Deep Learning for
Clinical Decision Support in Oncology

Dissertation
Zur Erlangung des akademischen Grades
Doktoringenieur (Dr.-Ing.)

vorgelegt der Fakultät für Informatik und Automatisierung
der Technischen Universität Ilmenau

von Alexander Katzmann, M.Sc.

1. Gutachter:	Univ.-Prof. Dr.-Ing. Horst-Michael Groß
2. Gutachter:	Univ.-Prof. Dr.-Ing. habil. Bernhard Preim
3. Gutachter:	Prof. Dr.-Ing. Horst Hahn

Tag der Einreichung:	4. Oktober 2021
Tag der wissenschaftlichen Aussprache:	24. Februar 2022

DOI: 10.22032/dbt.51864
URN: urn:nbn:de:gbv:ilm1-2022000100

Abstract

Over the last decades, medical imaging methods, such as computed tomography (CT), have become an indispensable tool of modern medicine, allowing for a fast, non-invasive inspection of organs and tissue. Thus, the amount of acquired healthcare data has rapidly grown, increased 15-fold within the last years, and accounts for more than 30 % of the world's generated data volume. In contrast, the number of trained radiologists remains largely stable. Thus, medical image analysis, settled between medicine and engineering, has become a rapidly growing research field. Its successful application may result in remarkable time savings and lead to a significantly improved diagnostic performance. Many of the work within medical image analysis focuses on radiomics, i. e. the extraction and analysis of hand-crafted imaging features. Radiomics, however, has been shown to be highly sensitive to external factors, such as the acquisition protocol, having major implications for reproducibility and clinical applicability.

Lately, deep learning has become one of the most employed methods for solving computational problems. With successful applications in diverse fields, such as robotics, physics, mathematics, and economy, deep learning has revolutionized the process of machine learning research. Having large amounts of training data is a key criterion for its successful application. These data, however, are rare within medicine, as medical imaging is subject to a variety of data security and data privacy regulations. Moreover, medical imaging data often suffer from heterogeneous quality, label imbalance, and label noise, rendering a considerable fraction of deep learning-based algorithms inapplicable.

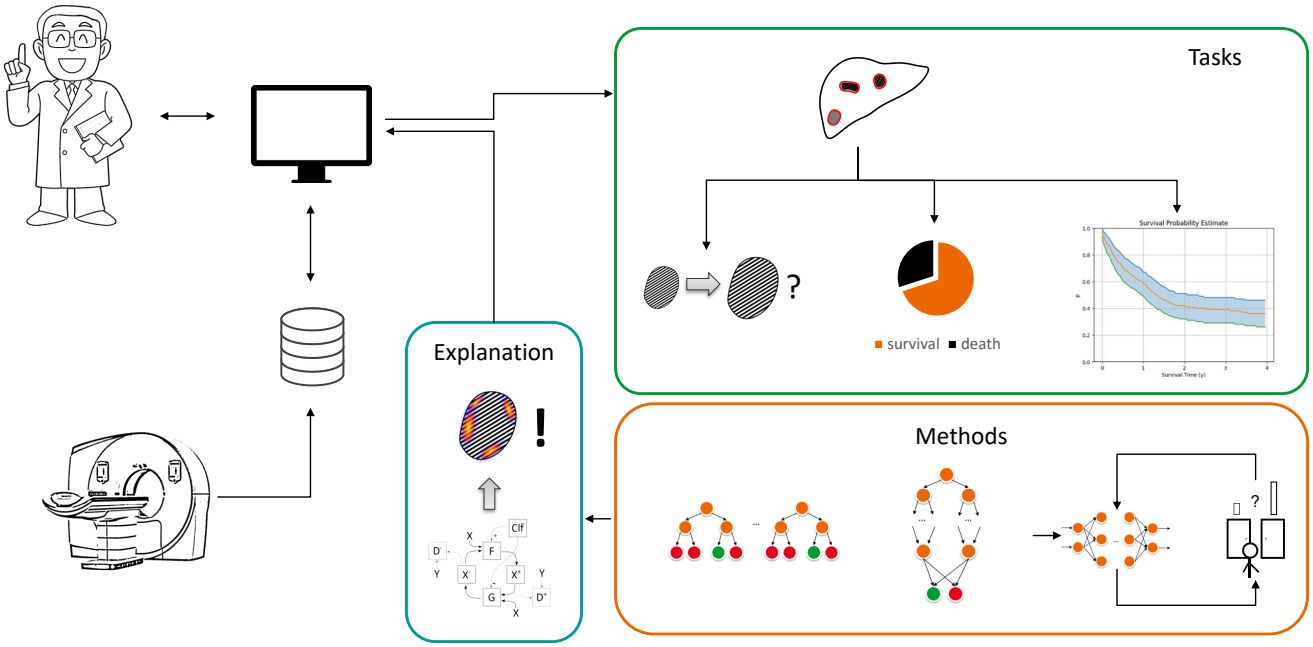
Settled in the field of CT oncology, this work addresses these issues, showing up ways to successfully handle medical imaging data using deep learning. It proposes novel methods for clinically relevant tasks, such as lesion growth and patient survival prediction, confidence estimation, meta-learning and classifier ensembling, and finally deep decision explanation, yielding superior performance in comparison to state-of-the-art approaches, and being applicable to a wide variety of applications. With this, the work contributes towards a clinical translation of deep learning-based algorithms, aiming for an improved diagnosis, and ultimately overall improved patient healthcare.

Zusammenfassung

In den letzten Jahrzehnten sind medizinische Bildgebungsverfahren wie die Computertomographie (CT) zu einem unersetzbaren Werkzeug moderner Medizin geworden, welche eine zeitnahe, nicht-invasive Begutachtung von Organen und Geweben ermöglichen. Die Menge an anfallenden Daten ist dabei rapide gestiegen, allein innerhalb der letzten Jahre um den Faktor 15, und aktuell verantwortlich für 30 % des weltweiten Datenvolumens. Die Anzahl ausgebildeter Radiologen ist weitestgehend stabil, wodurch die medizinische Bildanalyse, angesiedelt zwischen Medizin und Ingenieurwissenschaften, zu einem schnell wachsenden Feld geworden ist. Eine erfolgreiche Anwendung verspricht Zeitersparnisse, und kann zu einer höheren diagnostischen Qualität beitragen. Viele Arbeiten fokussieren sich auf "Radiomics", die Extraktion und Analyse von manuell konstruierten Features. Diese sind jedoch anfällig gegenüber externen Faktoren wie dem Bildgebungsprotokoll, woraus Implikationen für Reproduzierbarkeit und klinische Anwendbarkeit resultieren.

In jüngster Zeit sind Methoden des “Deep Learning” zu einer häufig verwendeten Lösung algorithmischer Problemstellungen geworden. Durch Anwendungen in Bereichen wie Robotik, Physik, Mathematik und Wirtschaft, wurde die Forschung im Bereich maschinellen Lernens wesentlich verändert. Ein Kriterium für den Erfolg stellt die Verfügbarkeit großer Datenmengen dar. Diese sind im medizinischen Bereich rar, da die Bilddaten strengen Anforderungen bezüglich Datenschutz und Datensicherheit unterliegen, und oft heterogene Qualität, sowie ungleichmäßige oder fehlerhafte Annotationen aufweisen, wodurch ein bedeutender Teil der Methoden keine Anwendung finden kann.

Angesiedelt im Bereich onkologischer Bildgebung zeigt diese Arbeit Wege zur erfolgreichen Nutzung von Deep Learning für medizinische Bilddaten auf. Mittels neuer Methoden für klinisch relevante Anwendungen wie die Schätzung von Läsionswachtum, Überleben, und Entscheidungkonfidenz, sowie Meta-Learning, Klassifikator-Ensembling, und Entscheidungsvisualisierung, werden Wege zur Verbesserungen gegenüber State-of-the-Art-Algorithmen aufgezeigt, welche ein breites Anwendungsfeld haben. Hierdurch leistet die Arbeit einen wesentlichen Beitrag in Richtung einer klinischen Anwendung von Deep Learning, zielt auf eine verbesserte Diagnose, und damit letztlich eine verbesserte Gesundheitsversorgung insgesamt.



Deep Learning for Clinical Decision Support in Oncology

Alexander Katzmann, M.Sc.

Supervisors: Prof. Dr.-Ing. Horst-Michael Groß[1]
Dr.-Ing. Michael Sühling[2]

[1]Neuroinformatics and Cognitive Robotics Lab
Technische Universität Ilmenau

[2]Computed Tomography Image Analytics Group
Siemens Healthineers

The styling of this book is based on The Legrand Orange Book, Version 2.4 (26/09/2018).

Original author:
Mathias Legrand (legrand.mathias@gmail.com)
with modifications by:
Vel (vel@latextemplates.com)

This work has received funding from the German Federal Ministry of Education and Research as part of the PANTHER project under grant agreement no. 13GW0163A.

First printing, April 2022

Acknowledgements

This work is dedicated to all who supported me on the long way until the final submission. Therefore, my very special thanks go to my supervisors Prof. Dr. Horst-Michael Groß, full professor at the Ilmenau University of Technology, and Dr. Michael Sühling, team lead at the CT R&D Image Analytics group at Siemens Healthineers for providing me with both academic and non-academic support in the creation of this work in various ways over all the years, thus finally making all that possible. I would like to further thank:

- My wife Antonia and my most beloved children Theodor and Emilia for supporting me in the creation of this work by encouraging me, relieving me, and having understanding for me, during the creation of this thesis often working until deep in the night,
- My mother Christine, my brother Robert and Lutz Ebhardt for their enormous help throughout all the years that we have spent together, and for giving me the strength I needed,
- My most loved friends Philipp, Stefanie, and Thomas for being the most supportive friends I could ever imagine,
- My parents in law for providing support to our family in any way and in times when it was urgently needed,
- My most honored colleagues for helping me with all the difficult brain teasers which arose over the years, and for letting me take part in their enormous knowledge about *every* possible subject,
- My university colleagues, especially Dipl.-Inf. Ronny Stricker for being my mentor and first address to approach over all my university years, Dr.-Ing. Markus Eisenbach for giving me valuable hints regarding deep neural networks, Thomas Schmiedel, M.Sc. and Tim Wengefeld, M.Sc. for various academic and non-academic discussions, and Dr.-Ing. Steffen Müller for being the universal genious he is,
- Mr. Matei Adrian Vidican, M.Sc. for helping me to take the decision to begin this work,
- Mr. Samuel Havadej, M.Sc., and Mr. Mayank Patwari, M.Sc., for discussing and challenging my research a variety of times, significantly helping to improve it,
- Prof. Dr. Thomas Flohr and Dr. André Hartung for the great opportunity to work for Siemens Healthineers and the support which I received there, and finally
- All those who I forgot to explicitly name for their tremendous support, and for understanding that I failed to mention their individual contribution here.

A Preface To Be Carefully Read

The following dissertation analyses the application of artificial intelligence and deep learning to medical image analysis and clinical decision support. Within the last years, artificial intelligence for clinical decision support has been a topic of steadily growing research interest. The applications range from lung cancer analysis, over cardiac and stroke assessment. This work is therefore dedicated to an assessment of the current state of knowledge on this area.

Background / Context

Artificial agents play an important role in many areas such as medical images and health (i.e. healthcare) [1], information retrieval or diagnosis [2]. However, it remains unclear which problems are best solved by such applications [3]. The reason behind the lack of good solutions lies primarily with computational complexity: it is difficult to find algorithms that can tackle many real-world problems successfully [4]–with the notable exception of medical image evaluation [5]. In the case of medical image processing, there exist only a handful of robust methods used within the field [6]. Most of these methods have proven effective in the past but still fall far short of what would be needed for efficient use in practical tasks [7].

It was found that the problem could be overcome through new approaches to statistical learning, based on deep neural networks. With these methods, the system learns to distinguish various classes (e. g. brain lesions or tumors) by building models of individual features from sparse and sparsely labeled data sets. These "feature architectures" can then give an approximate representation of the target feature space via deep feedforward circuits without training.

Because it does not require extensive preprocessing of images, these models do not require additional memory for intermediate values between the image and previous ones and thus do not require more computational power compared to existing techniques. Another benefit is that while classification systems require extensive memory for small datasets, deep networks also scale up well to large datasets. The author developed several systems under different optimization parameters using various learning algorithms. As the number of training examples grows, these models converge at an adequate level of performance. These results suggest that deep learning is a promising tool for biomedical imaging classification.

–

The above lines have been written by the Generative Pre-trained Transformer architecture GPT-2 [Radford et al. 2019], including paragraphing and references. GPT-2 is a natural language processing model for text generation and prediction. It was given the beginning of the first paragraph and continued with a well-structured text demonstrating a profound knowledge of the subject as well as scientific writing. Obviously, this is an impressive example of what current artificial intelligence[1] is already able to achieve, leading us to the highly important question of how this potential can be leveraged for improving human life. While within the scope of this thesis it will not be possible to fully fathom out every single aspect, this work aims for giving at least a part of the answer to this highly important and valuable question by focussing on the subfield of *medical image analysis*.

[1] The used model *only* had 1.5 billion parameters, with its successor GPT-3 already having 175 billion (117x) and the currently largest model Wu Dao 2.0 being at 1.75 trillion parameters (1,167x).

Contents

I Introduction

1 Introduction 3

1.1 Motivation 4

1.2 Outline 5

1.3 Innovative Contributions 7

1.4 Publications 7

II Scientific Background

2 Clinical Application 13

2.1 Cancer Statistics 15

2.1.1 Lung Cancer 15

2.1.2 Colorectal Cancer 16

2.2 Clinical Workflow 17

3 Radiomics 19

4 Deep Neural Networks 23

III Oncological Decision Support

5 Liver Lesion Growth Prediction 29
5.1 Introduction 29
5.2 Methods 31
5.2.1 Autoencoder Network Architecture 31
5.2.2 Predictor Network Architecture 32
5.3 Experiments 33
5.3.1 Dataset 33
5.3.2 Classifier Baseline 34
5.3.3 Deep Network Training 35
5.3.4 Results 36
5.4 Discussion 36
5.5 Conclusion 38

6 One-Year Survival Estimation 39
6.1 Introduction 39
6.2 Methods 40
6.2.1 Preprocessing 40
6.2.2 Baseline Classifier Design 41
6.2.3 Sparse Characterization 41
6.3 Experiments 42
6.3.1 Dataset 42
6.3.2 Results 43
6.4 Discussion 45
6.5 Conclusion 46

7 Deep Survival Regression 47
7.1 Medical Background 48
7.2 Related Work 49
7.3 Methods 49
7.3.1 Network Architecture 50
7.3.2 Loss Definition 50
7.4 Experiments 51
7.4.1 Datasets 52
7.4.2 Results 54
7.5 Discussion 54
7.6 Conclusion 55

IV Meta-Methods & Decision-Explanation

8 Deep Confidence Estimation 61
8.1 Introduction 62
8.1.1 Background 64
8.2 Base approach 64
8.3 Methods 66
8.3.1 Metamemory Importance Sampling 66
8.3.2 Model augmentation 67
8.4 Experiments 68
8.4.1 CIFAR-10 68
8.4.2 CIFAR-100 69
8.4.3 Radiological image data 69
8.5 Discussion 71
8.6 Conclusion 72

9 Bootstrapping Methods 73
9.1 Bootstrapping 73
9.2 Classifier Architectures 74
9.2.1 Bootstrapped Path Shaking 75
9.2.2 Deep Random Forests 76
9.2.3 Bootstrapping Ensembles 80
9.2.4 Experiments 81
9.2.5 Results 83
9.2.6 Discussion 85
9.3 Deep Survival Forests 87
9.3.1 Experiments 88
9.3.2 Results 89
9.3.3 Discussion 89
9.4 Conclusion 91

10 Deep Decision Explanation 93
10.1 Introduction 93
10.1.1 Theoretical Considerations 94
10.1.2 Related Work 96
10.2 Deep Decision Explanation using Cycle-GANs 98
10.2.1 Method 98
10.2.2 Relevancy Visualization 102
10.3 Experiments 102
10.3.1 Datasets 103
10.3.2 Quantiative Evaluation 103
10.3.3 User Study 104

10.3.4 Results 104
10.4 Discussion **107**
10.4.1 Limitations 107
10.5 Conclusion **108**

V Conclusion

11 Summary & Critical Evaluation 115
11.1 Outcome Prediction in Oncology **115**
11.2 Meta-Methods and Decision-Explanation **117**

