesions interpreted as probably benign," *Radiology*, vol. 171, no. 3, pp. 645–646, Jun. 1989.

[78] B. L. Sprague, R. E. Gangnon, V. Burt, A. Trentham-Dietz, J. M. Hampton, R. D. Wellman, K. Kerlikowske, and D. L. Miglioretti, "Prevalence of Mammographically Dense Breasts in the United States," *JNCI: Journal of the National Cancer Institute*, vol. 106, no. 10, Oct. 2014.

[79] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018," *CA: a cancer journal for clinicians*, vol. 68, no. 1, pp. 7–30, Jan. 2018.

[80] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, M. Broeders, G. Gennaro, P. Clauser, T. H. Helbich, M. Chevalier, T. Tan, T. Mertelmeier, M. G. Wallis, I. Andersson, S. Zackrisson, R. M. Mann, and I. Sechopoulos, "Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists," *Journal of the National Cancer Institute*, vol. 111, no. 9, pp. 916–922, Sep. 2019.

[81] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzebski, T. Fevry, J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, L. L. Y. Lin, K. Ho, J. D. Weinstein, B. Reig, Y. Gao, H. Toth, K. Pysarenko, A. Lewin, J. Lee, K. Airola, E. Mema, S. Chung, E. Hwang, N. Samreen, S. G. Kim, L. Heacock, L. Moy, K. Cho, and K. J. Geras, "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening," *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1184–1194, Apr. 2020.

[82] M. Kriege, C. T. M. Brekelmans, C. Boetes, P. E. Besnard, H. M. Zonderland, I. M. Obdeijn, R. A. Manoliu, T. Kok, H. Peterse, M. M. A. Tilanus-Linthorst, S. H. Muller, S. Meijer, J. C. Oosterwijk, L. V. A. M. Beex, R. A. E. M. Tollenaar, H. J. de Koning, E. J. T. Rutgers, J. G. M. Klijn, and Magnetic Resonance Imaging Screening Study Group, "Efficacy of MRI and mammography for breast-cancer screening in

women with a familial or genetic predisposition," *The New England Journal of Medicine*, vol. 351, no. 5, pp. 427–437, Jul. 2004.

[83] E. Warner, D. B. Plewes, K. A. Hill, P. A. Causer, J. T. Zubovits, R. A. Jong, M. R. Cutrara, G. DeBoer, M. J. Yaffe, S. J. Messner, W. S. Meschino, C. A. Piron, and S. A. Narod, "Surveillance of BRCA1 and BRCA2 mutation carriers with magnetic resonance imaging, ultrasound, mammography, and clinical breast examination," *JAMA*, vol. 292, no. 11, pp. 1317–1325, Sep. 2004.

[84] C. K. Kuhl, K. Strobel, H. Bieling, C. Leutner, H. H. Schild, and S. Schrading, "Supplemental Breast MR Imaging Screening of Women with Average Risk of Breast Cancer," *Radiology*, vol. 283, no. 2, pp. 361–370, May 2017.

[85] S. Bickelhaupt, F. B. Laun, J. Tesdorff, W. Lederer, H. Daniel, A. Stieber, S. Delorme, and H.-P. Schlemmer, "Fast and Noninvasive Characterization of Suspicious Lesions Detected at Breast Cancer X-Ray Screening: Capability of Diffusion-weighted MR Imaging with MIPs," *Radiology*, vol. 278, no. 3, pp. 689–697, Mar. 2016.

[86] I. Thomassin-Naggara, C. De Bazelaire, J. Chopier, M. Bazot, C. Marsault, and I. Trop, "Diffusion-weighted MR imaging of the breast: Advantages and pitfalls," *European Journal of Radiology*, vol. 82, no. 3, pp. 435–443, Mar. 2013.

[87] K. Sun, X. Chen, W. Chai, X. Fei, C. Fu, X. Yan, Y. Zhan, K. Chen, K. Shen, and F. Yan, "Breast Cancer: Diffusion Kurtosis MR Imaging—Diagnostic Accuracy and Correlation with Clinical-Pathologic Factors," *Radiology*, vol. 277, no. 1, pp. 46–55, May 2015.

[88] W. A. Berg, L. Gutierrez, M. S. NessAiver, W. B. Carter, M. Bhargavan, R. S. Lewis, and O. B. Ioffe, "Diagnostic Accuracy of Mammography, Clinical Examination, US, and MR Imaging in Preoperative Assessment of Breast Cancer," *Radiology*, vol. 233, no. 3, pp. 830–849, Dec. 2004.

[89] C. Boetes, S. P. Strijk, R. Holland, J. O. Barentsz, R. F. Van Der Sluis, and J. H. Ruijs, "False-negative MR imaging of malignant breast tumors," *European Radiology*, vol. 7, no. 8, pp. 1231–1234, 1997.

[90] T. Wan, B. N. Bloch, D. Plecha, C. L. Thompson, H. Gilmore, C. Jaffe, L. Harris, and A. Madabhushi, "A Radio-genomics Approach for Identifying High Risk Estrogen Receptor-positive Breast Cancers on DCE-MRI: Preliminary Results in Predicting OncotypeDX Risk Scores," *Scientific Reports*, vol. 6, Feb. 2016.

[91] S. Yamamoto, W. Han, Y. Kim, L. Du, N. Jamshidi, D. Huang, J. H. Kim, and M. D. Kuo, "Breast Cancer: Radiogenomic Biomarker Reveals Associations among Dynamic Contrast-enhanced MR Imaging, Long Noncoding RNA, and Metastasis," *Radiology*, vol. 275, no. 2, pp. 384–392, May 2015.

[92] Y. Zhu, H. Li, W. Guo, K. Drukker, L. Lan, M. L. Giger, and Y. Ji, "Deciphering Genomic Underpinnings of Quantitative MRI-based Radiomic Phenotypes of Invasive Breast Carcinoma," *Scientific Reports*, vol. 5, no. 1, pp. 1–10, Dec. 2015.

[93] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, Aug. 1995, pp. 278–282 vol.1.

[94] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, Jan. 1989.

[95] S. Geman, E. Bienenstock, and R. Doursat, "Neural Networks and the Bias/Variance Dilemma," *Neural Computation*, vol. 4, no. 1, pp. 1–58, Jan. 1992.

[96] B. Neal, S. Mittal, A. Baratin, V. Tantia, M. Scicluna, S. Lacoste-Julien, and I. Mitliagkas, "A Modern Take on the Bias-Variance Tradeoff in Neural Networks," *arXiv:1810.08591 [cs, stat]*, Dec. 2019.

[97] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[98] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[99] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[100] D. Kumar, A. Wong, and D. A. Clausi, "Lung Nodule Classification Using Deep Features in CT Images," in *2015 12th Conference on Computer and Robot Vision*, Jun. 2015, pp. 133–138.

[101] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," *Medical image analysis*, vol. 42, pp. 1–13, 2017.

[102] B. Sahiner, Heang-Ping Chan, N. Petrick, D. Wei, M. Helvie, D. Adler, and M. Goodsitt, "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images," *IEEE Transactions on Medical Imaging*, vol. 15, no. 5, pp. 598–610, Oct. 1996.

[103] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778.

[104] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, Jul. 2017, pp. 2261–2269.

[105] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[106] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[107] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 779–788.

[108] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[109] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[110] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. Lille, France: JMLR.org, Jul. 2015, pp. 448–456.

[111] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance Normalization: The Missing Ingredient for Fast Stylization," *arXiv:1607.08022 [cs]*, Nov. 2017.

[112] S. G. Orel, N. Kay, C. Reynolds, and D. C. Sullivan, "BI-RADS Categorization As a Predictor of Malignancy," *Radiology*, vol. 211, no. 3, pp. 845–850, Jun. 1999.

[113] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," *The Lancet Oncology*, vol. 3, no. 3, pp. 159–165, Mar. 2002.

[114] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with Deep Learning," *Scientific Reports*, vol. 8, no. 1, pp. 1–7, Mar. 2018.

[115] W. Lotter, G. Sorensen, and D. Cox, "A Multi-scale CNN and Curriculum Learning Strategy for Mammogram Classification," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, ser. Lecture Notes in Computer Science, M. J. Cardoso, T. Arbel, G. Carneiro, T. Syeda-Mahmood, J. M. R. Tavares, M. Moradi, A. Bradley, H. Greenspan, J. P. Papa, A. Madabhushi, J. C. Nascimento, J. S. Cardoso, V. Belagiannis, and Z. Lu, Eds. Cham: Springer International Publishing, 2017, pp. 169–177.

[116] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Apr. 2015.

[117] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 3431–3440.

[118] S. Nikolov, S. Blackwell, R. Mendes, J. De Fauw, C. Meyer, C. Hughes, H. Askham, B. Romera-Paredes, A. Karthikesalingam, C. Chu, D. Carnell, C. Boon, D. D'Souza, S. A. Moinuddin, K. Sullivan, D. R. Consortium, H. Montgomery, G. Rees, R. Sharma, M. Suleyman, T. Back, J. R. Ledsam, and O. Ronneberger, "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy," *arXiv:1809.04430 [physics, stat]*, Sep. 2018.

[119] T. C. Hollon, B. Pandian, A. R. Adapa, E. Urias, A. V. Save, S. S. S. Khalsa, D. G. Eichberg, R. S. D'Amico, Z. U. Farooq, S. Lewis, P. D. Petridis, T. Marie, A. H. Shah, H. J. L. Garton, C. O. Maher, J. A. Heth, E. L. McKean, S. E. Sullivan, S. L. Hervey-Jumper, P. G. Patil, B. G. Thompson, O. Sagher, G. M. McKhann, R. J. Komotar, M. E. Ivan, M. Snuderl, M. L. Otten, T. D. Johnson, M. B. Sisti, J. N. Bruce, K. M. Muraszko, J. Trautman, C. W. Freudiger, P. Canoll, H. Lee, S. Camelo-Piragua, and D. A. Orringer, "Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks," *Nature Medicine*, vol. 26, no. 1, pp. 52–58, Jan. 2020.

[120] P. Kickingereder, F. Isensee, I. Tursunova, J. Petersen, U. Neuberger, D. Bonekamp, G. Brugnara, M. Schell, T. Kessler, M. Foltyn, I. Harting, F. Sahm, M. Prager, M. Nowosielski, A. Wick, M. Nolden, A. Radbruch, J. Debus, H.-P. Schlemmer, S. Heiland, M. Platten, A. von Deimling, M. J. van den Bent, T. Gorlia, W. Wick, M. Bendszus, and K. H. Maier-Hein, "Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: A multicentre, retrospective study," *The Lancet Oncology*, vol. 20, no. 5, pp. 728–740, May 2019.

[121] J. D. Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C. O. Hughes, R. Raine, J. Hughes, D. A. Sim,

C. Egan, A. Tufail, H. Montgomery, D. Hassabis, G. Rees, T. Back, P. T. Khaw, M. Suleyman, J. Cornebise, P. A. Keane, and O. Ronneberger, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, pp. 1342–1350, Sep. 2018.

[122] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[123] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 379–387.

[124] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *Computer Vision – ECCV 2016*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37.

[125] C. H. Polman, S. C. Reingold, B. Banwell, M. Clanet, J. A. Cohen, M. Filippi, K. Fujihara, E. Havrdova, M. Hutchinson, L. Kappos, F. D. Lublin, X. Montalban, P. O'Connor, M. Sandberg-Wollheim, A. J. Thompson, E. Waubant, B. Weinshenker, and J. S. Wolinsky, "Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria," *Annals of Neurology*, vol. 69, no. 2, pp. 292–302, Feb. 2011.

[126] F. C. Detterbeck, D. J. Boffa, and L. T. Tanoue, "The New Lung Cancer Staging System," *Chest*, vol. 136, no. 1, pp. 260–271, Jul. 2009.

[127] P. R. Greipp, J. S. Miguel, B. G. Durie, J. J. Crowley, B. Barlogie, J. Bladé, M. Boccadoro, J. A. Child, H. Avet-Loiseau, R. A. Kyle, J. J. Lahuerta, H. Ludwig, G. Morgan, R. Powles, K. Shimizu, C. Shustik, P. Sonneveld, P. Tosi, I. Turesson, and J. Westin, "International Staging System for Multiple Myeloma," *Journal of Clinical Oncology*, vol. 23, no. 15, pp. 3412–3420, May 2005.

[128] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut Learning in Deep Neural Networks," *arXiv:2004.07780 [cs, q-bio]*, Apr. 2020.

[129] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLOS ONE*, vol. 10, no. 3, p. e0118432, Mar. 2015.

[130] N. H. Shah, A. Milstein, and P. Steven C. Bagley, "Making Machine Learning Models Clinically Useful," *JAMA*, vol. 322, no. 14, pp. 1351–1352, Oct. 2019.

[131] Y. Guo, Y.-Q. Cai, Z.-L. Cai, Y.-G. Gao, N.-Y. An, L. Ma, S. Mahankali, and J.-H. Gao, "Differentiation of clinically benign and malignant breast lesions using diffusion-weighted imaging," *Journal of magnetic resonance imaging: JMRI*, vol. 16, no. 2, pp. 172–178, Aug. 2002.

[132] S. Bickelhaupt, D. Paech, P. Kickingereder, F. Steudle, W. Lederer, H. Daniel, M. Götz, N. Gählert, D. Tichy, M. Wiesenfarth, F. B. Laun, K. H. Maier-Hein, H.-P. Schlemmer, and D. Bonekamp, "Prediction of malignancy by a radiomic signature from contrast agent-free diffusion MRI in suspicious breast lesions found on screening mammography," *Journal of magnetic resonance imaging: JMRI*, vol. 46, no. 2, pp. 604–616, Aug. 2017.

[133] Z. Marin, K. A. Batchelder, B. C. Toner, L. Guimond, E. G. Gerasimova-Chechkina, A. R. Harrow, A. Arneodo, and A. Khalil, "Mammographic evidence of microenvironment changes in tumorous breasts," *Medical Physics*, vol. 44, no. 4, pp. 1324–1336, 2017.

[134] H. Li, Y. Zhu, E. S. Burnside, E. Huang, K. Drukker, K. A. Hoadley, C. Fan, S. D. Conzen, M. Zuley, J. M. Net, E. Sutton, G. J. Whitman, E. Morris, C. M. Perou, Y. Ji, and M. L. Giger, "Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set," *NPJ breast cancer*, vol. 2, 2016.

[135] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *Journal of Medical Imaging (Bellingham, Wash.)*, vol. 3, no. 3, p. 034501, Jul. 2016.