12 Outlook 119

VI Appendix

Bibliography 123

A Additional Background 141

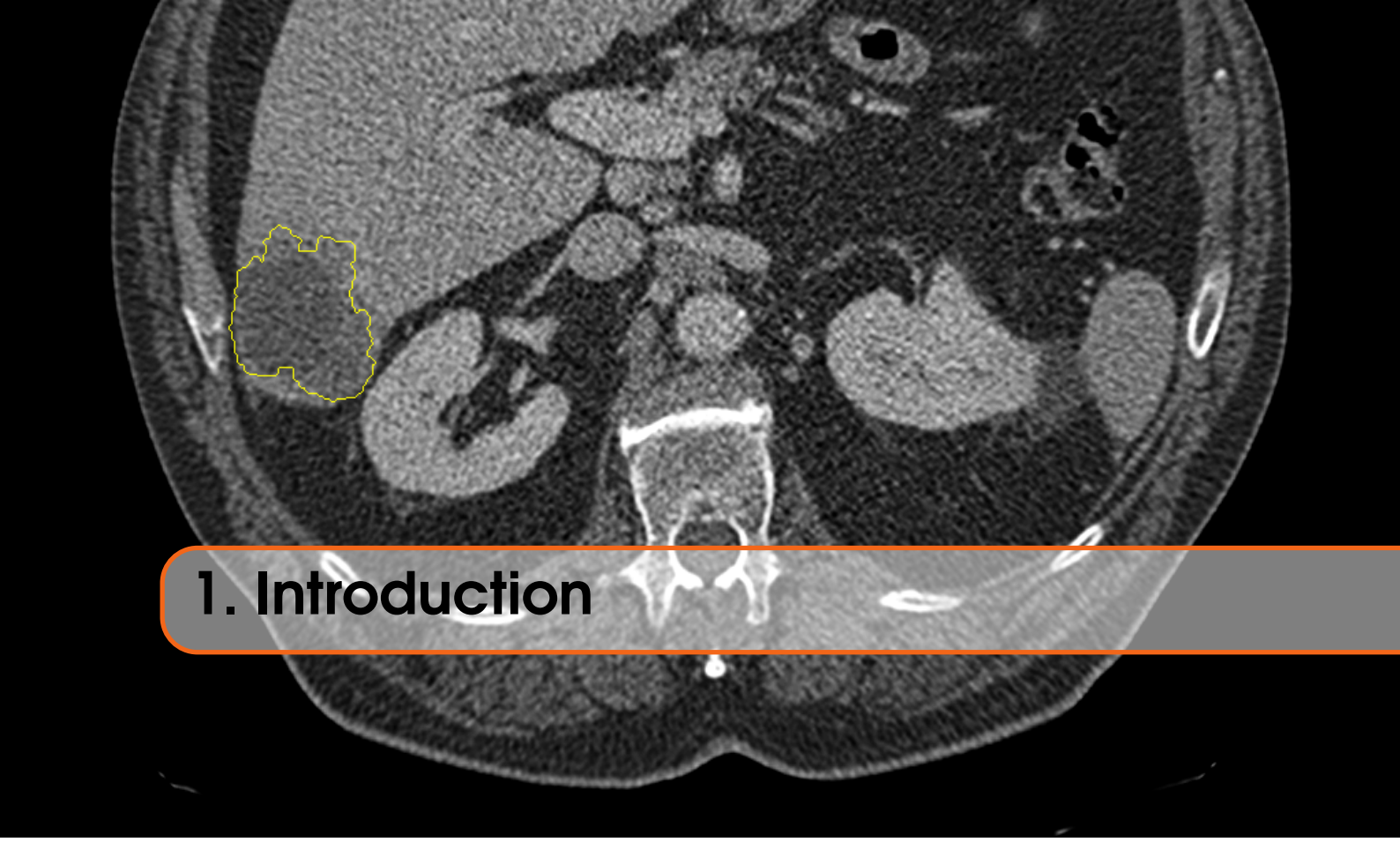
B Tables & Figures 143

C Metrics & Measures 159

Index 165

Introduction

1 Introduction 3
1.1 Motivation
1.2 Outline
1.3 Innovative Contributions
1.4 Publications



1. Introduction

Deep Learning has become one of the most employed methods for solving computational problems of any kind. At the latest with the high-profile successes in image recognition [Krizhevsky et al. 2012], Computer Chess and Computer Go [Silver et al. 2016, 2017, 2018], convolutional neural networks have become a topic of steadily growing interest and were applied to a variety of applications, including complex board [Anthony et al. 2020] and video games [Torrado et al. 2018], face recognition [Sun et al. 2015], person detection [Eisenbach et al. 2016] and re-identification [Ahmed et al. 2015], e-commerce [Shankar et al. 2017], fluid dynamic simulation [Wang et al. 2020], and even exoplanet [Shallue et al. 2018] & galaxy discovery [González et al. 2018] as well as planetary defense [NVIDIA Corporation 2016]. While some approaches are mostly academic, deep learning *is applied* to a number of very interesting daily, as well as highly specialized engineering environments. It is employed for automated solutions in fields previously considered *non-automatable*, thus steadily redefining this term [Jumper et al. 2021; Rao et al. 2018; Santos et al. 2021].

This work aims to shed some light on recent applications of machine learning in one of the most challenging, while interesting, fields of application. While known to be traditionally conservative[1], in recent years deep learning has evenly found its way into this domain. Thankfully, it is a field that can strongly benefit from it, while creating a significant and highly valuable contribution to human wellbeing. With a special focus on exemplary applications, this work aims to demonstrate some of the most outstanding opportunities of **Deep Learning in Medical Image Analysis**.

[1] Medicine requires high standards for the introduction of new methods. E. g. is any kind of human trial, including the use of automated decision systems for diagnosis, internationally subject to a variety of highly-strict regulations. The Helsinki declaration, the ethical codex of the World Medical Association (WMA), explicitly demands to not do any medical research with humans in case of non-predictable risks or unsure benefits. [World Medical Association 2013]

1.1 Motivation

Medical image analysis has become one of the most rapidly growing fields in medicine. Medical imaging allows the practitioner to inspect organs and tissue, i. e. the inner workings of the body, while simultaneously being only minimally invasive. Over the last century, medical imaging has become an indispensable tool in modern medicine. The amount of acquired healthcare data has grown from 153 exabytes in 2013 to 2,314 exabytes by 2020, i. e. 15-fold ([Statista 2021], cf. Chapter 2.2), currently being responsible for around 30 % of the world's generated data volume, and having an estimated annual growth of 36 % [RBC Capital Markets 2021]. Simultaneously, the number of trained radiologists who are needed for clinical interpretation has remained largely stable. As a result of this imbalance, medical image analysis has become more and more important for the automation of the clinical reading process, specifically taking care of simple and repetitive tasks, which in turn allows the radiologist to focus on the assessment of clinical manifestations which are less simple to automate.

As was shown by Cowan et al. [2013], the median reading times for abdominal computed tomography scans, although typically containing multiple hundreds of image slices, currently lie as low as only 14 minutes, including the creation of a clinical reading report. 25 % of all clinical pelvis/abdomen scans are read and reported in less than only 9 minutes[2]. As is demonstrated by these low turnaround times, an automated case preparation by the means of medical image analysis, e. g. by segmenting clinically relevant findings, highlighting suspicious image regions, or even triaging cases by urgency, can significantly help radiologists to save valuable time, resulting in lower error rates, higher throughput and ultimately improved patient healthcare.

With the rise of increasing computational capabilities, the role of medical image analysis in clinical assessment has become significantly larger and is now involved in many clinical imaging tasks. Still, most clinical decision support software is based on static algorithms, using the fixed parameterizations which proved to be valuable during their implementation and evaluation process.

In contrast, this work will emphasize the role and potential of data-driven clinical decision support using deep neural networks (**DNNs**). Using purely data-driven approaches, such as deep neural networks, allows for a variety of highly interesting advantages, taking into account the current technological developments, such as the continuous capability of using steadily growing data pools, scalability with respect to the available hardware, a known tolerance against missing and incomplete data in case of sufficiently large datasets, and finally: the possibility of learning problems *without any need for hand-crafted problem engineering* [Shen et al. 2017].

Unfortunately, medical imaging data is often sparse due to data privacy, data security, and regulatory issues. Thus, due to the enormous amount of model parameters in data-driven approaches, many of the above-mentioned advantages do only partly apply, or can even turn around, which may ultimately lead to an unknown result quality. This is especially true if classes are irregularly (**label imbalance**) or only weakly (**small data**) represented, or if the provided labels are sparse or contain errors (**label noise**). Although medical imaging datasets often suffer from these problems, the outstanding prospects of a clinical application clearly overweigh. This work will therefore specifically address these

[2] Assuming a slice thickness of 1 mm over a range of approximately 50 cm, this is equal to the assessment of around one full image per second, not including the time for reporting.

issues and will aim to show up ways for researchers to handle typical issues with medical imaging data. Moreover, it will present approaches for confidence estimation and decision explanation, being integral components for a translation into clinical practice. Although a concluding analysis of all possible approaches would be far beyond the scope of this work, this thesis will aim to highlight the outstanding importance of the field, as well as its potential for improving the wellbeing of millions of people all over the world, encouraging current and future researchers to engage in exploring and illuminating this highly important topic.

1.2 Outline

This work is structured in five parts (I–V), each of them tackling a specific segment of the above-mentioned points. While each part will illuminate the general topic from a different perspective, they are organized in such a way that it is possible to read each of them independently, and refer to each other where adequate:

- **Part I** will be a general introduction into the topic. It will give an overview, point out the innovative contributions of this thesis, and lists the relevant work that was published during its creation, and which was fundamental for the contents of this thesis.

- **Part II** dives into the concrete field of application. It specifically gives an introduction to relevant clinical terms in the field of CT oncology and the clinical workflow, as well as an outline of the field of classical radiomics based on multivariate statistics, and the application of deep neural networks to medical imaging.

- **Parts III–IV** cover the methodological contributions of this thesis and thus will be the main part of this work (see Fig. 1.1 for an outline of the discussed applications within this thesis).

  - More specifically, **Part III** introduces algorithms for lesion growth and one-year survival prediction, as well as a novel architecture for deep survival regression. All algorithms are analyzed and discussed thoroughly regarding performance, limitations, and applicability in a clinical environment.

  - **Part IV** will tackle the specific issues which arise from an application of deep learning on medical imaging data, such as the few data and unknown-certainty issues which were briefly discussed in the previous section. First, a method for confidence estimation is proposed, which can be used to improve training on small- to medium-sized datasets. Secondly, model ensembling techniques are discussed, resulting in an algorithm for significantly improved performance on small to medium-sized data sets by utilizing ideas from the well-known random forest classifier architecture. Third, a novel approach for decision explanation on medical imaging data is proposed and compared to various state-of-the-art algorithms, showing the superiority of the approach for medical imaging data in comparison to existing methods.

- Finally, **Part V** contains a critical evaluation of the achievements of this work. In particular, it will be discussed to which extent open issues in current research could successfully be addressed, where it was possible to make a significant contribution, and which issues could not yet be satisfactorily covered, and thus need further investigation. Finally, the ethical implications of an application of neural networks to clinical decision support are examined, aiming to conclude with a preliminary answer to the question of what can be expected in this highly interesting field of research within the near future, and how the work presented in this thesis may actively contribute to it.

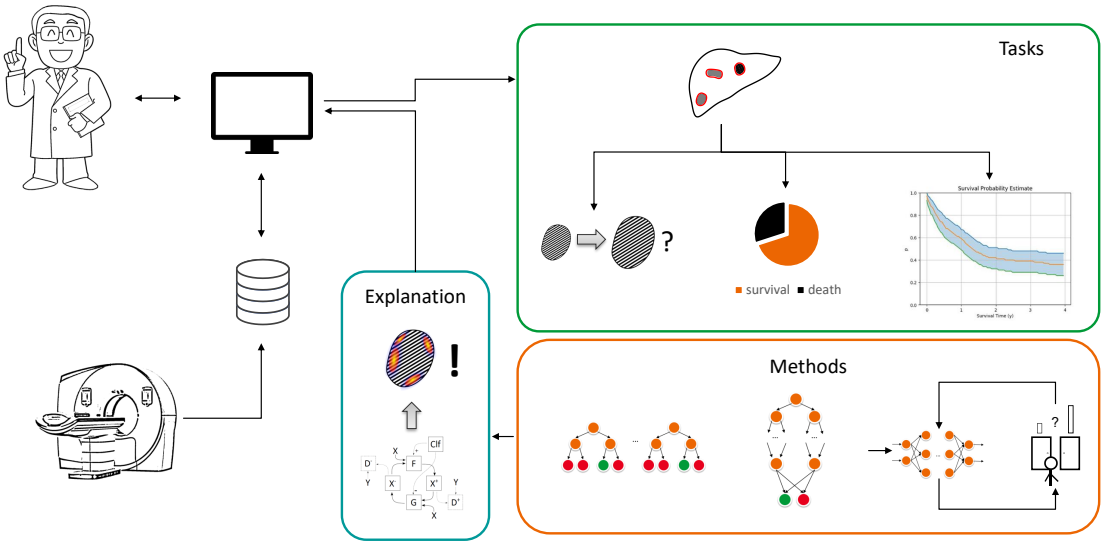


Figure 1.1: Overview on the topics of this thesis. First, a variety of deep learning-based methods for clinical decision support is presented (top-right), including methods for lesion growth and patient survival prediction. Subsequently, this work discusses meta-methods and advanced architectures (bottom-right), yielding improved classifier performance as well as relevant additional information in medical imaging scenarios. Finally, this work will present a method for classifier decision explanation (bottom-middle), being a key component towards a better comprehensibility of deep learning-based solutions in clinical decision support.

1.3 Innovative Contributions

The main scientific contributions of this work comprise:

- A successful application of **deep learning-based** methods for **lesion growth and one-year survival prediction** in metastatic colorectal cancer patients, being amongst the first realized solutions within this direction (see Chapters 5 and 6),
- **A method for non-proportional hazards-based deep survival regression** (Chapter 7) building on the results of the previous method,
- A successful application of **confidence estimation of deep image classifiers in medical imaging few-data scenarios**, and a demonstration of how it can be used for curriculum learning schedules for improved test time performance (Chapter 8),
- An innovative **combination of meta-classifier methods and deep neural networks** for efficient classification, regression and survival time estimation in few-data scenarios with enhanced accuracy (Chapter 9),
- **A novel method for the high-quality visualization of classifier decisions** in few-data scenarios by using cycle-consistent activation maximization (Chapter 10).

1.4 Publications

Some of the ideas discussed in this work have been presented in international journal and conference contributions, comprising:

- **Katzmann, A., Mühlberg, A., Sühling, M., Nörenberg, D., Holch, J. W., & Groß, H. M. (2018, July). TumorEncode - Deep Convolutional Autoencoder for Computed Tomography Tumor Treatment Assessment.** In 2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

  ⇒ This paper analyzes the application of deep convolutional autoencoders for liver lesion growth prediction (Chapter 5).

- **Katzmann, A., Muehlberg, A., Sühling, M., Noerenberg, D., Holch, J. W., Heinemann, V., & Groß, H. M. (2018). Predicting lesion growth and patient survival in colorectal cancer patients using deep neural networks.** In 2018 International Conference on Medical Imaging with Deep Learning.

  ⇒ This publication extends the former ideas to an application for patient overall survival prediction (Chapter 6).

- **Katzmann, A., Mühlberg, A., Sühling, M., Nörenberg, D., Maurus, S., Holch, J. W., ... & Groß, H. M. (2019, October). Computed Tomography Image-Based Deep Survival Regression for Metastatic Colorectal Cancer Using a Non-proportional Hazards Model.** In International Workshop on PRedictive Intelligence In MEdicine (pp. 73-80). Springer, Cham.

  ⇒ Within this paper an implementation of a fully deep convolutional survival estimator is analyzed, which extends on the ideas of the commonly employed Cox proportional hazards model (Chapter 7).

- **Katzmann, A., Mühlberg, A., Sühling, M., Nörenberg, D., & Groß, H. M. (2019, April). Deep Metamemory-A Generic Framework for Stabilized One-Shot Confidence Estimation in Deep Neural Networks and its Application on Colorectal Cancer Liver Metastases Growth Prediction.** In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) (pp. 1298-1302). IEEE.

  ⇒ In this publication an application of confidence estimation in medical imaging scenarios is discussed, demonstrating the potential to be used for curriculum learning-inspired training schedules for improved test time performance (Chapter 8).

- **Katzmann, A., Muehlberg, A., Suehling, M., Nörenberg, D., Holch, J. W., & Gross, H. M. (2020, April). Deep Random Forests for Small Sample Size Prediction with Medical Imaging Data.** In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) (pp. 1543-1547). IEEE.

  ⇒ Within this publication, a novel meta-training framework for handling small and medium-sized datasets is presented, showing wide applicability across multiple medical imaging scenarios, which is going to be discussed in Chapter 9.

- **Katzmann, A., Taubmann, O., Ahmad, S., Mühlberg, A., Sühling, M., & Groß, H. M. (2021). Explaining clinical decision support systems in medical imaging using cycle-consistent activation maximization.** In Neurocomputing, 458, 141-156.

  ⇒ This paper proposes a novel framework for medical decision explanation based on Cycle-Consistent Activation Maximization, and will be further explained in Chapter 10.

Clinical evaluations of the above results have been presented at international conferences, including:

- **Nörenberg, D., Huber, T., Maurus, S., Jäger, N., Katzmann, A., Mühlberg, A., Moltz, J., Heinemann, V. & Holch, J.. (2019). Deep learning based radiomics and its usage in prediction for metastatic colorectal cancer.** European Congress on Radiology (ECR) 2019.

- **Nörenberg, D., Huber, T., Maurus, S., Kazmierczak, P., Jäger, N., Katzmann, A., Moltz, J., Ricke, J., Heinemann, V. & Holch, J.. (2018). Deep Learning Based Radiomics and Its Usage in Prediction for Metastatic Colorectal Cancer.** Annual Meeting of the Radiologic Association of North America (RSNA) 2018.