[136] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. P. M. van Stiphout, P. Granton, C. M. L. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. W. L. Aerts, "Radiomics: Extracting more information from medical images using advanced feature analysis," *European Journal of Cancer (Oxford, England: 1990)*, vol. 48, no. 4, pp. 441–446, Mar. 2012.

[137] H. D. White, R. M. Norris, M. A. Brown, P. W. Brandt, R. M. Whitlock, and C. J. Wild, "Left ventricular end-systolic volume as the major determinant of survival after recovery from myocardial infarction," *Circulation*, vol. 76, no. 1, pp. 44–51, Jul. 1987.

[138] Authors/Task Force members, P. M. Elliott, A. Anastasakis, M. A. Borger, M. Borggrefe, F. Cecchi, P. Charron, A. A. Hagege, A. Lafont, G. Limongelli, H. Mahrholdt, W. J. McKenna, J. Mogensen, P. Nihoyannopoulos, S. Nistri, P. G. Pieper, B. Pieske, C. Rapezzi, F. H. Rutten, C. Tillmanns, and H. Watkins, "2014 ESC Guidelines on diagnosis and management of hypertrophic cardiomyopathy: The Task Force for the Diagnosis and Management of Hypertrophic Cardiomyopathy of

the European Society of Cardiology (ESC)," *European Heart Journal*, vol. 35, no. 39, pp. 2733–2779, Oct. 2014.

[139] C. A. Miller, P. Jordan, A. Borg, R. Argyle, D. Clark, K. Pearce, and M. Schmitt, "Quantification of left ventricular indices from SSFP cine imaging: Impact of real-world variability in analysis methodology and utility of geometric modeling," *Journal of magnetic resonance imaging: JMRI*, vol. 37, no. 5, pp. 1213–1222, May 2013.

[140] R. M. Norris, H. D. White, D. B. Cross, C. J. Wild, and R. M. Whitlock, "Prognosis after recovery from myocardial infarction: The relative importance of cardiac dilatation and coronary stenoses," *European Heart Journal*, vol. 13, no. 12, pp. 1611–1618, Dec. 1992.

[141] J. N. Cohn, R. Ferrari, and N. Sharpe, "Cardiac remodeling—concepts and clinical implications: A consensus paper from an international forum on cardiac remodeling," *Journal of the American College of Cardiology*, vol. 35, no. 3, pp. 569–582, Mar. 2000.

[142] T. Lindsey and J.-J. Lee, "Automated Cardiovascular Pathology Assessment Using Semantic Segmentation and Ensemble Learning," *Journal of Digital Imaging*, Jan. 2020.

[143] P. Medrano-Gracia, B. R. Cowan, B. Ambale-Venkatesh, D. A. Bluemke, J. Eng, J. P. Finn, C. G. Fonseca, J. A. C. Lima, A. Suinesiaputra, and A. A. Young, "Left ventricular shape variation in asymptomatic populations: The Multi-Ethnic Study of Atherosclerosis," *Journal of Cardiovascular Magnetic Resonance: Official Journal of the Society for Cardiovascular Magnetic Resonance*, vol. 16, p. 56, Jul. 2014.

[144] X. Zhang, B. Ambale-Venkatesh, D. A. Bluemke, B. R. Cowan, J. P. Finn, A. H. Kadish, D. C. Lee, J. A. C. Lima, W. G. Hundley, A. Suinesiaputra, A. A. Young, and P. Medrano-Gracia, "Information maximizing component analysis of left ventricular remodeling due to myocardial infarction," *Journal of Translational Medicine*, vol. 13, no. 1, p. 343, Nov. 2015.

[145] S. Bickelhaupt, F. Steudle, D. Paech, A. Mlynarska, T. A. Kuder, W. Lederer, H. Daniel, M. Freitag, S. Delorme, H.-P. Schlemmer, and F. B. Laun, "On a fractional order calculus model in diffusion weighted breast imaging to differentiate between malignant and benign breast lesions detected on X-ray screening mammography," *PloS One*, vol. 12, no. 4, p. e0176077, 2017.

[146] S. Bickelhaupt, J. Tesdorff, F. B. Laun, T. A. Kuder, W. Lederer, S. Teiner, K. Maier-Hein, H. Daniel, A. Stieber, S. Delorme, and H.-P. Schlemmer, "Independent value of image fusion in unenhanced breast MRI using diffusion-weighted and morphological T2-weighted images for lesion characterization in patients with

recently detected BI-RADS 4/5 x-ray mammography findings," *European Radiology*, vol. 27, no. 2, pp. 562–569, Feb. 2017.

[147] M. Giannelli and N. Toschi, "On the use of trace-weighted images in body diffusional kurtosis imaging," *Magnetic Resonance Imaging*, vol. 34, no. 4, pp. 502–507, May 2016.

[148] M. Nolden, S. Zelzer, A. Seitel, D. Wald, M. Müller, A. M. Franz, D. Maleike, M. Fangerau, M. Baumhauer, L. Maier-Hein, K. H. Maier-Hein, H. P. Meinzer, and I. Wolf, "The Medical Imaging Interaction Toolkit: Challenges and advances: 10 years of open-source development," *International Journal of Computer Assisted Radiology and Surgery*, vol. 8, no. 4, pp. 607–620, Jul. 2013.

[149] "Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. - PubMed - NCBI," https://www.ncbi.nlm.nih.gov/pubmed/26501536.

[150] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, ser. Lecture Notes in Computer Science, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham: Springer International Publishing, 2016, pp. 424–432.

[151] L. Wang, C.-Y. Lee, Z. Tu, and S. Lazebnik, "Training Deeper Convolutional Networks with Deep Supervision," *arXiv:1505.02496 [cs]*, May 2015.

[152] S. Ciatto, N. Houssami, D. Ambrogetti, S. Bianchi, R. Bonardi, B. Brancato, S. Catarzi, and G. G. Risso, "Accuracy and underestimation of malignancy of breast core needle biopsy: The Florence experience of over 4000 consecutive biopsies," *Breast Cancer Research and Treatment*, vol. 101, no. 3, pp. 291–297, Mar. 2007.

[153] I. J. Dahabreh, L. S. Wieland, G. P. Adam, C. Halladay, J. Lau, and T. A. Trikalinos, *Core Needle and Open Surgical Biopsy for Diagnosis of Breast Lesions: An Update to the 2009 Report*, ser. AHRQ Comparative Effectiveness Reviews. Rockville (MD): Agency for Healthcare Research and Quality (US), 2014.

[154] H. M. Verkooijen, P. H. Peeters, E. Buskens, V. C. Koot, I. H. Borel Rinkes, W. P. Mali, and T. J. van Vroonhoven, "Diagnostic accuracy of large-core needle biopsy for nonpalpable breast disease: A meta-analysis," *British Journal of Cancer*, vol. 82, no. 5, pp. 1017–1021, Mar. 2000.

[155] J. H. Youk, E.-K. Kim, M. J. Kim, J. Y. Lee, and K. K. Oh, "Missed breast cancers at US-guided core needle biopsy: How to reduce them," *Radiographics: A Review Publication of the Radiological Society of North America, Inc*, vol. 27, no. 1, pp. 79–94, 2007 Jan-Feb.

[156] M. Wang, X. He, Y. Chang, G. Sun, and L. Thabane, "A sensitivity and specificity comparison of fine needle aspiration cytology and core needle biopsy in evaluation of suspicious breast lesions: A systematic review and meta-analysis," *Breast (Edinburgh, Scotland)*, vol. 31, pp. 157–166, Feb. 2017.

[157] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, Sep. 1988.

[158] "Assessment of BI-RADS category 4 lesions detected with screening mammography and screening US: Utility of MR imaging. - PubMed - NCBI," https://www.ncbi.nlm.nih.gov/pubmed/25271857.

[159] S. Bickelhaupt, D. Paech, F. B. Laun, F. Steudle, T. A. Kuder, A. Mlynarska, M. Bach, W. Lederer, S. Teiner, S. Schneider, M. E. Ladd, H. Daniel, A. Stieber, A. Kopp-Schneider, S. Delorme, and H.-P. Schlemmer, "Maximum intensity breast diffusion MRI for BI-RADS 4 lesions detected on X-ray mammography," *Clinical Radiology*, vol. 72, no. 10, pp. 900.e1–900.e8, Oct. 2017.

[160] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. de Marvao, T. Dawes, D. P. O'Regan, B. Kainz, B. Glocker, and D. Rueckert, "Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation," *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 384–395, Feb. 2018.

[161] D. Wu, G. Li, J. Zhang, S. Chang, J. Hu, and Y. Dai, "Characterization of Breast Tumors Using Diffusion Kurtosis Imaging (DKI)," *PLOS ONE*, vol. 9, no. 11, p. e113240, Nov. 2014.

[162] V. Golkov, A. Dosovitskiy, J. I. Sperl, M. I. Menzel, M. Czisch, P. Samann, T. Brox, and D. Cremers, "Q-Space Deep Learning: Twelve-Fold Shorter and Model-Free Diffusion MRI Scans," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1344–1351, May 2016.

[163] V. Golkov, T. Sprenger, J. I. Sperl, M. I. Menzel, M. Czisch, P. G. Sämann, and D. Cremers, "Model-free novelty-based diffusion MRI," *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016.

[164] S. Koppers and D. Merhof, "Direct Estimation of Fiber Orientations Using Deep Learning in Diffusion Imaging," in *Machine Learning in Medical Imaging*, ser. Lecture Notes in Computer Science, L. Wang, E. Adeli, Q. Wang, Y. Shi, and H.-I. Suk, Eds. Cham: Springer International Publishing, 2016, pp. 53–60.

[165] T. Schultz, "Learning a Reliable Estimate of the Number of Fiber Directions in Diffusion MRI," in *Medical Image Computing and Computer-Assisted Intervention*

– *MICCAI 2012*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, vol. 7512, pp. 493–500.

[166] N. Dhungel, G. Carneiro, and A. P. Bradley, "The Automated Learning of Deep Features for Breast Mass Classification from Mammograms," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham: Springer International Publishing, 2016, vol. 9901, pp. 106–114.

[167] Z. Liu, S. Wang, D. Dong, J. Wei, C. Fang, X. Zhou, K. Sun, L. Li, B. Li, M. Wang, and J. Tian, "The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges," *Theranostics*, vol. 9, no. 5, pp. 1303–1322, Feb. 2019.

[168] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding Transfer Learning for Medical Imaging," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 3347–3357.

[169] "What the radiologist should know about artificial intelligence – an ESR white paper," *Insights into Imaging*, vol. 10, Apr. 2019.

[170] L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, T. Arbel, H. Bogunovic, A. P. Bradley, A. Carass, C. Feldmann, A. F. Frangi, P. M. Full, B. van Ginneken, A. Hanbury, K. Honauer, M. Kozubek, B. A. Landman, K. März, O. Maier, K. Maier-Hein, B. H. Menze, H. Müller, P. F. Neher, W. Niessen, N. Rajpoot, G. C. Sharp, K. Sirinukunwattana, S. Speidel, C. Stock, D. Stoyanov, A. A. Taha, F. van der Sommen, C.-W. Wang, M.-A. Weber, G. Zheng, P. Jannin, and A. Kopp-Schneider, "Why rankings of biomedical image analysis competitions should be interpreted with care," *Nature Communications*, vol. 9, no. 1, pp. 1–13, Dec. 2018.

[171] S. Kohl, D. Bonekamp, H.-P. Schlemmer, K. Yaqubi, M. Hohenfellner, B. Hadaschik, J.-P. Radtke, and K. Maier-Hein, "Adversarial Networks for the Detection of Aggressive Prostate Cancer," *arXiv:1702.08014 [cs]*, Feb. 2017.

[172] P. Schelb, S. Kohl, J. P. Radtke, M. Wiesenfarth, P. Kickingereder, S. Bickelhaupt, T. A. Kuder, A. Stenzinger, M. Hohenfellner, H.-P. Schlemmer, K. H. Maier-Hein, and D. Bonekamp, "Classification of Cancer at Prostate MRI: Deep Learning versus Clinical PI-RADS Assessment," *Radiology*, vol. 293, no. 3, pp. 607–617, Oct. 2019.

[173] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, P. Bilic, P. F. Christ, R. K. G. Do, M. Gollub, J. Golia-Pernicka, S. H. Heckers, W. R. Jarnagin, M. K. McHugo, S. Napel, E. Vorontsov, L. Maier-Hein, and M. J. Cardoso, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv:1902.09063 [cs, eess]*, Feb. 2019.

[174] S. Wang, M. Zhou, Z. Liu, Z. Liu, D. Gu, Y. Zang, D. Dong, O. Gevaert, and J. Tian, "Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation," *Medical Image Analysis*, vol. 40, pp. 172–183, Aug. 2017.

[175] N. Wang, C. Bian, Y. Wang, M. Xu, C. Qin, X. Yang, T. Wang, A. Li, D. Shen, and D. Ni, "Densely Deep Supervised Networks with Threshold Loss for Cancer Detection in Automated Breast Ultrasound," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, ser. Lecture Notes in Computer Science, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 641–648.

[176] G. Pons, J. Martí, R. Martí, S. Ganau, and J. A. Noble, "Breast-lesion Segmentation Combining B-Mode and Elastography Ultrasound," *Ultrasonic Imaging*, vol. 38, no. 3, pp. 209–224, May 2016.

[177] M. H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, A. K. Davison, R. Marti, n. Moi Hoon Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, A. K. Davison, and R. Marti, "Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1218–1226, Jul. 2018.

[178] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser, S. Kadoury, T. Konopczynski, M. Le, C. Li, X. Li, J. Lipkovà, J. Lowengrub, H. Meine, J. H. Moltz, C. Pal, M. Piraud, X. Qi, J. Qi, M. Rempfler, K. Roth, A. Schenk, A. Sekuboyina, E. Vorontsov, P. Zhou, C. Hülsemeyer, M. Beetz, F. Ettlinger, F. Gruen, G. Kaissis, F. Lohöfer, R. Braren, J. Holch, F. Hofmann, W. Sommer, V. Heinemann, C. Jacobs, G. E. H. Mamani, B. van Ginneken, G. Chartrand, A. Tang, M. Drozdzal, A. Ben-Cohen, E. Klang, M. M. Amitai, E. Konen, H. Greenspan, J. Moreau, A. Hostettler, L. Soler, R. Vivanti, A. Szeskin, N. Lev-Cohain, J. Sosna, L. Joskowicz, and B. H. Menze, "The Liver Tumor Segmentation Benchmark (LiTS)," *arXiv:1901.04056 [cs]*, Jan. 2019.

[179] K. Roth, J. Hesser, and T. Konopczyński, "Mask Mining for Improved Liver Lesion Segmentation," *arXiv:1908.05062 [cs, eess]*, Feb. 2020.