Throughout this work, it was furthermore possible to file patents for some of the implemented systems, including[3]:

- **Feature-Enhanced Computed Tomography for Deep Learning yielding Quantitative Imaging Biomarkers.** Muehlberg, A., Kaergel, R., Katzmann, A., Suehling, M., EP 3570288, US 2019/0355114. Published: 20.11.2019

- **Sparse Lung Nodule Characterization for Differential Diagnosis from CT images.** Katzmann, A., Kratzke, L. Muehlberg, A., Suehling, M., EP 3576020, US 2019-0370969. Published: 04.12.2019

- **Dispersion-based Tumor Analytics System.** Muehlberg, A., Katzmann, A., Durlak, F., Suehling, M., EP 3792871, US 2021/0082569. Published: 17.03.2021

- **Geometric Deep Learning-based Whole Tumor Burden Analytics System.** Muehlberg, A., Taubman, O., Katzmann, A., Suehling, M., EP 3836157, US 2021/0183514. Published: 16.06.2021

While this thesis has a focus on deep learning-based methods for medical image processing, more classical machine learning-based techniques can and have been similarly applied to the medical imaging datasets found in this work. They typically have specific pros and cons, which will be discussed in more detail in Chapter 3. While these works have strong interconnections with the work discussed in this thesis or do even include parts of it, the contributions below either have a more classical and/or medical, rather than methodological focus than this work and therefore can not be covered in detail within this thesis. Notable work which was created in parallel to this thesis includes:

- **Muehlberg, A., Katzmann, A., Heinemann, V., Kaergel, R., Wels, M., Taubmann, O., ... & Rémy-Jardin, M. (2020). The Technome - A Predictive Internal Calibration Approach for Quantitative Imaging Biomarker Research.** Scientific Reports 10(1) (Nature Publishing Group), 1-15.

  ⇒ Within this work, a novel system for automated biomarker calibration is presented, taking into account application-specific external factors which are compensated, leading to a higher reproducability

- **Mühlberg, A., Holch, J.W., Heinemann, V., Huber, T., Moltz, J., Maurus, S., Jäger, N., Liu, L., Froelich, M.F., Katzmann, A. and Gresser, E., 2021. The relevance of CT-based geometric and radiomics analysis of whole liver tumor burden to predict survival of patients with metastatic colorectal cancer.** European Radiology, 31(2), pp.834-846.

  ⇒ This paper proposes new biomarkers and highlights the role of geometric radiomics features for the prediction of patient survival in metastatic colorectal cancer patients.

[3] Additional patents have been filed for work with relation to this thesis (3 US, 3 EP, and 1 CN), but at the time of this publication have not yet been published.

- **Muehlberg, A., Kaergel, R., Katzmann, A., Suehling, M., et al. (2021) Unraveling the Interplay of Image Formation, Data Representation and Learning in CT-based COPD Phenotyping Automation: The Need for a Meta-Strategy**. Medical Physics

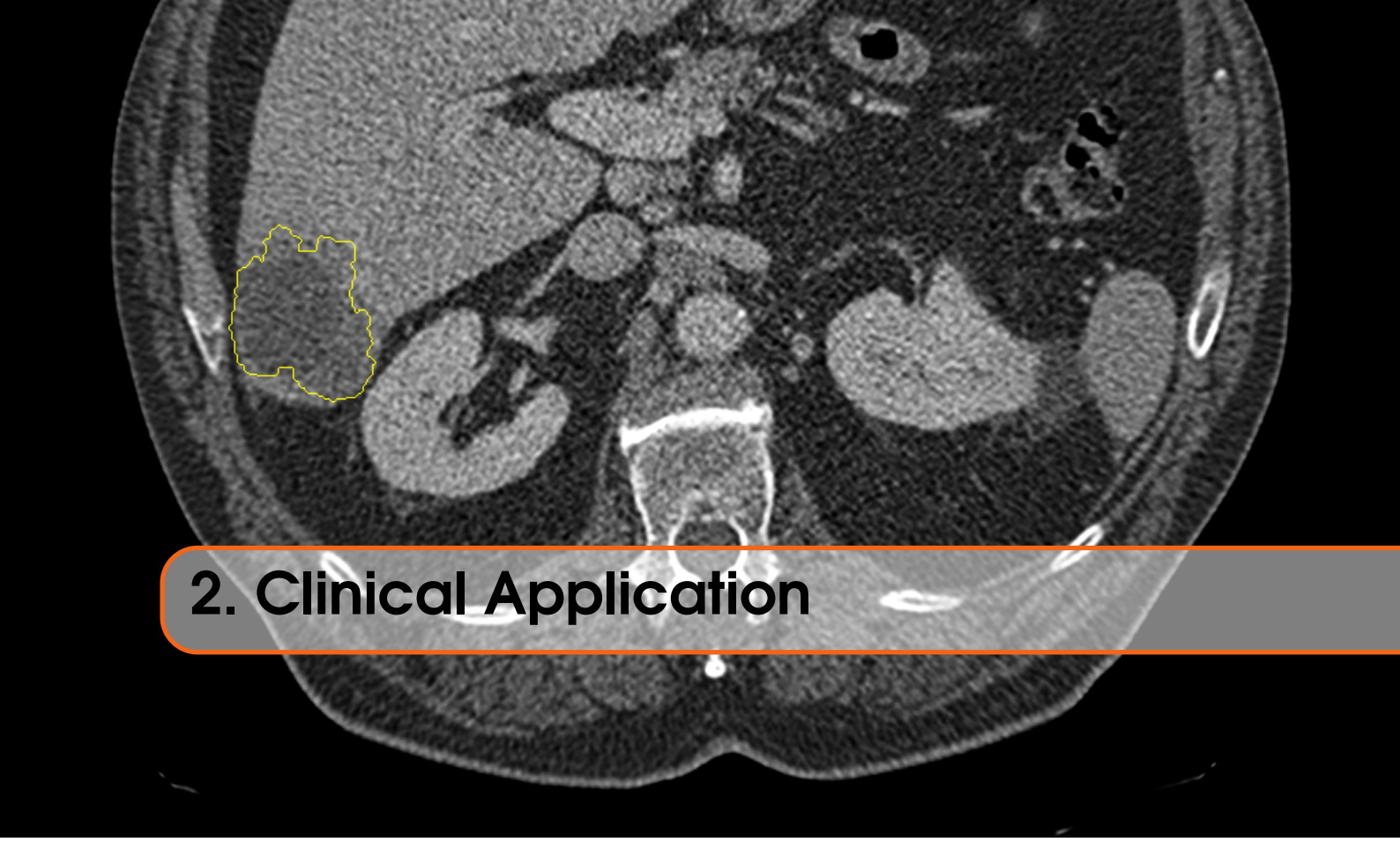
  $\Rightarrow$ Within this paper, state-of-the-art approaches for COPD phenotyping are analyzed and complemented by novel classical, as well as deep learning-guided methods.

II Scientific Background

2 Clinical Application 13
2.1 Cancer Statistics
2.2 Clinical Workflow

3 Radiomics 19

4 Deep Neural Networks 23



2. Clinical Application

"The science dealing with the preserving of health and with preventing and treating disease or injury" – according to the Cambridge Online Dictionary, this definition constitutes a comprehensive specification of the term *medicine* [Cambridge Online Dictionary 2021]. Evidently, if taking this wide definition, there is an overwhelming amount of potential medical applications which could be the subject of a doctoral thesis on medical image analysis. Although none of them can be quantified as more or less important, this work should aim for a clinical scenario that addresses as many people as possible, while having the potential of a large positive impact on their life quality.

With more than 17 million cases and over 9.5 million deaths per year, cancer currently poses the second leading cause of death after cardiovascular diseases, by 2040 is expected to even have a total number of 27.5 million cases and over 16.3 million deaths yearly. In 2018, the estimated years of life lost due to cancer were estimated to be as high as 15.3 years per case, and can cause treatment costs of as high as 150,000 USD per patient, contributing to overall healthcare costs of more than four times the amount of any other common health condition [AARP 2021; American Cancer Society 2021a; NIH-NCI 2021]. As clearly demonstrated by these numbers, a successful *early-stage treatment* of cancer would not only result in significant growth in life expectancy and quality of millions of people but could also reduce the consequential health costs by up to hundreds of billions of dollars globally, which in turn would be available for fighting other diseases.

From a medical image analysis perspective, cancer-related diseases are a rather promising target: Most cancer forms are routinely monitored by continious and fine-granular applications of non-invasive imaging techniques, such as ultrasound (US), and magnetic resonance (MRI), or computed tomography (CT) imaging. Furthermore, imaging data is routinely archived, and the strongly normed clinical treatment protocols require comprehensive documentation of the treatment process, which in turn can be used as information for the estimation of clinically relevant variables through classification and regression. Although utilizing a different physical mechanism, and in contrast to US imaging, MRI

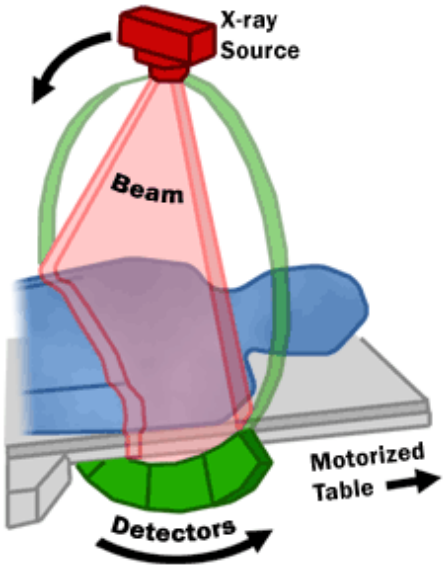


Figure 2.1: This figure shows the basic concept of CT imaging. The patient is slowly moved through a ring construction, called *gantry*, holding X-ray source and detectors, being opposed to each other. The gantry, and thus X-ray source and detectors, are revolving around the patient, yielding linewise attenuation curves which are concatenated to a *sinogram*. After collecting multiple sinograms, they can be transformed into a volumetric array, i. e. the final CT image, by using an image reconstruction technique, such as the Radon transform [TOFT 1996], the filtered backprojection [Brooks et al. 1975] or the iterative reconstruction [Willemink et al. 2013a,b]. Source: [FDA 2018]

and CT both yield volumetric images by slice-wise scanning of the patient (cf. Fig. 2.1). If available, volumetric data is generally to be preferred, due to the significantly larger amount of usable information. Amongst these two, computed tomography is particularly promising, as CT images are normed to a common scale called "Hounsfield Units", or *HU values*[1]. HU values are a measure of radiodensity and are normed in such a way that an HU value of -1000 represents the attenuation of air, and 0 represents the attenuation of distilled water at standard pressure and temperature (i. e. 273.15 K at an absolute pressure of 1 bar). In contrast to other imaging techniques, such as US or MRI, CT images can thus be interpreted as *absolute*, neglecting the influence of patient position and geometry, measurement apparatus positioning, the imaging device or scanning parameterization, etc., i. e. the *scanning protocol*[2]. While MRI in direct comparison generally provides better soft-tissue contrast, devices are significantly more costly, and thus generally less affordable and available in smaller clinics, rural areas, or emerging and developing countries. US imaging, on the other hand, being rather inexpensive, is subject to strong variations across

[1] Hounsfield units are named after Sir Godfrey Hounsfield who suggested the scale, and who for developing the computed tomography imaging technique together with Allan MacLeod Cormack received a Nobel Prize for Physiology or Medicine. He was later honored as a Commander of the Order of the British Empire (CBE) and a Fellow of the Royal Society (FRS) of London.

[2] In fact, the scanning protocol *is* important for the finally measured HU values to a relevant degree, as was demonstrated by Muehlberg, Katzmann, et al. [Mühlberg et al. 2020]. However, especially in comparison to other scanning modalities, such as MRI or US, these variations are rather low, and within a wide range of applications can be neglected.

views, such as beam width, side lobe, reverberation, and comet tail artifacts [Feldman et al. 2009], generally non-quantitative, and thus less suitable for an automated image analysis process. Thus, in the course of this work, CT imaging data is used. However, most of the approaches discussed in this work can likely be transferred to other imaging modalities, too.

In the following chapters, various methods will be proposed for a mostly automated clinical decision support, leading to estimates and diagnoses regarding cancer growth, treatment response and recovery chances, and even patient survival expectations. Clearly, these methods are currently far off from an actual clinical application. Next to the ethical implications (see Chapter 11) it would be possible to list a vast variety of reasons for this, but the two most prominent, also being the strongest market barriers, are the unsolved questions of:

(a) the consequences of potentially resulting treatment errors, and,
(b) the questions of liability in case of such an event.

In clinically approved products, a common way to combat this is to provide multiple proposals, but require the clinician to actively take the final decision, e. g. by choosing specific presentation styles which enforce this. While the automation of non-diagnosis-related tasks, such as simple measurements or contouring, is less strictly regulated and thus can be seen as state of the art already, the legal framework for more sophisticated solutions, slowly follows the state of the art, reflected by the FDA Artificial Intelligence and Machine Learning-based Software as a Medical Device (SaMD) Action Plan [Food et al. 2021]. However, quality is a highly important factor: There are demonstrated cases of deep learning-based systems clearly outperforming even experienced practitioners, which have consequently received FDA approval [Benjamens et al. 2020].

It has to be noted that any diagnostic proposal might finally influence the clinician's decision and may thus have the potential of inducing harm to the patient. Within the context of this work therefore *all* evaluations have been conducted on retrospectively acquired data for which an ethical council approval was given in advance.

2.1 Cancer Statistics

Amongst the various tumor entities, lung and colorectal cancer (**CRC**) are of particular interest, as they are the two leading causes of cancer-related deaths in modern societies. In the U.S. alone, lung and colorectal cancer were responsible for more than 140,000 and 50,000 deaths, respectively, in 2019 only [American Cancer Society 2017a]. Both, lung and colorectal cancer, tend to show significantly reduced survival rates especially if distant metastases are present. Commonly metastases can be found in the lung, brain, and liver, where they may lead to organ compression, organ failure, and ultimately death [Holch et al. 2017; Misiakos et al. 2011]. Both clinical entities go hand in hand with a significantly reduced patient lifetime, i. e. 14.9 and 15.6 years of life lost on average for lung and colorectal cancer, respectively [NIH-NCI 2021].

2.1.1 Lung Cancer

Lung cancer can be clinically categorized into multiple sub-classes. The most important classification is small-cell (**SCLC**) vs. non-small-cell lung cancer (**NSCLC**), as the

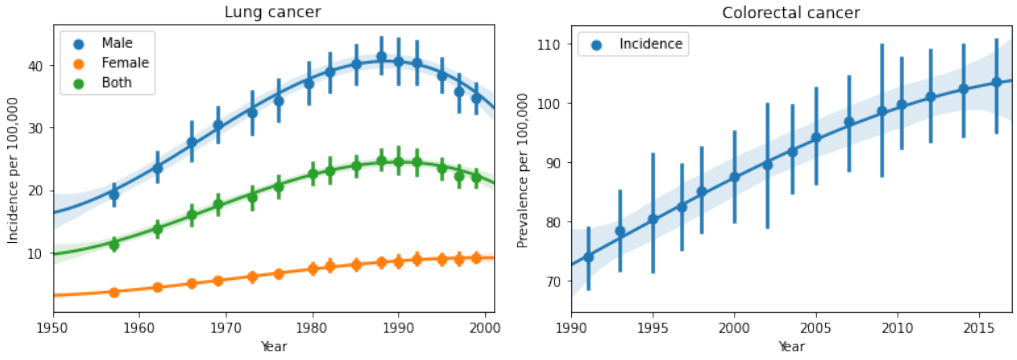


Figure 2.2: **Left:** Incidence of lung cancer per 100,000 inhabitants for man, women and both over time, averaged over 175 countries for which data was available. As can be clearly seen, the lung cancer incidence steadily drops for men and rises for women, both being directly linked to smoking behavior. **Right:** Prevalence of colorectal cancer cases over time, evidently showing a significant increase globally. Data from [Roser et al. 2015].

clinical treatment between both differs. Due to its much higher prevalence, NSCLC is the clinically most relevant form, accounting for 80-85 % of all lung cancers, in contrast to only 10-15 % with SCLC. SCLC, however, is significantly more deadly. While the average 5-year survival rate for *localized* NSCLC lies at around 61%, it reduces to only 27% for *localized* SCLC. More than that, SCLC tends to metastasize early, i. e. becomes *non-localized*, leading to an average 5-year survival rate over all stages of only 7 %, compared to 25 % percent for NSCLC [American Cancer Society 2021b,c,d,e]. Globally lung cancer numbers remain mostly stable, although incidence on average starts to decrease with a strong imbalance between male and female persons, largely due to a change in the number of smoked cigarettes ([Flanders et al. 2003], cf. Fig. 2.2). Within this work, automated lung cancer assessment is mostly discussed in the Chapters 9 and 10.2.

2.1.2 Colorectal Cancer

In contrast to lung cancer, localized colorectal cancer (**CRC**) can typically be treated very successfully. This is usually done by a full resection of the primarius. Unfortunately, however, only 39% of all colorectal cancer patients are diagnosed at an early, non-metastatic stage, as at the beginning of the disease colorectal cancer tends to show only a few unspecific or even no symptoms at all.