[180] G. Chlebus, A. Schenk, J. H. Moltz, B. van Ginneken, H. K. Hahn, and H. Meine, "Automatic liver tumor segmentation in CT with fully convolutional neural networks

and object-based postprocessing," *Scientific Reports*, vol. 8, no. 1, pp. 1–7, Oct. 2018.

[181] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv:1710.05006 [cs]*, Jan. 2018.

[182] A. A. Shvets, V. I. Iglovikov, A. Rakhlin, and A. A. Kalinin, "Angiodysplasia Detection and Localization Using Deep Convolutional Neural Networks," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2018, pp. 612–617.

[183] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam, "MaskLab: Instance Segmentation by Refining Object Detection With Semantic and Direction Features," in *CVPR*, Jun. 2018.

[184] A. Shrivastava and A. Gupta, "Contextual Priming and Feedback for Faster R-CNN," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, vol. 9905, pp. 330–348.

[185] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What Can Help Pedestrian Detection?" in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, Jul. 2017, pp. 6034–6043.

[186] J. Uhrig, E. Rehder, B. Fröhlich, U. Franke, and T. Brox, "Box2Pix: Single-Shot Instance Segmentation by Assigning Pixels to Object Boxes," in *IEEE Intelligent Vehicles Symposium (IV)*, 2018.

[187] Z. Zhang, S. Qiao, C. Xie, W. Shen, B. Wang, and A. L. Yuille, "Single-Shot Object Detection with Enriched Semantics," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, Jun. 2018, pp. 5813–5821.

[188] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, Jul. 2017, pp. 936–944.

[189] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, vol. 11211, pp. 833–851.

[190] C. Peng, T. Xiao, Z. Li, Y. Jiang, X. Zhang, K. Jia, G. Yu, and J. Sun, "MegDet: A Large Mini-Batch Object Detector," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, Jun. 2018, pp. 6181–6189.

[191] N. Dvornik, K. Shmelkov, J. Mairal, and C. Schmid, "BlitzNet: A Real-Time Deep Network for Scene Understanding," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, Oct. 2017, pp. 4174–4182.

[192] T. Araújo, G. Aresta, A. Galdran, P. Costa, A. M. Mendonça, and A. Campilho, "UOLO - Automatic Object Detection and Segmentation in Biomedical Images," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, ser. Lecture Notes in Computer Science, D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madabhushi, Eds. Cham: Springer International Publishing, 2018, pp. 165–173.

[193] M. P. Shah, S. N. Merchant, and S. P. Awate, "MS-Net: Mixed-Supervision Fully-Convolutional Networks for Full-Resolution Segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, ser. Lecture Notes in Computer Science, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 379–387.

[194] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99.

[195] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[196] A. Masood, B. Sheng, P. Li, X. Hou, X. Wei, J. Qin, and D. Feng, "Computer-Assisted Decision Support System in Pulmonary Cancer detection and stage classification on CT images," *Journal of Biomedical Informatics*, vol. 79, pp. 117–128, Mar. 2018.

[197] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, and S. Shetty, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Medicine*, vol. 25, no. 6, pp. 954–961, Jun. 2019.

[198] T. Weikert, T. Akinci D'Antonoli, J. Bremerich, B. Stieltjes, G. Sommer, and A. W. Sauter, "Evaluation of an AI-Powered Lung Nodule Algorithm for Detection and 3D Segmentation of Primary Lung Tumors," https://www.hindawi.com/journals/cmmi/2019/1545747/, p. e1545747, 2019.

[199] W. De Wever, S. Ceyssens, L. Mortelmans, S. Stroobants, G. Marchal, J. Bogaert, and J. A. Verschakelen, "Additional value of PET-CT in the staging of lung cancer: Comparison with CT alone, PET alone and visual correlation of PET and CT," *European Radiology*, vol. 17, no. 1, pp. 23–32, Jan. 2007.

[200] S. S. Shim, K. S. Lee, B.-T. Kim, M. J. Chung, E. J. Lee, J. Han, J. Y. Choi, O. J. Kwon, Y. M. Shim, and S. Kim, "Non-small cell lung cancer: Prospective comparison of integrated FDG PET/CT and CT alone for preoperative staging," *Radiology*, vol. 236, no. 3, pp. 1011–1019, Sep. 2005.

[201] A. Kandathil, F. U. Kay, Y. M. Butt, J. W. Wachsmann, and R. M. Subramaniam, "Role of FDG PET/CT in the Eighth Edition of TNM Staging of Non–Small Cell Lung Cancer," *RadioGraphics*, vol. 38, no. 7, pp. 2134–2149, Nov. 2018.

[202] P. Goldstraw, K. Chansky, J. Crowley, R. Rami-Porta, H. Asamura, W. E. E. Eberhardt, A. G. Nicholson, P. Groome, A. Mitchell, V. Bolejack, P. Goldstraw, R. Rami-Porta, H. Asamura, D. Ball, D. G. Beer, R. Beyruti, V. Bolejack, K. Chansky, J. Crowley, F. Detterbeck, W. E. Erich Eberhardt, J. Edwards, F. Galateau-Sallé, D. Giroux, F. Gleeson, P. Groome, J. Huang, C. Kennedy, J. Kim, Y. T. Kim, L. Kingsbury, H. Kondo, M. Krasnik, K. Kubota, A. Lerut, G. Lyons, M. Marino, E. M. Marom, J. van Meerbeeck, A. Mitchell, T. Nakano, A. G. Nicholson, A. Nowak, M. Peake, T. Rice, K. Rosenzweig, E. Ruffini, V. Rusch, N. Saijo, P. Van Schil, J.-P. Sculier, L. Shemanski, K. Stratton, K. Suzuki, Y. Tachimori, C. F. Thomas, W. Travis, M. S. Tsao, A. Turrisi, J. Vansteenkiste, H. Watanabe, Y.-L. Wu, P. Baas, J. Erasmus, S. Hasegawa, K. Inai, K. Kernstine, H. Kindler, L. Krug, K. Nackaerts, H. Pass, D. Rice, C. Falkson, P. L. Filosso, G. Giaccone, K. Kondo, M. Lucchi, M. Okumura, E. Blackstone, F. Abad Cavaco, E. Ansótegui Barrera, J. Abal Arca, I. Parente Lamelas, A. Arnau Obrer, R. Guijarro Jorge, D. Ball, G. K. Bascom, A. I. Blanco Orozco, M. A. González Castro, M. G. Blum, D. Chimondeguy, V. Cvijanovic, S. Defranchi, B. de Olaiz Navarro, I. Escobar Campuzano, I. Macía Vidueira, E. Fernández Araujo, F. Andreo García, K. M. Fong, G. Francisco Corral, S. Cerezo González, J. Freixinet Gilart, L. García Arangüena, S. García Barajas, P. Girard, T. Goksel, M. T. González Budiño, G. González Casaurrán, J. A. Gullón Blanco, J. Hernández Hernández, H. Hernández Rodríguez, J. Herrero Collantes, M. Iglesias Heras, J. M. Izquierdo Elena, E. Jakobsen, S. Kostas, P. León Atance, A. Núñez Ares, M. Liao, M. Losanovscky, G. Lyons, R. Magaroles, L. De Esteban Júlvez, M. Mariñán Gorospe, B. McCaughan, C. Kennedy, R. Melchor Íñiguez,

L. Miravet Sorribes, S. Naranjo Gozalo, C. Álvarez de Arriba, M. Núñez Delgado, J. Padilla Alarcón, J. C. Peñalver Cuesta, J. S. Park, H. Pass, M. J. Pavón Fernández, M. Rosenberg, E. Ruffini, V. Rusch, J. Sánchez de Cos Escuín, A. Saura Vinuesa, M. Serra Mitjans, T. E. Strand, D. Subotic, S. Swisher, R. Terra, C. Thomas, K. Tournoy, P. Van Schil, M. Velasquez, Y. L. Wu, and K. Yokoi, "The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer," *Journal of Thoracic Oncology*, vol. 11, no. 1, pp. 39–51, Jan. 2016.

[203] K. Yan, M. Bagheri, and R. M. Summers, "3D Context Enhanced Region-Based Convolutional Neural Network for End-to-End Lesion Detection," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, ser. Lecture Notes in Computer Science, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 511–519.

[204] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[205] Y. LeCun, "1.1 Deep Learning Hardware: Past, Present, and Future," in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, Feb. 2019, pp. 12–19.

[206] K. Nagpal, D. Foote, Y. Liu, P.-H. C. Chen, E. Wulczyn, F. Tan, N. Olson, J. L. Smith, A. Mohtashamian, J. H. Wren, G. S. Corrado, R. MacDonald, L. H. Peng, M. B. Amin, A. J. Evans, A. R. Sangoi, C. H. Mermel, J. D. Hipp, and M. C. Stumpe, "Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer," *npj Digital Medicine*, vol. 2, no. 1, pp. 1–10, Jun. 2019.

[207] A. B. Rosenkrantz, J. S. Babb, S. S. Taneja, and J. M. Ream, "Proposed Adjustments to PI-RADS Version 2 Decision Rules: Impact on Prostate Cancer Detection," *Radiology*, vol. 283, no. 1, pp. 119–129, Apr. 2017.

[208] B. G. Muller, J. H. Shih, S. Sankineni, J. Marko, S. Rais-Bahrami, A. K. George, J. J. M. C. H. de la Rosette, M. J. Merino, B. J. Wood, P. Pinto, P. L. Choyke, and B. Turkbey, "Prostate Cancer: Interobserver Agreement and Accuracy with the Revised Prostate Imaging Reporting and Data System at Multiparametric MR Imaging," *Radiology*, vol. 277, no. 3, pp. 741–750, Dec. 2015.

[209] A. B. Rosenkrantz, L. A. Ginocchio, D. Cornfeld, A. T. Froemming, R. T. Gupta, B. Turkbey, A. C. Westphalen, J. S. Babb, and D. J. Margolis, "Interobserver

Reproducibility of the PI-RADS Version 2 Lexicon: A Multicenter Study of Six Experienced Prostate Radiologists," *Radiology*, vol. 280, no. 3, pp. 793–804, Sep. 2016.

[210] P. Skaane, K. Engedal, and A. Skjennald, "Interobserver Variation in the Interpretation of Breast Imaging: Comparison of mammography, ultrasonography, and both combined in the interpretation of palpable noncalcified breast masses," *Acta Radiologica*, vol. 38, no. 4, pp. 497–502, Jul. 1997.

[211] J. N. Mandrekar, "Measures of Interrater Agreement," *Journal of Thoracic Oncology*, vol. 6, no. 1, pp. 6–7, Jan. 2011.

[212] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. San Francisco, California, USA: AAAI Press, Feb. 2017, pp. 1919–1925.

[213] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander, and N. Silberman, "Learning From Noisy Labels by Regularized Estimation of Annotator Confusion," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 11 236–11 245.

[214] Z. Zhang and M. Sabuncu, "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 8778–8788.

[215] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[216] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond PASCAL: A benchmark for 3D object detection in the wild," in *IEEE Winter Conference on Applications of Computer Vision*. Steamboat Springs, CO, USA: IEEE, Mar. 2014, pp. 75–82.

[217] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. M. A. Eslami, D. Jimenez Rezende, and O. Ronneberger, "A Probabilistic U-Net for Segmentation of Ambiguous Images," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 6965–6975.

[218] T. M. Liddell and J. K. Kruschke, "Analyzing ordinal data with metric models: What could possibly go wrong?" *Journal of Experimental Social Psychology*, vol. 79, pp. 328–348, 2018.

[219] M. Feindt, "A Neural Bayesian Estimator for Conditional Probability Densities," *arXiv:physics/0402093*, Feb. 2004.

[220] P. McCullagh, "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 42, no. 2, pp. 109–142, 1980.

[221] M. Ghodrati, A. Farzmahdi, K. Rajaei, R. Ebrahimpour, and S.-M. Khaligh-Razavi, "Feedforward object-vision models only tolerate small image variations compared to human," *Frontiers in Computational Neuroscience*, vol. 8, 2014.

[222] F. C. Schmeel, "Variability in quantitative diffusion-weighted MR imaging (DWI) across different scanners and imaging sites: Is there a potential consensus that can help reducing the limits of expected bias?" *European Radiology*, vol. 29, no. 5, pp. 2243–2245, May 2019.

[223] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017 - 20th International Conference, Proceedings*. Springer Verlag, 2017, pp. 417–425.

[224] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, Jul. 2017, pp. 5967–5976.

[225] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic Backpropagation and Approximate Inference in Deep Generative Models," p. 9.

[226] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *International Conference for Learning Representations*, May 2014.

[227] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "HeMIS: Hetero-Modal Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, ser. Lecture Notes in Computer Science, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham: Springer International Publishing, 2016, pp. 469–477.

[228] M. C. Roethke, T. A. Kuder, T. H. Kuru, M. Fenchel, B. A. Hadaschik, F. B. Laun, H.-P. Schlemmer, and B. Stieltjes, "Evaluation of Diffusion Kurtosis Imaging Versus Standard Diffusion Imaging for Detection and Grading of Peripheral Zone Prostate Cancer," *Investigative Radiology*, vol. 50, no. 8, pp. 483–489, Aug. 2015.

[229] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

[230] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential Model-Based Optimization for General Algorithm Configuration," in *Learning and Intelligent Optimization*, C. A. C. Coello, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, vol. 6683, pp. 507–523.

[231] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning Augmentation Strategies From Data," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 113–123.

[232] T. Elsken, J. H. Metzen, and F. Hutter, "Neural Architecture Search: A Survey," *Journal of Machine Learning Research*, vol. 20, p. 21, 2019.

[233] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han, G. Yao, Y. Gao, Y. Zhang, Y. Wang, F. Hou, J. Yang, G. Xiong, J. Tian, C. Zhong, J. Ma, J. Rickman, J. Dean, B. Stai, R. Tejpaul, M. Oestreich, P. Blake, H. Kaluzniak, S. Raza, J. Rosenberg, K. Moore, E. Walczak, Z. Rengel, Z. Edgerton, R. Vasdev, M. Peterson, S. McSweeney, S. Peterson, A. Kalapara, N. Sathianathen, C. Weight, and N. Papanikolopoulos, "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 Challenge," *arXiv:1912.01054 [cs, eess]*, Dec. 2019.

[234] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*. Stanford, CA, USA: IEEE, Oct. 2016, pp. 565–571.

[235] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul. 2017, pp. 1175–1183.

[236] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning Where to Look for the Pancreas," in *1st Conference on Medical Imaging with Deep Learning*, vol. 1, 2018, p. 10.