For metastatic colorectal cancer (**mCRC**), i. e. if distant metastases are present, these can typically be found in organs such as the liver, brain, and lung, leading to a significantly reduced life expectancy. This is clearly expressed by numbers: While the average 5-year survival for localized CRC lies at around 90 %, it drops to 71 % if locally spreading, and to only 14 % if distant metastases are present [Clinical Oncology 2021]. Globally, the amount of colorectal cancer cases is steadily increasing (see Fig. 2.2). While multiple authors suggest a link to changed lifestyle (e. g. [Hausen 2012; Thanikachalam et al. 2019]) for which recent research also implies an influence on the treatment outcome after diagnosis (cf. [Zutphen et al. 2017]), other reasons also include a significant increase in life expectancy, as well as larger screening efforts and a better diagnosis. Within this work, colorectal cancer data will in particular be used in Chapters 5, 6, 7, 8 and 9.

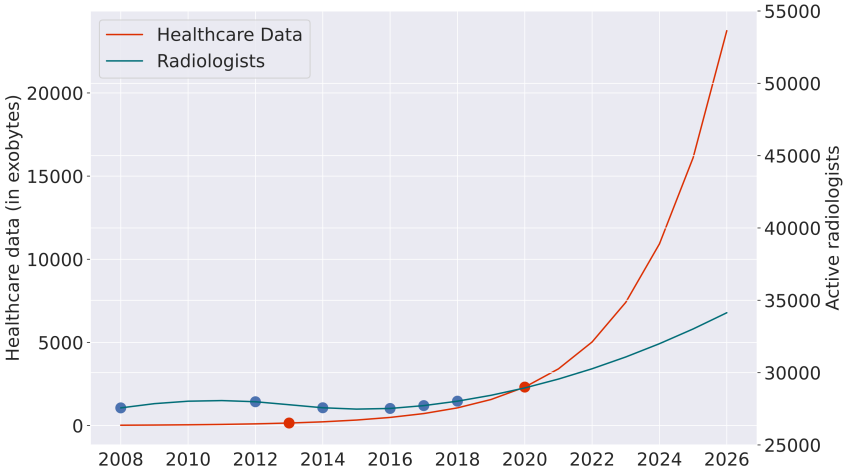


Figure 2.3: Number of active radiologists in the U.S. (blue) and amount of acquired healthcare data (red) per year. Notably, while the number of trained radiologists remains largely stable, the amount of healthcare data is rapidly increasing, clearly creating a need for algorithmically supporting and automating tasks for increasing the clinical throughput, while simultaneously ensuring an equal or even improved level of treatment quality. Data from [AAMC 2021; RBC Capital Markets 2021]

2.2 Clinical Workflow

Depending on the tumor entity, the clinical workflow may differ. For both, lung and colorectal cancer, the disease is regularly monitored via computed tomography imaging, with typical intervals being 8 to 12 weeks between two subsequent scans. For colorectal cancer, whenever possible the first step of treatment is the full resection of the primarius. However, as mentioned above, over 60 % of all CRC patients already have metastases at the date of first diagnosis (**DOFD**). The most common site of mCRC metastases is the liver, being particularly susceptible as the main filter organ of the human body, followed by the lung and brain. Around 50 % of all CRC patients develop liver metastases in contrast to 15-20 % with lung and only 1-4 % with brain metastases [Damiens et al. 2012; Mongan et al. 2009]. As a result, treating colorectal cancer typically involves at least half-year scans over three years even if resection was successful and if no metastases are present. As liver metastases are common, a fine-granular assessment of the liver is obligatory. If metastases are present, these are usually treated by local or systemic chemotherapy, radiation, or resection, if possible. Advanced therapy options may include hyperthermia, radiofrequency ablation, or cryotherapy.

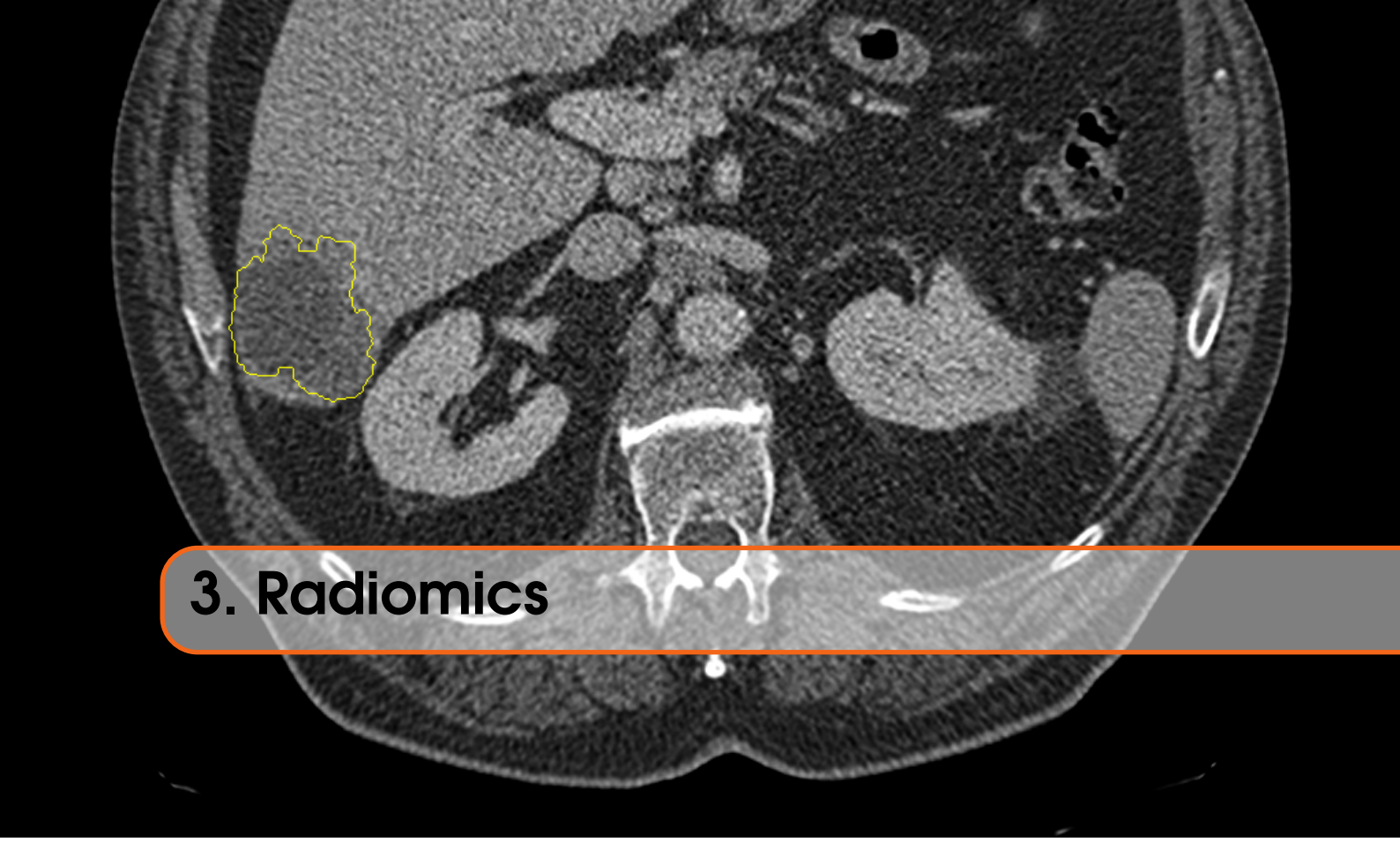
Similarly, the primary treatment for non-small-cell lung cancer is the resection of the primarius and infiltrated tissue, although an interventional resection is significantly less likely, which is clearly reflected by the statistics above. Generally, treatment of lung cancer is more difficult and typically involves less surgical, but instead systemic treatment.

For both entities, all lesions are monitored in regular follow-ups. This includes the scanning, measurement, registration, and re-identification of each lesion. Each of these tasks provides considerable potential for automation, which is currently only partly leveraged. The subsequent clinical assessment involves an estimate of future progression, which – although being based on treatment protocols and clinical experience – is qualitative in its

nature, meaning that similar to the previous processes it may significantly vary in interpretation, reliability, and quality. This was for example underlined by De Vries et al. [2014], demonstrating that the clinical outcome of cancer patients is significantly influenced by the practitioner's psychological constitution, and Waite et al. [2019], demonstrating that radiologists mostly automatically focus on relevant regions in medical images by using video-oculography and that this process was faster, more precise and more accurate for experienced radiologists.

As shown in Fig. 2.3, the amount of medical imaging data is steadily increasing, and clearly surpasses the amount of newly trained radiologists. As was discussed in Sec. 1.1, a typical radiological assessment for a CT scan comprising multiple hundred images on average is conducted in less than 15 minutes. Thus, as already pointed out at the beginning of Chapter 1, both throughput and quality could strongly benefit from an automated case preparation which independently composes a diagnostic draft by the means of computer vision and deep learning. With the automation of clinical tasks, these algorithms have the potential to significantly improve healthcare by enabling the radiologist to spend more time on the actual case assessment, with broad automation especially due to the increasing amount of data becoming more and more needed.

As discussed initially, deep learning has had its largest successes when trained on equally large amounts of *well-structured* training data. In contrast to this, the amount of publicly available and especially well-structured data in medical imaging is typically scarce, as these data are difficult to obtain, and often have heterogeneous quality (cf. Chapter 1.1). Therefore, this work will focus on the successful implementation of deep learning-based solutions for medical imaging when only a few data are available. It will propose novel architectures for working with this data, analyze their strengths and shortcomings, and finally demonstrate how to even get a deeper insight into highly complex decision-making processes if only small datasets are available.



3. Radiomics

While this work will focus on deep learning-based systems, it has to be noted that especially in the field of **radiomics**[1] has given a new thrust to the research in medical image analysis, and will at multiple times serve as a comparison within this work. The term "radiomics" was originally introduced by Kumar and Aerts [Kumar et al. 2012]. Radiomics is a neologism built upon "radiology" and "-omics", chosen in analogy to other disciplines such as genomics and proteomics, and expressing the *systematic* analysis of radiological imaging data. While the impact was only limited at that time, after the later "Nature Communications" publication [Aerts et al. 2014], radiomics became one of the most active fields in medical research. In a nutshell, as depicted in Fig. 3.1, given a set of images $\boldsymbol{X}$, radiomics is characterized by the successive application of:

1. **Filtering** - Apply a set of predefined preprocessing operations $\boldsymbol{P}$ independently to each input image (typically image filters),

2. **Feature extraction** - For each image and each filter, extract a set of predefined image features $\boldsymbol{F}$, using an image mask if given,

3. **Feature selection** - Select several features based on a predefined selection criterion, or by systematically varying selection criteria,

4. **Predictor training** - Train a predictor based on the selected subset of the $|\boldsymbol{P}| \times |\boldsymbol{F}|$ features of the $|\boldsymbol{X}|$ images, either using a predefined machine learning model or by systematically varying models.

[1] While often the term is written with a capital letter ("Radiomics"), with becoming more widespread the capital is typically omitted and the word is written in lowercase only. In the following, the lowercase writing will be used.

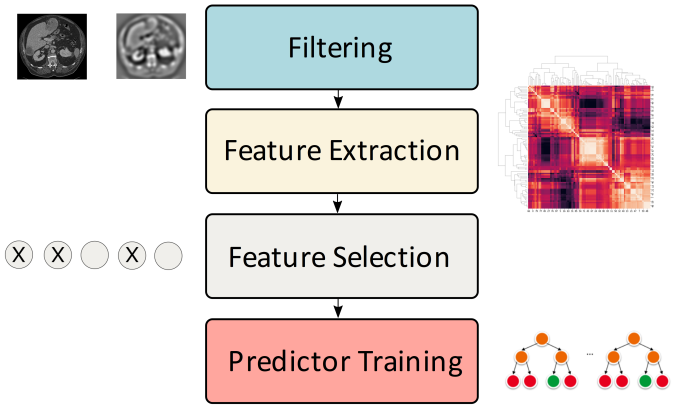


Figure 3.1: Overview on the basic radiomics process, consisting of the systematic application of image filtering, feature extraction, feature selection and finally predictor training.

Radiomics is thus not a novel methodology itself, but instead describes relevant tools for the process of a *systematic* analysis, rather than case-to-case feature design, selection, and classifier training. Radiomics provides a variety of benefits, most prominently, the possibility of a combination of hand-crafted, potentially meaningful features on one side, and a simple and easily-understandable machine learning model, e. g. logistic regression, on the other. Such a combination allows for analysis with strong human comprehensibility, thus being a striking advantage over deep learning-based classifiers, which tend to be difficult to explain and understand. In many cases, the combination of simple features can yield a comparable or even equally good quality[2]. Similarly, with a small known and successfully reproduced feature set, such as the 4-feature Aerts signature [Aerts et al. 2014], radiomics can be applied to very small datasets for which deep learning might be infeasible, too.

Although radiomics captivates with an easier understanding and comprehensibility, the adequacy of the chosen features remains unsure yet. First, there is no general guarantee that radiomics features are sensible for the problem at hand. In contrast to pure, data-driven methods, radiomics features are extracted in advance, thus important information might have been erased in the feature extraction step already. Secondly, radiomics typically comprises a large amount of hundreds or even thousands of features, many of them highly correlated, making the model prone to false discovery, and unstable model parameters (cf. multicollinearity[3]), and overfitting. If region-based features are used[4], the region annotation itself may pose a major source of variance. In summary, radiomics results have shown to vary strongly, depending on the used modality, the scanning protocol, the patient

[2] A prominent example for this is the so-called Apgar-Score for health assessment in newborn [Apgar 1952]. For radiology, similar approaches exist, such as the LA950 threshold for lung emphysema detection [Hueper et al. 2015].

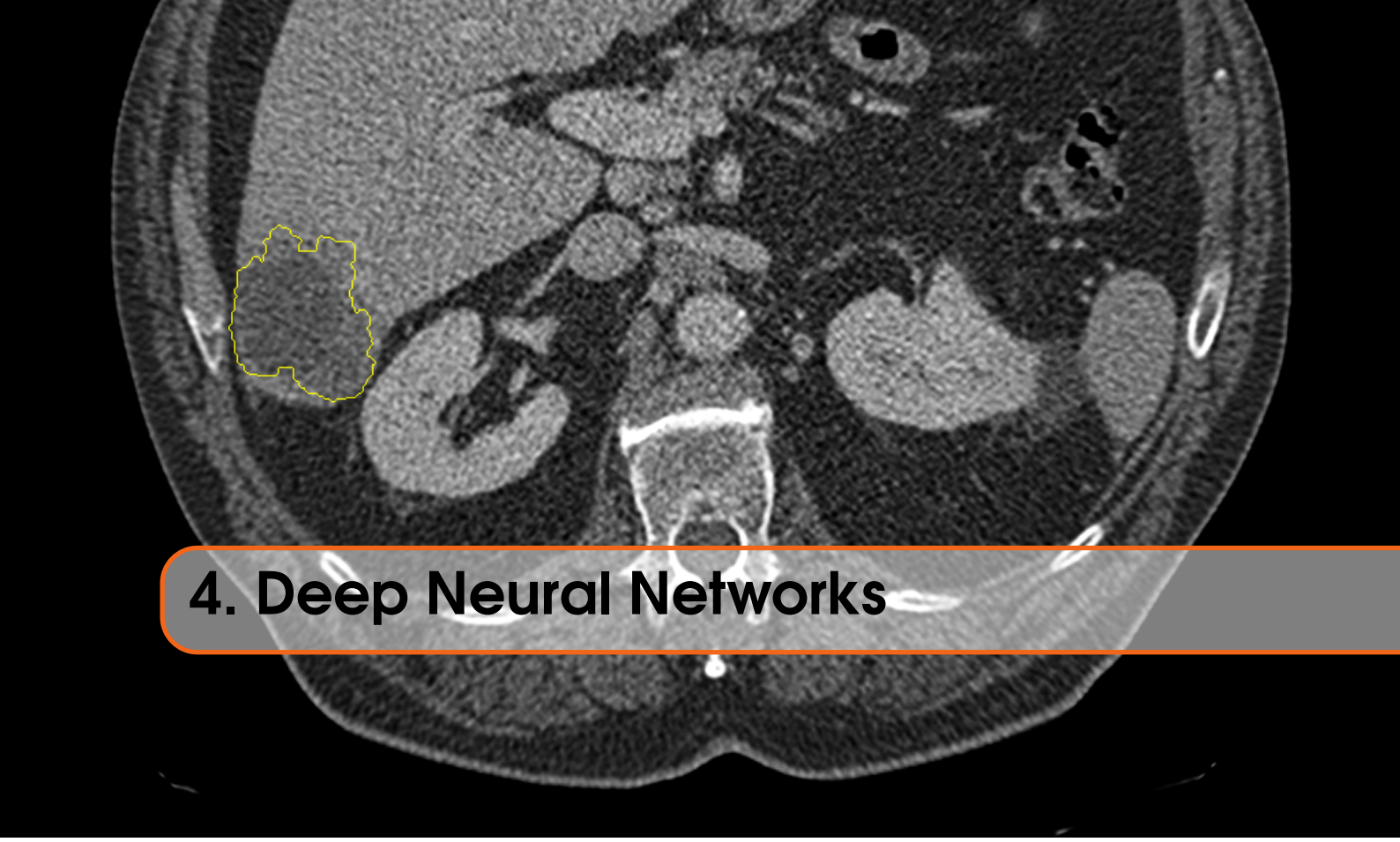
[3] Multicollinearity describes the situation that multiple predictor variables are strongly correlated and thus can predict each other. As a result, model parameters are unstable, as an indefinite number of model parameterizations would yield similar results.

[4] Radiomics is often applied to a particular region-of-interest (ROI) within the image, which is typically defined by a binary mask and allows to infer features such as the compactness or sphericity, with the former also being part of the 4-feature signature from Aerts et al. [2014].

geometry, and many more factors [Moltz 2019; Mühlberg et al. 2020, 2021b], and were often found to be non-reproducible. This is underlined by meta-studies demonstrating that positive results are highlighted and negative findings downplayed in as high as 97.8 % of the published work on radiomics [Song et al. 2020]. Lastly, in contrast to the above-mentioned advantage, the comprehensibility problem is mainly redirected towards the feature construction process. Especially for more advanced but often used features, such as the Haralick gray level co-occurrence matrix (GLCM) or the neighboring gray level dependence matrix (NGLDM), an intuitive understanding of the features is not necessarily given. This especially becomes relevant when radiomics is not only used for feature candidate exploration but when a larger amount of features is combined for classification, yielding a model complexity that can become evenly difficult to understand like other machine learning models such as deep neural networks.

In summary, radiomics has been demonstrated to be a powerful toolbox for the automated analysis of medical imaging data. Its advantages especially lie in the comparatively higher comprehensibility and its potential for simple and feasible clinical application when only a few features are used and few data are available. Radiomics has successfully demonstrated its applicability to a large variety of medical imaging scenarios and even clinical products. It has indisputably given an enormous new impetus to the field of medical image analysis.

A major downside of radiomics is its very strong sensitivity to environmental parameters, such as the used acquisition protocol, the scanner physics, the patient geometry, and similar [Mühlberg et al. 2020]. Its comprehensibility only stays high as long as the number of used features stays low. As radiomics uses pre-defined hand-crafted features, however, the features do not necessarily represent the full, clinically important image variance, often making the use of a larger number of features obligatory, and thus the applicability to small datasets more difficult. Finally, radiomics mostly represent the classical machine learning domain. In the last years, however, deep learning has started to become the state of the art in most machine learning applications, with a demonstrably higher performance on a variety of problems. In accordance, the increase in publication numbers on deep learning for medical image analysis has clearly surpassed that of radiomics, with over 90 % of submissions on the *International Conference on Medical Image Computing and Computer Assisted Intervention* (MICCAI), the largest conference in the field, being about deep learning since 2020.



4. Deep Neural Networks

In recent years, as for example pointed out in [Katzmann et al. 2018b], Deep Neural Networks (**DNNs**) have been employed for a variety of medical imaging applications, such as tumor [Havaei et al. 2017], multiple sclerosis [Brosch et al. 2015] and whole-organ segmentation [Roth et al. 2015], vessel tracking [Wu et al. 2016], and many more. Some of the proposed architectures even found a larger distribution outside of the medical domain, such as the well-known U-Net from [Ronneberger et al. 2015], now being a standard approach for image segmentation in very different fields, such as robotics. As emphasized in Part I, most breakthroughs in the field of artificial intelligence in the last decade can be attributed to deep learning-based approaches. Notably, its power to infer meaningful latent variables in a pure, data-driven fashion allows it to extract relevant information from images by constructing the most informative features itself, which in direct comparison to radiomics is likely the largest advantage, as it can leverage arbitrary information, and may ultimately lead to a significantly improved estimation accuracy (cf. [Mühlberg et al. 2020]).

On the other hand, training DNNs comes with some disadvantages, most of them being a result of model complexity. DNNs typically contain hundreds of thousands, millions, or even billions [Radford et al. 2019] of parameters that have to be trained prior to a successful prediction by the model. In classical statistics, the so-called “One in ten rule” [Harrell Jr et al. 1984] states that the number of samples should be typically at least ten times the number of parameters that are to be estimated from the distribution at hand – a requirement which in the deep learning domain is rarely met in practice. While datasets in some computer vision domains contain at least some thousands of samples, acquiring these amounts for medical scenarios is rather difficult. Even in medical specialization centers the amount of patients with a specific medical condition does rarely exceed the number of hundreds per year, typically with only a subset of these patients being eligible for the question at hand (cf. Chapters 5 and 6). As described in Chapter 1.1, further reasons include data security, data regulation, and data privacy issues, which – although

having outstandingly relevant reasons – are an obstacle to the collection of large and freely accessible multicentric medical imaging databases from a data scientist's point of view, reinforcing the problem of finding optimal model parameters, from a mathematical perspective being an underdetermined problem. In turn, the model provides no guarantee that solutions are optimal or even near to being optimal. In fact, recent studies have emphasized that deeper networks tend to generalize better than shallow networks with comparable parameter amounts (e. g. [Szegedy et al. 2015]), and, for the particular case of natural language processing architectures, it could be shown that models with an extremely large parameter space even perform well in few-shot learning environments (e. g. [Brown et al. 2020]). However, while this can reduce the issue at hand, it still needs explicit addressing. As a result, many of the algorithms presented in this work either directly address (cf. Chapters 5 and 6) or even explicitly focus (cf. Chapter 8) on the trainability of DNNs with only small datasets.

As pointed out above, one of the largest issues for the application of DNNs in medical applications is the overall low comprehensibility. While many applications do not necessarily require an in-depth understanding of the actual processing as long as possible mistakes are limited in their consequences, this is clearly not the case in a medical scenario, in which wrong decisions can easily lead to a significantly shortened patient survival or death. In fact, even the top-5 strongest research university hospitals in the US *do not employ* a single fully-automatic machine learning-based system in clinical practice, particularly mentioning skepticism as a reason [Sennaar 2021]. Thus, creating comprehensible systems is therefore seen as a major goal toward a market transition of DNN-based algorithms into the daily clinical workflow (cf. explainable artificial intelligence, **XAI** [Linardatos et al. 2021]). For this reason, multiple chapters within this work will explicitly take into account algorithmic comprehensibility. Finally, with Chapter 10, a whole section is specifically dedicated to this highly important issue.

II Oncological Decision Support

5 Liver Lesion Growth Prediction .. 29
5.1 Introduction
5.2 Methods
5.3 Experiments
5.4 Discussion
5.5 Conclusion

6 One-Year Survival Estimation ... 39
6.1 Introduction
6.2 Methods
6.3 Experiments
6.4 Discussion
6.5 Conclusion

7 Deep Survival Regression 47
7.1 Medical Background
7.2 Related Work
7.3 Methods
7.4 Experiments
7.5 Discussion
7.6 Conclusion

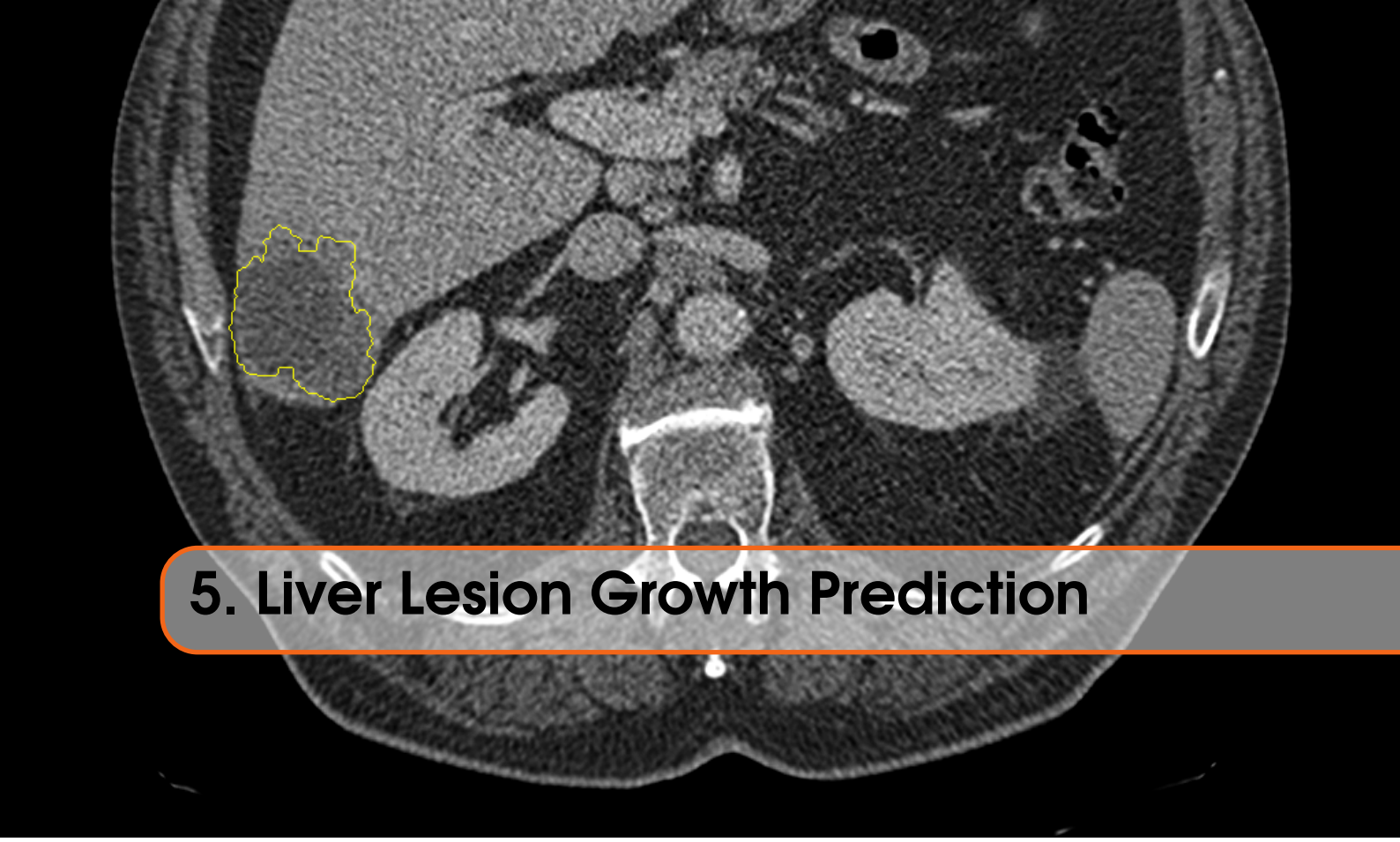
In Chapter 2, we have already discussed a variety of reasons that oncological image classification is one of the most promising targets for medical image analysis (cf. Chapter 2), e. g. as data is quantitative, and major parts of the clinical assessment can be automated, such as lesion measurement and re-identification. The latter arguments both aim at patient throughput. While turnaround times are a clinically highly relevant factor, however, quality is even more important. As already discussed in Chapter 2.2, the clinical expertise across practitioners may vary. For example, shown in a study from Moltz et al. [2011], manual liver lesion delineation in CT images can easily result in a high inter-observer variability, showing volume differences of up to 35 % across annotators. Recent research has further demonstrated that both accuracy and performance in complex human decision-making processes can depend on a variety of highly non-related factors, such as daytime, shift, fatigue, and even the presence or absence of the expectation of a full meal (cf. [Alshabibi et al. 2020; Cowley et al. 1997; Danziger et al. 2011]). Thus, as pointed out in [Katzmann et al. 2018b], a reliable estimation based on deep neural networks "may enable better therapy planning, deeper insight into tumor growth dynamics, [and] a greater patient turnover". Taken together, there is a clear need for an automated, quantitative and deterministic assessment. The following part will therefore cover multiple applications of deep neural networks for automated treatment outcome prediction in oncology using computed tomography imaging data.

Like also pointed out in [Katzmann et al. 2018b], it should be noted that within a clinical workflow, as emphasized by Glimelius et al. [2013] and Van Cutsem et al. [2016], a radiological assessment amongst other information typically includes an evaluation of:

- visual lesion appearance (e. g. shape, size, density),
- a histopathological assessment,
- oncologically relevant blood values and biomarkers, such as haemoglobin, antibodies, tumor markers (CA19.9, ...), etc.,
- the patient's demographic data (age, gender, ...),
- the patient's medical history.

The design of a clinical treatment plan comprises the collection of these and other information, as each may strongly influence the treatment plan. Thus, besides ethical arguments (cf. Part II), a thorough, prospective, clinical evaluation cannot be contained within this work, as a full, clinical assessment of an oncological disease requires the strong, continuous, multidisciplinary expertise of various clinical specialties.

In contrast, the following part is dedicated to the identification of opportunities for using relevant image information which previously had been unused. As the visual assessment is a key criterion for the creation of the treatment plan and its adaption over the course of the disease, the following chapters will analyze how medical image analysis through deep neural networks can be used in a clinical few data scenario to extract relevant diagnostic information, such as an estimate on future tumor growth and overall patient survival by leveraging yet unused image information through a data-driven approach.



5. Liver Lesion Growth Prediction

First, this chapter will analyze the prediction of lesion progression. More specifically, the scenario of liver lesion growth prediction for patients with metastatic colorectal cancer (mCRC, cf. Chapter 2.1.2) is chosen. As discussed in Chapter 2.2, a fine-granular assessment of liver lesions is obligatory for mCRC patients, as, while the primarius can often be resected, liver metastases pose a significant risk to life.

Recent research has shown clear correlations between the visual tumor appearance in medical imaging data, the future disease progression, and the overall patient survival (e. g. [Aerts et al. 2014], cf. Chapter 3). For leveraging this potential in a data-driven fashion, with [Katzmann et al. 2018b] a novel algorithm is proposed, using structural image information for tumor treatment response prediction, and being able to predict liver lesion growth with high reliability, thus having the potential to be a relevant step towards a semi-automatic oncological estimation of future disease progression.

5.1 Introduction

The clinical manual for the radiological assessment of liver lesions is the so-called *Response Evaluation Criteria in Solid Tumors* (**RECIST**, [Eisenhauer et al. 2009]). RECIST contains multiple criteria and allows for an overall patient assessment over the course of the disease. Within this work, the so-called RECIST lesion assessment for single lesions will be most important, is based on the maximum 2D lesion diameter, also called **RECIST diameter**, within one slice[1]. An example is depicted in Fig. 5.1. The single lesion assessment defines the lesion status by setting the current measurement in contrast to previous time points. The possible results and their criteria are depicted in Tab. 5.1.

In the clinical assessment, RECIST serves as an indispensable tool. Still, it has several shortcomings. First, it can be subject to large inter-observer variabilities [Rothe et al.

[1] RECIST was created to be available for various imaging modalities. As some of these, e. g. X-ray, generate only 2D images, the RECIST diameter is defined in such a way that it uses a 2D instead of a 3D measure. It thus assumes isotropic growth of lesions, which typically is approximately given for liver lesions.

Lesion status	Status Code	Criterion
Complete Response	CR	Dissappearance of lesion
Partial Response	PR	Shrinkage of at least 30%
Stable disease	SD	Neither significant growth nor shrinkage
Progressive disease	PD	Growth of at least 20%

Table 5.1: RECIST single lesion assessment criteria and status codes based on a comparison of the RECIST diameter between two successive timepoints.

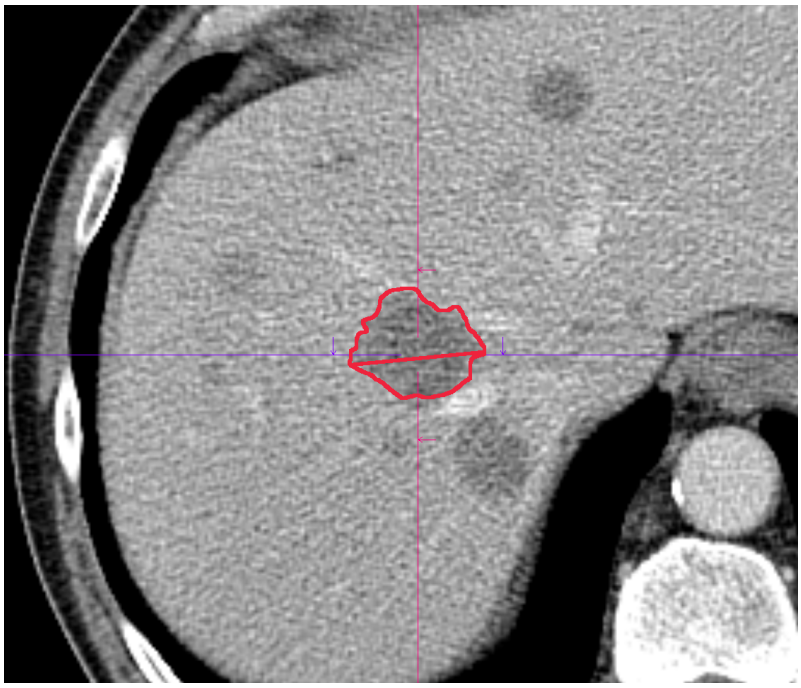


Figure 5.1: Liver lesion with manually delineated outline and RECIST lesion diameter.