[237] R. McKinley, R. Meier, and R. Wiest, "Ensembles of Densely-Connected CNNs with Label-Uncertainty for Brain Tumor Segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, ser. Lecture Notes in Computer Science, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham: Springer International Publishing, 2019, pp. 456–465.

[238] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz,

R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[239] Z. Wu, C. Shen, and A. van den Hengel, "Bridging Category-level and Instance-level Semantic Image Segmentation," *arXiv:1605.06885 [cs]*, May 2016.

[240] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[241] M. Wiesenfarth, A. Reinke, B. A. Landman, M. J. Cardoso, L. Maier-Hein, and A. Kopp-Schneider, "Methods and open-source toolkit for analyzing and visualizing challenge results," *arXiv:1910.05121 [cs, stat]*, Dec. 2019.

[242] L. Heinrich, J. Funke, C. Pape, J. Nunez-Iglesias, and S. Saalfeld, "Synaptic Cleft Segmentation in Non-isotropic Volume Electron Microscopy of the Complete Drosophila Brain," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, ser. Lecture Notes in Computer Science, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 317–325.

[243] "Federated Learning powered by NVIDIA Clara," https://devblogs.nvidia.com/federated-learning-clara/, Dec. 2019.

[244] "Joint Imaging Platform," https://jip.dktk.dkfz.de/jiphomepage/.

[245] N. Rieke, J. Hancox, W. Li, F. Milletari, H. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. M. Summers, A. Trask, D. Xu, M. Baust, and M. J. Cardoso, "The Future of Digital Health with Federated Learning," *arXiv:2003.08119 [cs]*, Mar. 2020.

[246] J. Petersen, P. F. Jäger, F. Isensee, S. A. A. Kohl, U. Neuberger, W. Wick, J. Debus, S. Heiland, M. Bendszus, P. Kickingereder, and K. H. Maier-Hein, "Deep Probabilistic Modeling of Glioma Growth," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, ser. Lecture Notes in Computer Science, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds. Cham: Springer International Publishing, 2019, pp. 806–814.

[247] S. A. A. Kohl, B. Romera-Paredes, K. H. Maier-Hein, D. J. Rezende, S. M. A. Eslami, P. Kohli, A. Zisserman, and O. Ronneberger, "A Hierarchical Probabilistic U-Net for Modeling Multi-Scale Ambiguities," *arXiv:1905.13077 [cs]*, May 2019.

[248] J. Haugeland, *Artificial Intelligence: The Very Idea*. Cambridge, Mass: MIT Press, 1985.

[249] G. Marcus and E. Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust*, first edition ed. New York: Pantheon Books, 2019.

[250] M. Garnelo and M. Shanahan, "Reconciling deep learning with symbolic artificial intelligence: Representing objects and relations," *Current Opinion in Behavioral Sciences*, vol. 29, pp. 17–23, Oct. 2019.

[251] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic Routing Between Capsules," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3856–3866.

[252] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with EM routing," in *International Conference on Learning Representations*, Feb. 2018.

[253] A. Kosiorek, S. Sabour, Y. W. Teh, and G. E. Hinton, "Stacked Capsule Autoencoders," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 15 512–15 522.

[254] Y. Bengio, "From System 1 Deep Learning to System 2 Deep Learning," Neurips, 2019.

[255] Y. Bengio, D.-H. Lee, J. Bornschein, T. Mesnard, and Z. Lin, "Towards Biologically Plausible Deep Learning," *arXiv:1502.04156 [cs]*, Aug. 2016.

[256] S. P. Johnson, *Object Perception*. Oxford University Press, Mar. 2013, pp. 337–379.

[257] E. J. Green and J. Quilty-Dunn, "What is an object file?" *The British Journal for the Philosophy of Science*, Dec. 2017.

[258] F. Crick, "The recent excitement about neural networks," *Nature*, vol. 337, no. 6203, pp. 129–132, Jan. 1989.

[259] A. H. Marblestone, G. Wayne, and K. P. Kording, "Toward an Integration of Deep Learning and Neuroscience," *Frontiers in Computational Neuroscience*, vol. 10, Sep. 2016.

[260] N. Caporale and Y. Dan, "Spike Timing–Dependent Plasticity: A Hebbian Learning Rule," *Annual Review of Neuroscience*, vol. 31, no. 1, pp. 25–46, 2008.

[261] S. Löwe, P. O Connor, and B. Veeling, "Putting An End to End-to-End: Gradient-Isolated Learning of Representations," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 3039–3051.

[262] S. Reardon, "Rise of Robot Radiologists," *Nature*, vol. 576, no. 7787, pp. S54–S58, Dec. 2019.

[263] M. P. Recht, M. Dewey, K. Dreyer, C. Langlotz, W. Niessen, B. Prainsack, and J. J. Smith, "Integrating artificial intelligence into the clinical practice of radiology: Challenges and recommendations," *European Radiology*, Feb. 2020.

[264] "Geoff Hinton: On Radiology," *https://www.youtube.com/watch?v=2HMPRXstSvQ*, 2016.

[265] R. Susskind and D. Susskind, "Technology Will Replace Many Doctors, Lawyers, and Other Professionals," *Harvard Business Review*, Oct. 2016.

[266] Z. Obermeyer and E. J. Emanuel, "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine," *The New England Journal of Medicine*, vol. 375, no. 13, pp. 1216–1219, Sep. 2016.

[267] G. Hinton, "Deep Learning—A Technology With the Potential to Transform Health Care," *JAMA*, vol. 320, no. 11, pp. 1101–1102, Sep. 2018.

[268] C. P. Langlotz, "Will Artificial Intelligence Replace Radiologists?" *Radiology: Artificial Intelligence*, vol. 1, no. 3, p. e190058, May 2019.

[269] J. J. Budovec, C. A. Lam, and C. E. Kahn, "Informatics in Radiology: Radiology Gamuts Ontology: Differential Diagnosis for the Semantic Web," *RadioGraphics*, vol. 34, no. 1, pp. 254–264, Nov. 2013.

[270] D. Schönberger, "Artificial intelligence in healthcare: A critical analysis of the legal and ethical implications," *International Journal of Law and Information Technology*, vol. 27, no. 2, pp. 171–203, Jun. 2019.

[271] "How Should AI Be Developed, Validated, and Implemented in Patient Care?" *AMA Journal of Ethics*, vol. 21, no. 2, pp. E125–130, Feb. 2019.

[272] European Group on Ethics in Science and New Technologies to the European Commission, European Commission, European Commission, and Directorate-General for Research and Innovation, *Statement on Artificial Intelligence, Robotics and 'autonomous' Systems: Brussels, 9 March 2018*, 2018.

[273] J. Guo and B. Li, "The Application of Medical Artificial Intelligence Technology in Rural Areas of Developing Countries," *Health Equity*, vol. 2, no. 1, pp. 174–181, Aug. 2018.

[274] J. Das, L. Woskie, R. Rajbhandari, K. Abbasi, and A. Jha, "Rethinking assumptions about delivery of healthcare: Implications for universal health coverage," *BMJ*, p. k1716, May 2018.

[275] H. Harvey, "Why AI will not replace radiologists," https://towardsdatascience.com/why-ai-will-not-replace-radiologists-c7736f2c7d80, Apr. 2018.

[276] E. J. Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*, first edition ed. New York: Basic Books, 2019.