2013]. Secondly, being a fully shape-based measure, it cannot leverage structural image information, such as the lesion heterogeneity, although being a known sign of biological activity (cf. [Ganeshan et al. 2013]). Finally, it poorly aligns with specific therapies, such as immuno-oncological treatments, which may show initial growth after treatment (cf. [Seymour et al. 2017]).

Finally, the RECIST assessment is a purely retrospective measure. Clinically, however, an *estimate of future progression* would be highly beneficial, as it would allow for an anticipatory escalation or de-escalation of the tumor therapy[2], by yielding a so-called *early assessment*. Therefore, this work will focus on a prediction of future lesion growth, as indicated by the RECIST lesion status, by using two subsequent time points[3].

Like mentioned in Chapter 3, with the work from Aerts et al. [2014], Kumar et al. [2012], and Lambin et al. [2012], automated tumor assessment using radiomics has become a highly active research field (e. g. [Bogowicz et al. 2017; Gillies et al. 2015; Leijenaar et al. 2013; Yip et al. 2016]). As was pointed in Chapter 4, while radiomics focuses on a combination of hand-crafted image features, deep neural networks, in contrast, can learn meaningful features in a purely data-driven fashion. Therefore, their value for liver lesion growth prediction will be assessed in the following.

At the time of its publication (cf. [Katzmann et al. 2018b]), the proposed algorithm was among the first to predict longitudinal liver lesion progression, presenting:

1. **A novel approach for CRC liver lesion growth prediction** from single-slice CT images of two separate time points (before- and within-treatment),

2. A successful demonstration of the applicability for **pre-treatment assessment** by using a single time-point only, and finally

3. A thorough evaluation showing **superiority over other radiological assessment measures** on the given data, such as the RECIST lesion diameter and volume.

5.2 Methods

Limited data poses a relevant obstacle to successful training. In [Katzmann et al. 2018b] therefore a two-step approach has been used. First, a convolutional sparse auto-encoder (cf. [Hinton et al. 2006; Krizhevsky et al. 2011, 2012; Ng et al. 2011]) has been trained for conditioning the network to create a sparse representation from the dataset. Secondly, the sparse representation is used for the actual classification task.

5.2.1 Autoencoder Network Architecture

A common way to combat overfitting, which might occur as a result of the imbalance between the number of model parameters and training data, is *data augmentation*, i. e. the

[2]Escalation or de-escalation in terms of tumor treatment describes the intensity of medication or treatment. The decision about escalation or de-escalation involves a tradeoff between estimated response and patient life quality. Notably, an oncological treatment plan typically involves an expertise-guided estimate of future response, naturally taking into account the RECIST measurements.

[3]It should be noted that subsequent time points may contain information on an *actually observed* progression, and thus that on this basis an extrapolation of future progress may partly be possible. The work in this and the following chapter, therefore, focuses on the *additional*, rather than *absolute value*, in the direct comparison. An in-depth analysis focussing on single-timepoint-based assessment is found in Chapter 8.

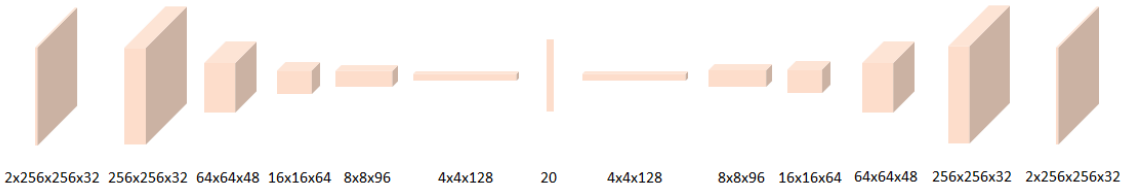


Figure 5.2: Autoencoder used for a training of sparse liver lesion representations. One slice of baseline and followup images are given as an input, with the model trying to reconstruct it at the output layer. Source: [Katzmann et al. 2018b] © 2018 IEEE

use of modifications such as image transformations to create virtually new data points. As discussed in [Katzmann et al. 2018b], the amount to which it can be applied is limited: "When having classification tasks, especially image transformations like varying rotation, shear, jittering, etc. can be used. However, even when combined with dropout [Hinton et al. 2012], batch normalization [Ioffe et al. 2015], 2D or 3D image augmentation, the degree to which augmentation results in additional performance is limited, as images are still highly correlated. The limit to data augmentation especially holds true for our task, as the underlying theory claims that phenotypical manifestations (e. g. specific tissue structure, size and shape of central necrosis, etc.) correlate with image structures and/or noise patterns. As these manifestations are currently a field of active research, it can not be said whether larger transformations are realistic in these terms, too. This reduces the amount to which image transformations can be done, so augmentation does not fully solve the problem of few data. Therefore, a main goal of the proposed approach was to keep the number of parameters as low as possible."

For this reason, in the following a 2D approach is employed, using axial slices (cf. Appendix A) in order to keep the amount of model parameters low[4]. As already pointed out above, the sparse autoencoder approach from Ng et al. [2011] is used for creating a meaningful latent space representation despite the low amount of available data. The autoencoder model architecture consisted of a simple ConvNet (cf. Chapter 9 Fig. 9.2), downsampling the input using leaky ReLU activation [Maas et al. 2013] and batch normalization [Ioffe et al. 2015], followed by a 20-neuron dense layer, i. e. the later sparse representation, and finally deconvolutional and upsampling layers in the reverse order to the initial convolutions. The overall architecture is depicted in Fig. 5.2, details can be found in the Appendix Tab. B.1.

5.2.2 Predictor Network Architecture

The above autoencoder is trained to create meaningful, low-dimensional representations of the input-space information. After its training, a second neural network, having only a very low parameter amount, is appended. Therefore, a very simple, empirically designed 2-layered network architecture is added after the low-dimensional representation of the autoencoder, which consists of 8 dense neurons followed by a two-neuron softmax output layer, corresponding to the possible predicted outcomes (growth/non-growth, see Sec. 5.3.1). This results in a total number of 218 parameters to be trained after autoencoder

[4] As the longitudinal resolution of CT scans is typically lower than the sagittal and coronal resolutions, amongst the possible projections using axial slices retains the largest amount of image information.

pretraining. While more complex architectures of the predictor network have also been tried, they did not result in a significant advantage over this simple architecture. Possible reasons for this include a) the low amount of training data (see Sec. 5.3.1), and b) a loss of information in the encoder part of the autoencoder network. The full predictor network architecture can be found in the Appendix Tab. B.2.

5.3 Experiments

The above architecture has been trained on a liver lesion dataset of mCRC patients (Sec. 5.3.1). As a comparison to the presented approach, multiple reference classifiers have been trained (Sec. 5.3.2–5.3.3), and were finally compared against each other (Sec. 5.3.4).

5.3.1 Dataset

The used dataset consisted of 321 volumetric CT scans from 135 patients, scanned between 12/2009 and 02/2017. It contained a total of 460 unique liver lesions with fully volumetric segmentations at an average of 2,38 time points per patient. In total, 1,344 liver lesion volumes were available, of which after pairing (t_i, t_{i+1}, t_{i+2})[5], 805 samples remained. 419 samples have been used for training and validation (325 positive, 94 negative) and 386 (304/82) for testing. The splits have been chosen randomly using a grouped split, such that all samples from the same patient are contained in either the train, the validation, or the test dataset. The data was acquired within the BMBF project PANTHER[6] and was continuously extended throughout the project. The PANTHER data will also used in Chapters 6, 7 and 8. As the data was acquired retrospectively, no unified scan protocol has been used. Thus, the acquired images suffer from a high level of heterogeneity, including variation in contrast enhancement, noise level, and voxel resolution.

For unification, the dataset has been resampled to an isotropic voxel size of 1x1x1 mm using cubic interpolation. As mentioned in Sec. 5.1, a 2D approach was used to keep the parameter amount low[7]. As most lesion diameters $\emptyset$ were in the interval $\left[\emptyset_{Pr(\emptyset)\leq 0.1}, \emptyset_{Pr(\emptyset)\leq 0.9}\right] = [11.3mm, 53.3mm]$, a window of 80mm x 80mm centered around the lesion center of mass has been extracted (cf. [Nibali et al. 2017]), using a resolution of 256x256 pixels to assure that no image information is lost[8]. To reduce the influence of different contrast enhancements due to non-unified contrast phases, all images underwent a histogram equalization. An example lesion at two timepoints and their histogram equalized counterpart is depicted in Fig. 5.3.

In accordance with the RECIST guideline, tumor diameter was measured as the longest diameter within one slice. The data was labeled according to RECIST lesion assessment criteria to discriminate lesion progression (PD) from non-progression (CR/PR/SD), i. e.:

$$y_i = \begin{cases} 1 & \text{if } d_{i,t+1}/d_{i,t} \geq 1.2 \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

[5] The method uses two time points for the prediction of future growth, which is derived using a third time point.

[6] The PANTHER project was a BMBF project conducted between 2016 and 2020 by a consortium formed of MeVis Breastcare, the Fraunhofer MEVIS, the KUM Munich, and Siemens Healthineers, with PANTHER being an acronym for "Patientenorientierte onkologische Therapieunterstützung", engl. patient-oriented oncological therapy support.

[7] Lately Perslev et al. [2019] proposed a method for semi-3D processing with similar benefits.

[8] Later approaches, such as [Katzmann et al. 2018a], have used smaller image resolutions (see Chapter 6).

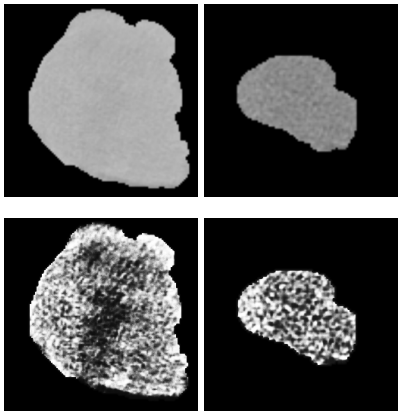


Figure 5.3: Histogram equalization for a lesion baseline (12/2015, left) and followup (04/2016, right) pair (top) with its respective histogram equalized variant (bottom). Source: [Katzmann et al. 2018b] © 2018 IEEE

for lesion diameters $d_{i,t}$ of lesions i at timepoint t. As the study at hand aimed for an assessment relative to the current time point, in contrast to the standard RECIST criterion, the current timepoint has been taken as the reference t for label assignment, rather than the time point of the best response t_{BR} (cf. [Eisenhauer et al. 2009]).

5.3.2 Classifier Baseline

Treatment response prediction is a relatively new field, and as pointed out in Chapter 1.1. Therefore, only a few works have yet taken into account tumor disease prognosis using machine learning, with no system to the time of this study to estimate the per-lesion progression of liver metastases. Relevant work with respect to the task at hand conducted a longitudinal analysis of lesions using radiomics features of volumetric CT data [Aerts et al. 2014], and was discussed in Chapter 3. It aimed, however, at non-small-cell lung cancer instead of mCRC[9]. Within this study therefore the clinical RECIST measure has been used to provide a baseline comparison. As some recent work (cf. [Hayes et al. 2016; Rothe et al. 2013; Xiao et al. 2015]), has identified lesion volume, rather than diameter, as more predictive, additionally lesion volume is compared. For each predictor, raw and delta features are provided based on input sets X_r for measures m_r with $x_{r,i,t} \in X_r$ as:

$$x_{r,i,t} = \begin{pmatrix} m_{r,i,t} \\ m_{r,i,t-1} \\ m_{r,i,t} - m_{r,i,t-1} \\ \frac{m_{r,i,t}}{m_{r,i,t-1}} \end{pmatrix} \tag{5.2}$$

with $m_{r,i,t}$ being the measure r (RECIST diameter or lesion volume) for sample i at timepoint t. For each dataset X_r, one classifier was trained. Hyperparameter optimization was done using 100 iterations of randomized search cross-validation with inner nested 10-fold grouped cross validation. Train-test split was done using an outer 10-fold grouped cross-validation using the patient identifier as the grouping parameter.

[9] A thorough comparison to the radiomics approach can later be found in Chapter 6.

5.3.3 Deep Network Training

The deep network was trained on an NVIDIA DGX-1 using Keras and Tensorflow [Chollet et al. 2015; Martín Abadi et al. 2015]. The data was divided into distinct datasets using label stratification to ensure a comparable label distribution between the training and test set. The sparse autoencoder for lesion representation was trained using mean absolute error. Afterward, the actual classification network was appended and trained using categorical crossentropy. For optimization, the NAdam optimizer[10] [Dozat 2016] was used. The learning rate η_i was annealed exponentially as a function of the current epoch i:

$$\eta_i = \eta_0 \cdot \left(\frac{\eta_{n-1}}{\eta_0}^{\frac{i}{n-1}^{\gamma}} \right) \tag{5.3}$$

with the initial learning rate set to $\eta_0 = 3 \cdot 10^{-4}$, and the target learning rate $\eta_{n-1} = 1 \cdot 10^{-7}$, the number of epochs n and $\gamma = 1.2$ being the learning rate exponent. For both, autoencoder and classifier training, adversarial training was used for regularization and test time performance enhancement as described in [Goodfellow et al. 2014]. Stratified sampling has been used to account for the faced label imbalance. Thus, for each sample i a sampling probability p_i with $m = 2$ classes has been assigned as:

$$p_i = \frac{\frac{1}{Pr(y=y_i)}}{\sum_{k=0}^{m} \frac{1}{Pr(y=y_k)}} \tag{5.4}$$

Inputs and outputs were scaled to the interval $[-.5, .5]$ using tanh activation to ensure a well-defined gradient. Additionally, mean image subtraction was used. Further, a modified version of the exact importance sampling from [Katharopoulos et al. 2017] was employed, multiplying the above sampling probabilities with the norm of the error gradient[11].

The autoencoder architecture was trained for 1,000 epochs. Due to the low dimensionality of the bottleneck layer, after training, the bias and variance errors did not differ remarkably. An example of a baseline-followup-pair before and after autoencoder preprocessing is given in Fig. 5.4. As depicted there, the autoencoder reconstruction quality as a result of the strong compression may be seen as rather low, implying that the autoencoder was *not* able to fully represent the original training data. However, as noted before, a rather coarse representation was in fact intended to avoid overfitting. As depending on the time point of treatment not necessarily two CT scans have been acquired already, two different classification networks have been trained. The first network is based on the low-dimensional representations of baseline/follow-up pairs as proposed above. The second network is trained analogously but feeds the duplicated baseline image as the input, i. e. serving as both baseline and follow-up to the autoencoder simultaneously.

Due to the low amount of available data, the training process of both networks converged after only a few epochs. The autoencoder layers were not fixed during training to allow for an adaption of network weights for the concrete training goal. Fixing the layers resulted in a lower performance [Katzmann et al. 2018b].

[10] NAdam combines the Adam optimizer from Kingma et al. [2014] with a Nesterov momentum.

[11] The multiplication results in the more frequent sampling of difficult samples, and in [Katzmann et al. 2018b] empirically resulted in improved performance. However, as the improvement was only marginal, this step has later on been dropped.

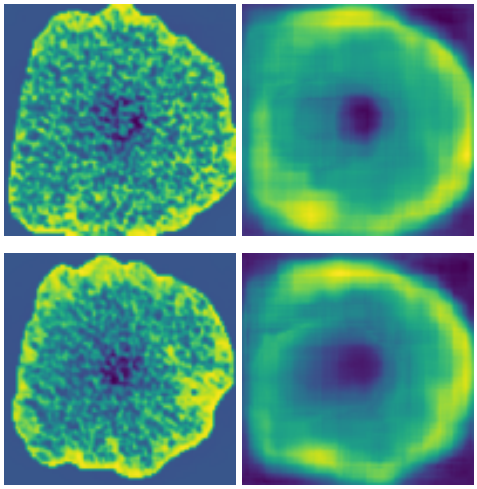


Figure 5.4: Baseline (top) and followup images (bottom) prior to (left) and after (right) autoencoder preprocessing. Source: [Katzmann et al. 2018b] © 2018 IEEE

5.3.4 Results

All trained classifiers were evaluated using F1-score, informedness (IFD), markedness (MKD), Matthews correlation coefficient (MCC), and area under the curve (AUC)[12][13], see Tab. 5.2. Significance was tested using 5,000 iterations of bootstrapping[14].

As was emphasized in [Katzmann et al. 2018b], the results show nearly equal values for RECIST- and volume-based prediction. Both were signficantly correlated with the ground-truth labels as stated by the $F1$, MCC and AUC values, having bootstrapped 95 % confidence intervals of $[.402, .486]$, $[.235, .328]$, and $[.671, .726]$ for RECIST-based, and $[.398, .482]$, $[.220, .323]$, and $[.652, .714]$ with volume-based prediction for $F1$, MCC and AUC, respectively, with all $p < .001$ (two-tailed t-test).

When having a look on the deep classifiers, with an F1 of .596, CI 95 % $[.450, .726]$, an MCC of .520 $[.356, .694]$, and an AUC of .814 $[.721, .896]$, the two-timepoint classifier clearly outperformed the RECIST diameter- and volume-based approaches. The single-timepoint classifier performed slightly worse, but still yielded superior results with .581 $[.455, .693]$, .423 $[.283, .540]$, and .787 $[.694, .865]$ for $F1$, MCC and AUC, respectively. The receiver operating characteristics (ROC) on the test data for the one- and two-timepoint classifiers are depicted in Fig. 5.5.

5.4 Discussion

As demonstrated by these results, both classifiers performed superior to, or on par with, the RECIST-based assessment, implying that lesion images may contain relevant visual information for the prediction of future tumor growth. Further, the results clearly demonstrate that structural image information might have additional value for estimating tumor treatment response over pure diameter- or volume-based assessment.

[12] The metrics are described in more detail in the Appendix in Tab. C.1.

[13] In the original study, also true positive and true negative rate, as well as positive and negative predictive value have been reported. However, due to the label imbalance, these metrics have only limited value for the concrete scenario and are omitted here.

[14] For a more detailed discussion on the bootstrapping methodology please refer to Chapter 9.1.

	Classifier (BL+FU)	Classifier (single)	RECIST	Volume
F_1	**.596**	.581	.444	.440
IFD	.436	**.486**	.250	.267
MKD	**.622**	.368	.224	.212
MCC	**.520**	.423	.282	.271
AUC	**.814**	.787	.698	.683

Table 5.2: Performance of the proposed approaches on the mCRC dataset in direct comparison to the predictors based on the radiological measurements. The highest result for each metric is marked bold. Source: [Katzmann et al. 2018b] © 2018 IEEE

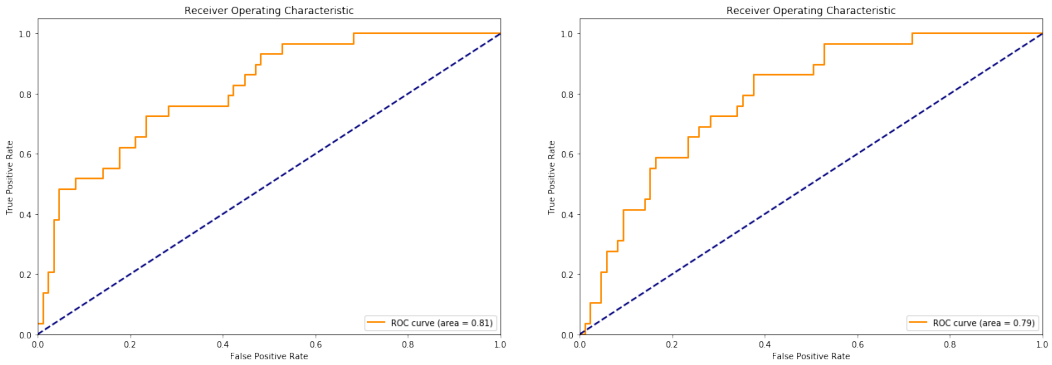


Figure 5.5: Receiver operating characteristics on the test set using the sparse autoencoder-based predictor with baseline and followup (left) and single-timepoint (right). Source: [Katzmann et al. 2018b] © 2018 IEEE

The study, however, also has multiple limitations which have to be noted at this point. First, this includes the limited amount of data. While the dataset in total consisted of 1,344 samples, these included only 460 different liver lesions stemming from only 135 patients. As a result, the numbers of independent training and test set samples were rather low. While for the baseline approach a thorough nested cross validation was conducted, the neural networks were trained using a fixed train/val/test split due to memory and time consumption, and thus should be repeated on larger datasets[15]. The need for this procedure, in particular, becomes clear when having a look at the size of the confidence intervals, spanning up to 35 % for MCC, and thus strongly relativizing the observed effects, in particular when taking into account the single-timepoint-based classifier results[16]. Further, the used dataset may have limited representativeness, as it was acquired by only one clinic, and amongst the acquired data only high-quality, thin-slice images have been included. In clinical practice, however, data may have lower quality, depending on the available scanner platforms and similar restrictions. Regarding the training procedure, it should be noted that with some of the meanwhile published works, such as BYOL [Grill et al. 2020] and PAWS [Assran et al. 2021], effective alternatives for classifier pretraining were proposed, whose value for the medical imaging domain should be analyzed in further studies.

Future work should make use of the available clinical and demographic data, such as blood values, age, and similar. Regarding the network architecture, it would be interesting to see the results of a fully-volumetric assessment. While not feasible within this study due to the amount of available data, a 3D assessment could include relevant additional information. However, as pointed out in [Katzmann et al. 2018b], as the extraction of 2D-slices is rather simple and segmentation can easily be done, a 2D-approach might well constitute an advantage if no fully-automated volumetric assessment is available, as it is only marginally more time-consuming than RECIST, but might yield significantly better performance in comparison to pure radiological assessment.

5.5 Conclusion

Treatment response prediction is an important part of *early assessment*, being a requirement for fast therapy adaption, which reduces costs and results in a patient-tailored treatment. Therefore, opening up an additional source of information by taking into account structural image information might be of considerable clinical value with respect to the highly important field of *precision medicine*[17]. While tumor size is known to be a relevant predictor of patient survival, having a diagnostic utility that allows for a prediction of tumor growth might in turn allow for an estimate of patient survival. However, patient survival is influenced by multiple clinical factors, too, such as age, fatigue, overall patient condition, comorbidities, etc. Thus, the direct link between both has yet to be established.

[15] This was done in later studies, and will be discussed in the following chapters.

[16] While with slightly lower performance, later studies have similarly demonstrated that single timepoints may allow for a proper estimate of future progression, cf. Chapter 8.

[17] Precision medicine denotes a medical treatment paradigm, aiming at a case-specific disease treatment, instead of a unified therapy, based on an in-depth anamnesis as well as thorough diagnostic testing.

6. One-Year Survival Estimation

As demonstrated in Chapter 5, structural image information can be used to predict tumor treatment response and might allow giving an estimate of the probability of future liver lesion growth. While treatment response in terms of lesion growth is an interesting outcome variable in itself, this chapter analyzes whether the structural information can further be used for predicting patient one-year survival, assuming that the visual lesion phenotype contains information on the underlying tumor genotype[1], and thus is associated with a specific treatment outcome. In contrast to lesion growth, the one-year survival chance immediately quantifies the impact of a tumor disease on the estimated lifetime and, thus, constitutes a complementary source of information for treatment planning, which may be decisive for the therapy choice or even category, i. e. curative, (neo-)adjuvant or palliative[2]. Therefore, its estimation is highly important for life quality and expectancy and is demonstrated in the following chapter based on [Katzmann et al. 2018a], presented at the International Conference on Medical Imaging with Deep Learning (MIDL) 2018.

6.1 Introduction

As was demonstrated in Chapter 5, image information can be used for predicting the treatment response, as measured by lesion growth. However, clinically it could be even more important to directly assess the expected effects of the disease on the patient survival time, rather than quantifying the growth or shrinkage of single lesions. In fact, depending on

[1] Tumor phenotype denotes the visual appearance of the tumor. In contrast, tumor genotype denotes the genome of the tumor, typically being analyzed as part of the tumor treatment process, as specific gene expressions have been shown to correlate with growth patterns.

[2] Curative therapy includes forms of therapy which aim to cure an underlying disease, such as surgical resection. Adjuvant therapy includes forms of systemic treatment which reduce the risk of progression or recurrence. Neoadjuvant therapy is conducted before the main treatment, e. g. systemic therapy to reduce tumor size before surgery. Palliative care, in contrast, aims to improve life quality, prolong life expectancy, and reduce suffering, but accepts death as the treatment outcome.

its localization and the patient's overall status, a singular lesion's growth is not necessarily correlated with a lower overall survival rate, while for the underlying tumor genotype a number of very marked correlations could be verified [Popat et al. 2005; Stoehlmacher et al. 2002; Teng et al. 2012]. The effect can be observed when taking into account specific treatments, such as immuno-oncological therapies, which are linked to a pattern called pseudoprogression and describes a process marked by the temporary growth of a lesion under treatment response [Chiou et al. 2015]. As pointed out in [Katzmann et al. 2018a] "[...] Oxnard et al. [2012] conclude that current criteria for progression may not adequately capture disease biology. Thus, having a high precision early lesion estimate on future growth would be of high clinical value, allowing to prematurely double-check, and thus potentially even prepone treatment decisions.".

In alignment with other work, such as [Aerts et al. 2014], within this work, it is assumed that the tumor phenotype contains additional information on the underlying tumor genotype, and therefore might be suitable for the identification of high-risk patients, associated with lower overall survival rates. Currently, estimating patient survival requires comprehensive clinical expertise and includes a high level of uncertainty. Therefore, the herein proposed method might be a relevant step towards a quantitative, better-substantiated estimate. While there has been recent work on patient survival prediction using deep learning, such as the work from Nie et al. [2016] and Yao et al. [2016, 2017], they were either trained on different modalities [Nie et al. 2016], or used histological data [Yao et al. 2016, 2017], and often employed a combination of deep learning and classical approaches. None of them has yet covered the assessment of mCRC lesions.

In the following sections thus a novel approach for CT liver lesion assessment is presented, and compared to the prediction using RECIST lesion diameter-based, as well as classical radiomics-based prediction. As in the prior study, liver lesion growth will be predicted, further amending the dataset by linking it to the respective clinical survival data. As a result, this study allows for an assessment of the performance of deep neural networks not only for tumor growth but also for patient one-year survival prediction by using a unified framework, demonstrating its superiority over other methods, such as radiomics- or RECIST lesion diameter-based prediction[3].

6.2 Methods

As pointed out earlier, both analyzed tasks are not yet near use in the clinical routine (see Chapter 2). Currently, only a few, non-applicable approaches in this direction exist (see Sec. 6.1). Thus, analogously to Chapter 5, a RECIST-based classifier will be trained. As no superiority of volume-based prediction could be shown in the previous study, it will be omitted herein. Further, a radiomics-based approach was added as proposed by Aerts et al. [2014] to better reflect the capabilities of state-of-the-art image analysis methods.

6.2.1 Preprocessing

The dataset in [Katzmann et al. 2018a] was first unified to reduce the data heterogeneity. Therefore, isotropic resampling was applied using cubic interpolation to ensure a homo-

[3]The original study also covered the use of saliency maps for the visual identification of relevant tumor growth patterns, potentially allowing for new insights in radiological image assessment. To the end of better consistency, however, decision explanation will instead be discussed in more detail in Chapter 10.

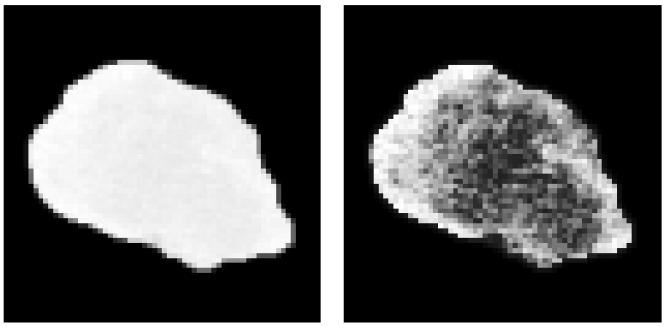


Figure 6.1: Example lesion (left) with applied histogram equalization (right). Clearly the image contrast is strongly enhanced, allowing the network to easier identify important structures while reducing data heterogeneity. Source: [Katzmann et al. 2018a].

geneous voxel size. This is especially important as the underlying hypothesis assumes specific structures to be predictive of specific outcomes. However, as CNN filter sizes are static, they always represent same-sized structures in the input space, and additional filters would have to be learned to represent scale invariance. Hence, heterogeneous voxel sizes are detrimental to efficient learning if only a few data are available. To this end, the same preprocessing as in [Katzmann et al. 2018b] was applied, i. e. extracting windows of 80 x 80 mm and segmenting the lesion to reduce the heterogeneity induced by the surrounding tissue, and final histogram equalization. An exemplary sample is depicted in Fig. 6.1.

6.2.2 Baseline Classifier Design

The baseline classification was conducted using a standard radiomics pipeline (cf. Chapter 3), consisting of feature z-normalization, ANOVA k-best feature selection and subsequent classification through a random forest. All hyperparameters have been extensively optimized by using 10,000 iterations of a randomized search cross validation on the training data. The pipeline was finally fit using the optimal parameter set and evaluated on the yet unseen testing data. This pipeline design allows for a wide variety of feature definitions, and is equally applicable to the RECIST- as well as the radiomics-based approach.

For the RECIST-based classification, the feature definition from Eq. 5.2 was re-used. The radiomics-based classification was constructed analogously to [Aerts et al. 2014] and employed the publicly available PyRadiomics reference implementation from [Griethuysen et al. 2017]. The radiomics features were extracted from the fully-volumetrically segmented lesions using the datasets A.2 and C.2 for tumor growth prediction and survival estimation, respectively. For each volume, the library extracted 1,209 features. Using the definition for delta-features from Eq. 5.2, this leads to a total of 4,836 features per lesion for each longitudinal pairing.

6.2.3 Sparse Characterization

Again, parameter amount stays an important issue. Thus, a first step was to drastically reduce the image dimensionality by employing a sparse convolutional neural autoencoder [Hinton et al. 2006; Krizhevsky et al. 2011, 2012; Ng et al. 2011] as was already done in Chapter 5. As the first timepoint of each pairing within this study always was a baseline image before therapy, the images are expected to differ significantly. Therefore, a two-lane

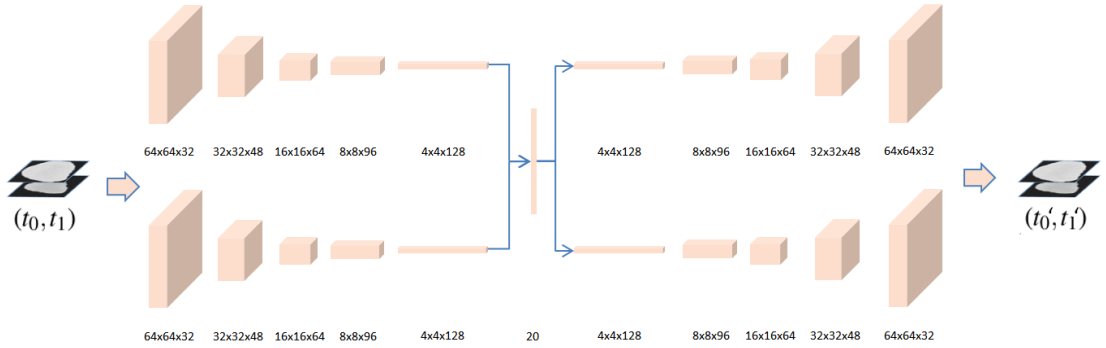


Figure 6.2: Architecture of the autoencoder network used in [Katzmann et al. 2018a]. Further details can be found in the Appendix in Tab. B.3.

design has been used, encoding the first and the second time point separately, but sharing a common sparse representation. Again a 2D representation has been chosen. Batch normalization [Ioffe et al. 2015] was applied after each convolution to accelerate the training process and preserve a better generalization performance. A structural graphic of the autoencoder can be found in Fig. 6.2. The complete architecture is depicted in the Appendix in Tabs. B.3–B.4. After autoencoder pretraining, analogously to [Katzmann et al. 2018b], the network was cut off after the sparse representation layer and a simple neural network was appended. As in the previous work, a 2-layer network with 8 and 2 neurons, introducing 218 parameters has been used, as it demonstrated to be reasonable for both outcome prediction tasks while providing a good generalization performance.

6.3 Experiments

6.3.1 Dataset

Within this study, two datasets have been merged to receive a dataset with both, (A) radiological images, as well as (B) clinical outcome parameters. As clinical data was not available for all radiological images and vice versa, the merged datasets (C/C.1/C.2) were significantly smaller than the original datasets. Analogously to the approach from [Katzmann et al. 2018b], two successive time points were used to predict the outcome variable. In contrast to [Katzmann et al. 2018b], only baseline scans, i. e. timepoint t_0, have been used as the first samples of each longitudinal pairing, significantly reducing the heterogeneity which results from timepoints t_i being either before or within therapy, and thus allowing for better reliability. To further reduce variance, only scans with a slice thickness of $\leq$3 mm have been used. A detailed overview of the employed datasets is given in Tab. 6.1.

The labels for tumor growth were assigned analogously to Chapter 5. One-year survival labels were extracted from the clinical data relative to the radiological scan date. Due to the one-year survival rate for colorectal cancer being at around 75% [Joachim et al. 2019], the resulting label distribution was rather imbalanced. An overview of the label distributions of the final datasets can be seen in Tab. 6.2.

Dataset	Description	N	$N_{patients}$	timepoints	lesions
A	Radiological Data (t_i)	1235	116	315	458
A.1	Radiological Data (t_0, t_1)	777	94	198	360
A.2	Radiological Data (t_0, t_1, t_2)	417	55	104	218
B	Clinical Data	135	135	-	-
C	Combined (t_i)	800	78	211	304
C.1	Combined (t_0, t_1)	496	61	132	231
C.2	Combined (t_0, t_1, t_2)	302	33	78	131

Table 6.1: Overview on the used datasets. Tumor growth prediction was conducted on dataset A.2, overall survival prediction on C.2. Pairing the radiological with the clinical data notably reduces the amount of available samples. Source: [Katzmann et al. 2018a]

Dataset	N	Positive	Negative
Tumor Growth	417	63	354
One-Year Survival	302	124	178

Table 6.2: Label distribution for tumor growth and one-year survival prediction datasets. Source: [Katzmann et al. 2018a]

Network Training

The network was trained using the Adam-optimizer with Nesterov momentum [Dozat 2016] using a batch size of 128 samples. For combatting the label imbalance, the importance sampling from [Katzmann et al. 2018b] has been used. The network was trained using a 4-fold cross validation grouped by the patient ID to ensure that all samples of a single patient are either contained in the train or test set. For each fold, a validation set containing a third of the training data was randomly split off.