[277] Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, and The National Academies of Sciences, Engineering, and Medicine, *Improving Diagnosis in Health Care*, E. P. Balogh, B. T. Miller, and J. R. Ball, Eds. Washington, D.C.: National Academies Press, Dec. 2015.

[278] M. A. Makary and M. Daniel, "Medical error—the third leading cause of death in the US," *BMJ*, vol. 353, May 2016.

[279] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, Apr. 2019.

[280] M. A. Bruno, E. A. Walker, and H. H. Abujudeh, "Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction," *RadioGraphics*, vol. 35, no. 6, pp. 1668–1676, Oct. 2015.

[281] E. J. Topol, *The Patient Will See You Now: The Future of Medicine Is in Your Hands*. New York: Basic Books, 2016.

[282] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, Jun. 2006, pp. 233–240.

[283] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham: Springer International Publishing, 2016, vol. 9901, pp. 424–432.

[284] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," in *Computer Vision – ECCV 2016*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 630–645.

[285] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7132–7141.

[286] Y. Wu and K. He, "Group Normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[287] S. Singh and S. Krishnan, "Filter Response Normalization Layer: Eliminating Batch Dependence in the Training of Deep Neural Networks," *arXiv:1911.09737 [cs, stat]*, Nov. 2019.

[288] A. L. Maas, A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[289] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 2818–2826.

[290] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The Importance of Skip Connections in Biomedical Image Segmentation," in *Deep Learning and Data Labeling for Medical Applications*. Springer, Sep. 2016, pp. 179–187.

[291] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8026–8037.

[292] F. Isensee, P. Jäger, J. Wasserthal, D. Zimmerer, J. Petersen, S. Kohl, J. Schock, A. Klein, T. Roß, S. Wirkert, P. Neher, S. Dinkelacker, G. Köhler, and K. Maier-Hein, "Batchgenerators - a python framework for data augmentation," Zenodo, Jan. 2020.

[293] L. Bennett, X. Zhoubing, J. E. Igelsias, M. Styner, T. R. Langerak, and A. Klein, *MICCAI Multi-Atlas Labeling Beyond the Cranial Vault–Workshop and Challenge*, 2015.

[294] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, R. Strand, F. Malmberg, Y. Ou, C. Davatzikos, M. Kirschner, F. Jung, J. Yuan, W. Qiu, Q. Gao, P. E. Edwards, B. Maan, F. van der Heijden, S. Ghose, J. Mitra, J. Dowling, D. Barratt, H. Huisman, and A. Madabhushi, "Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge," *Medical Image Analysis*, vol. 18, no. 2, pp. 359–373, Feb. 2014.

[295] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre, M. Jorge Cardoso, N. Cawley, O. Ciccarelli, C. A. M. Wheeler-Kingshott, S. Ourselin, L. Catanese, H. Deshpande,

P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S. K. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L. O. Iheme, D. Unay, S. Jain, D. M. Sima, D. Smeets, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, P.-L. Bazin, P. A. Calabresi, C. M. Crainiceanu, L. M. Ellingsen, D. S. Reich, J. L. Prince, and D. L. Pham, "Longitudinal multiple sclerosis lesion segmentation: Resource and challenge," *NeuroImage*, vol. 148, pp. 77–102, Mar. 2017.

[296] A. E. Kavur, N. S. Gezer, M. Barış, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, B. Baydar, D. Lachinov, S. Han, J. Pauli, F. Isensee, M. Perkonigg, R. Sathish, R. Rajan, S. Aslan, D. Sheet, G. Dovletov, O. Speck, A. Nürnberger, K. H. Maier-Hein, G. B. Akar, G. Ünal, O. Dicle, and M. A. Selver, "CHAOS Challenge – Combined (CT-MR) Healthy Abdominal Organ Segmentation," *arXiv:2001.06535 [cs, eess]*, Jan. 2020.

[297] R. Trullo, C. Petitjean, B. Dubray, and S. Ruan, "Multiorgan segmentation using distance-aware adversarial networks," *Journal of medical imaging*, 2019.

List of Figures

List of Tables

Acronyms

ACDC Automated Cardiag Diagnosis Challenge. 40, 45

ADC Apparent Diffusion Coefficient. 11, 28, 31, 32, 36, 37, 50, 52, 56, 57, 98, 99, 191

AKC Apparent Kurtosis Coefficient. 11, 28, 31, 32, 36–38, 50, 52, 56, 57, 98, 99, 191

AP Average Precision. 129

ARV Abnormal Right Ventricle. 43

AUC Area Under the Receiver Operating Characetristic Curve. 34, 36–38, 54, 55, 57, 103, 129, 130

AUPRC Area under the Precision Recall Curve. 129

AVP Average Viewpoint Precision. 92

BI-RADS Breast-Imaging Reporting and Data System. 14, 15, 28, 38, 39

CAD Computer Assisted Diagnosis. 2, 19, 41

CNN Convolutional Neural Network. 5, 6, 21, 22, 26, 40, 49–53, 55–57, 61, 87, 97, 98, 100, 118, 121

CT Computed Tomography. 6, 11, 62, 119

DCIS Ductal Carcinoma in Situ. 13, 15, 16, 39

DCM Dilated Cardiomyopathy. 43, 46

DDC Deep Diffusion Coefficient. 51–53, 55–57

DKI Diffusion Kurtosis Imaging. 11, 29, 31, 32, 34, 37, 50, 51, 56, 57, 97–99, 101, 103

DWI Diffusion Weighted Magentic Resonance Imaging. 5, 6, 10, 11, 16, 17, 27–30, 32, 37, 39, 40, 49, 50, 52, 54, 57, 60, 62, 85, 98, 101, 118, 119, 191

ED End Dyastolic. 42–47

ES End Systolic. 42–47

FPN Feature Pyramid Network. 25, 68–70, 85, 91

FROC Free-Response Receiver Operating Characteristic. 130

HCM Hypertrophic Cardiomyopathy. 43

IDC Invasive Ductal Carcinoma. 14

ILC Invasive Lobular Carcinoma. 14

IoU Intersection over Union. 25

LCIS Lobular Carcinoma in Situ. 13

LVC Left Ventricular Cavity. 40, 42, 44–46

LVM Left Ventricular Myocardium. 40, 42, 44–46

mAP Mean Average Precision. 74, 75, 81, 91, 92, 129, 130

MBDA Model-based Domain Adaptation. 98–101, 103

MICCAI Medical Image Computing and Computer Assisted Intervention. 3, 49, 88, 104, 105

MINF Myocardial Infarction. 43, 46

MLP Multilayer Perceptron. 20, 21, 45, 46

MRI Magnetic Resonance Imaging. 34, 40, 118

NMS Non-maximum Surpression. 25

pAUC partial AUC. 34, 36, 129

PET Positron Emission Tomography. 12

PET-CT Positron Emission Tomography - Computed Tomography. 6, 12, 62, 119

PRC Precision Recall Curve. 129

ROC Receiver Operating Characetristic. 33, 36, 57, 129

ROI Region of Interest. 5, 28, 32, 36, 37, 51, 53–55, 57, 64, 93, 118

ROIs Regions of Interest. 2, 5, 6, 28, 38, 50, 52, 61, 77, 81, 118

RPN Region Proposal Network. 25, 91

RVC Right Ventricular Cavity. 40, 42, 44–46

SotA State of the Art. 19, 22, 24, 25

WBC Weighted Box Clustering. 94