6.3.2 Results

All classifiers were evaluated using a variety of informed and non-informed metrics[4]. The results for tumor growth and one-year survival prediction are depicted in Tabs. 6.3 and 6.4. Rows with a significant superiority of the deep learning-based classifier with respect to the best available reference classifier are separately highlighted (*; $p < 0.05$; two-tailed z-test). Empirical confidence intervals were computed as proposed by Efron [1987] using 10,000 iterations of bootstrapping[5]. The receiver operating characteristics for growth and survival prediction are visualized in Fig. 6.3.

As seen in Tab. 6.3, the proposed approach significantly outperformed the other tested approaches in the tumor growth prediction task for TPR, NPV, F1, IFD, and MCC with $p < .05$, and did further provide the best AUC (.784, $CI_{95} = [.735, .833]$, n.s.). While the radiomics-based prediction achieved a significantly higher TNR, this seems to be a result of the class weighting, as implied by the other metrics. Differences in PPV, MKD, and

[4]The differentiation of informed and non-informed metrics stems from [Powers 2011]. A detailed description of each used metric can be found in the Appendix C.

[5]The bootstrapping methodology will be discussed in more detail in Chapter 9.1.

	DL		RECIST		Radiomics		sig.
	μ	**CI 95**	**μ**	**CI 95**	**μ**	**CI 95**	
TPR	**.743**	[.648,.831]	.430	[.302,.552]	.363	[.245,.483]	*
TNR	.768	[.730,.806]	.864	[.827,.901]	**.912**	[.881,.940]	
PPV	.366	[.292,.439]	.359	[.250,.468]	.425	[.292,.560]	
NPV	**.944**	[.918,.965]	.894	[.861,.926]	.890	[.858,.922]	*
F1	**.490**	[.412,.561]	.390	[.280,.490]	.390	[.268,.505]	*
IFD	**.511**	[.405,.606]	.296	[.167,.425]	.277	[.153,.404]	*
MKD	.311	[.239,.390]	.258	[.148,.383]	.318	[.183,.452]	
MCC	**.400**	[.314,.480]	.273	[.159,.400]	.294	[.166,.420]	*
AUC	.784	[.735,.833]	.744	[.674,.810]	.737	[.669,.803]	

Table 6.3: Results on tumor growth prediction using the deep learning (DL) approach as proposed in [Katzmann et al. 2018a] in comparison to RECIST diameter- and radiomics-based prediction. Results are highlighted in bold if they are significantly better than both other approaches. Significant superiority of the DL approach is depicted as * ($p < .05$, two-tailed z-test). Source: [Katzmann et al. 2018a]

	DL		RECIST		Radiomics		sig.
	μ	**CI 95**	**μ**	**CI 95**	**μ**	**CI 95**	
TPR	.462	[.368,.547]	**.717**	[.593,.638]	.566	[.482,.648]	
TNR	**.927**	[.882,.963]	.612	[.538,.683]	.620	[.550,.689]	*
PPV	**.815**	[.721,.902]	.561	[.482,.646]	.507	[.425,.592]	*
NPV	.712	[.655,.768]	.756	[.684,.830]	.670	[.599,.739]	
F1	.586	[.497,.667]	.630	[.557,.695]	.534	[.459,.608]	
IFD	.387	[.290,.484]	.332	[.225,.434]	.182	[.072,.288]	
MKD	**.528**	[.419,.634]	.321	[.218,.426]	.178	[.069,.277]	*
MCC	**.449**	[.344,.541]	.321	[.219,.423]	.180	[.063,.288]	*
AUC	.710	[.645,.773]	.688	[.629,.740]	.568	[.504,.628]	

Table 6.4: Results on one-year patient survival prediction using the deep learning (DL) approach as proposed in [Katzmann et al. 2018a] in comparison to RECIST diameter- and radiomics-based prediction. Results are highlighted in bold if they are significantly better than both other approaches. Significant superiority of the DL approach is depicted as * ($p < .05$, two-tailed z-test). Source: [Katzmann et al. 2018a]

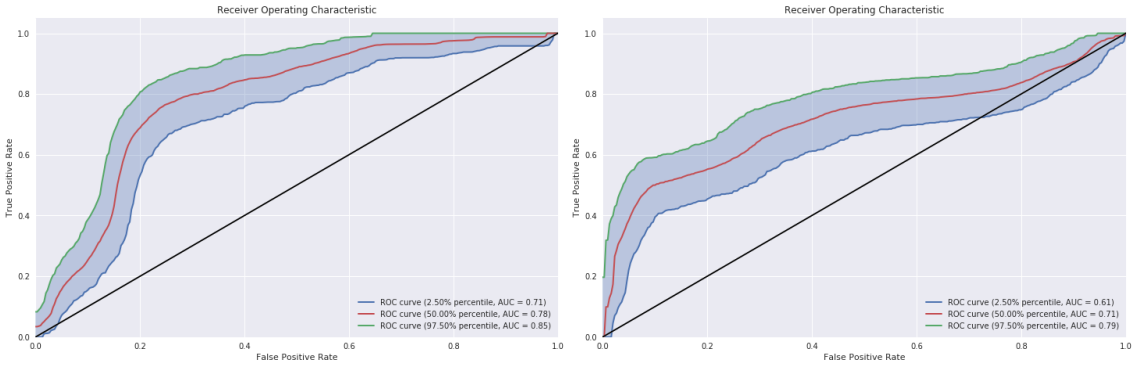


Figure 6.3: Receiver Operating Characteristic for tumor growth (left) and survival prediction (right) with the proposed deep learning approach. The shapes imply that both tasks contain samples of strongly different difficulties.

AUC were non-significant. Both, RECIST and Radiomics achieved comparable MCC and AUC values with .273/.294 for MCC and .744/.737 for AUC, respectively.

Similar results could be observed for the one-year survival prediction task, with significant superiority of the deep learning-based approach for TNR, PPV, MKD, and MCC, each with $p < .05$. Further, the method again provided the highest AUC ($AUC = .710, CI_{95} = [.645, .773]$, n. s.), and had a significantly higher informedness than the radiomics-based approach. Both, DL- and RECIST-based predictions clearly outperformed the Radiomics-based approach in terms of IFD, MKD, MCC, and AUC with $p < .05$.

6.4 Discussion

As demonstrated by the results in Sec. 6.3.2, the proposed method performed clearly best for both tumor growth and one-year survival prediction. While significant superiority could be shown for both target variables in terms of MCC, it was not possible to demonstrate a significantly higher AUC, which might be caused by the low amount of available data. However, the significance could be shown for a variety of metrics on both datasets.

Regarding the radiomics approach, it achieved a reasonable performance for lesion growth prediction, being on par with RECIST diameter-based performance, but was clearly inferior to RECIST- or DL-based prediction for one-year survival prediction. Although shown to be beneficial for a variety of tasks (cf. [Aerts et al. 2014; Huang et al. 2016; Li et al. 2016]), the results were rather unsatisfactorily. However, this might have been a result of the limited amount of training data, too, as the number of features M was significantly higher than the number of available samples N ($M >> N$, cf. multicollinearity), making a successful feature selection significantly more difficult. Generally, the radiomics approach would be expected to perform at least on par with RECIST-based assessment, as, as pointed out in [Katzmann et al. 2018a], the 2D lesion diameter, being equal to the RECIST diameter, is part of the feature set, and has, in fact, had the highest feature importance in each of the trained RF classifiers. Radiomics, however, was also shown to be highly influenced by different scanning protocols, across vendors, and even across models (see Chapter 3, cf. [Leijenaar et al. 2013; Mackin et al. 2015; Mühlberg et al. 2020]).

The study at hand has some limitations which should be pointed out. First, it suffers

from a high amount of heterogeneity, including different scanners and scan protocols, as well as the different therapies, e. g. 5FU, FOLFOX, XELOX, and various surgical interventions, possibly interfering with both feature-based as well as deep learning-based approaches. Repeating the study on a more homogeneous and/or significantly larger dataset therefore might go hand in hand with major improvements in the quality of estimates.

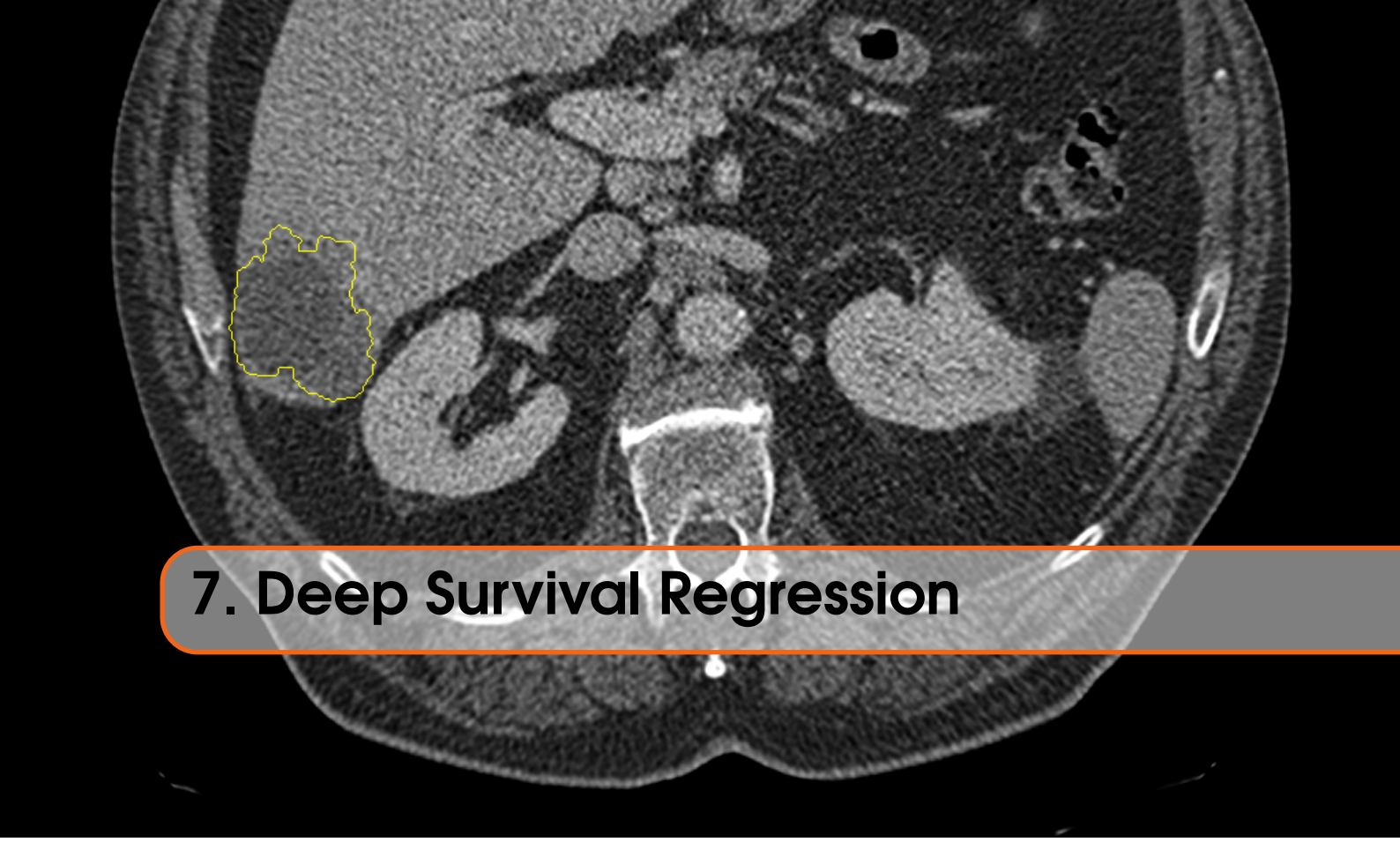
A more general problem of the application of deep neural networks for medical image classification lies in the low transparency of the decision-making process, being an issue which is covered in more detail in Chapter 10.

It can be noticed that the achieved performance is slightly lower than in Chapter 5. This results from multiple factors: First, in this study, it was decided that only baseline scans and their successor serve as a prediction base. This is relevant as it a) reduces the amount of available data, and b) might come with a worse baseline quality, as the follow-up data stems from specialized clinics and thus generally has a higher quality than the baseline data, which often stems from resident radiologists. Secondly, the limitation of maximum slice thickness has been added, also reducing the amount of available data. Both decisions, however, come with a significantly improved reliability of the results, as they enforce a clearer environment in which the algorithm might be applied. Additionally, the first study was conducted using a single test set, while the results in this Chapter are based on cross-validation, and are thus clearly more representative. The shortcomings of a slightly lower performance are therefore acceptable, as they come with significantly higher reliability.

6.5 Conclusion

Within this chapter, a method has been proposed which has demonstrated that not only the growth pattern of liver lesions can be assessed using structural image information, but that furthermore a *direct link* to the overall patient survival can be established. It is well known that lesion growth goes hand in hand with effects such as organ compression and therefore on average leads to a reduced patient survival (cf. Chapter 2). However, the results demonstrated that additional criteria over actual growth are involved, too, as pure size-based measures such as those included in the radiomics feature set or the RECIST diameter-based assessment are clearly outperformed by the deep learning-based approach.

As concluded in [Katzmann et al. 2018a]: “The results for our proposed deep learning approach imply that the radiological tumor phenotype itself encodes information beneficial for predicting tumor progress as well as the patient’s final outcome. [...] It is rather of clinical value to understand *which* structural specifics actually are predictive for tumor growth or the final outcome, and how to practically acquire and interpret these values within the clinical workflow.” Especially when combined with recent work in the field of decision explanation (cf. Chapter 10), the proposed work could clearly contribute towards an automated deduction of visually identifiable biomarkers, e. g. in an educational setting, and might allow practitioners to apply a patient-centric, specifically tailored medical treatment. It thus well aligns with the paradigm of precision medicine and could help to improve oncological healthcare.



7. Deep Survival Regression

In the last chapters, we have seen that structural image information can be used to predict lesion growth and one-year survival for liver metastases in colorectal cancer patients. While such a prediction might allow for better treatment planning and an earlier therapy adjustment, especially the granularity of the yet proposed method has room for improvement. Although being used in a variety of other studies (cf. [Mühlberg et al. 2020, 2021a], the threshold of one-year survival in fact is mostly arbitrary and introduces an artificial binarization of the continuous problem. Even if assuming a perfect one-year survival classification, it would not be possible to differentiate between patient survival times of only a few days or multiple months. Thus, for fine-granular risk stratification, a continuous approach is needed.

When using deep networks, the architecture could be modified in such a way that it represents continuous values, e. g. by using a linear output neuron or some arbitrary kind of fuzzification, and combining it with a regression loss function, e. g. mean squared error (MSE). If used on small data, deep regressors tend to show poor performance if not additionally constrained, choosing one of the two trivial loss-minimizing solutions to either a) perfectly predict each training sample without generalization (low bias/high variance), or b) to significantly over-smoothen the outputs, mostly predicting the distribution's mean (high bias/high variance).

In classical machine learning, typically, parameterized quantitative models are used for survival estimation, most notably the so called *Cox proportional hazards model* (**CPH**, [Cox 1972]) for predicting patient survival expectations based on time-dependent hazard probabilities. Until today this model can be seen as the state-of-the-art for survival estimation. The CPH models survival times as a result of hazard probabilities $h(t|X)$, consisting of a time-dependent, but covariate-independent *base hazard probability* h_0, which is multiplied by a time-independent, but covariate-dependent *exponential term* $\exp(\beta X)$:

$$h(t|X) = h_0(t) \cdot \exp(\beta_1 x_1, ..., \beta_n x_n) = h_0(t) \cdot \exp(\beta \cdot X) \tag{7.1}$$

for a patient at time t with features X and feature weights β. For fitting the model, this term is decomponized by first fitting the feature weights β, maximizing the rank concordance[1] of survival times, and afterwards fitting the base hazard probability h_0.

Although providing a practical method for risk stratification, due to this procedure the CPH does not always provide a good fit for absolute valued predictions. Some of the CPH assumptions, particularly the non-time-dependence of covariates [Desquilbet et al. 2005] and the multiplicativity of the hazard term [Aalen 1989], were shown to not be granted in the general case. Especially to address the latter issue, Aalen [1989] proposed a modified version of the approach by using an additive hazard term:

$$h(t|X) = h_0(t) + h_1(t)x_1 + ... + h_n(t)x_t \tag{7.2}$$

In contrast to the CPH, Aalen's approach took into account that covariates might constitute time-dependent hazard distributions. However, both the CPH as well as Aalen's model require hand-crafted feature design, and provide poor generalization performance if the number of features is high in comparison to the number of samples, being a result of multi-collinearity. To this end, in [Katzmann et al. 2019b] an image-based architecture for deep survival regression has been proposed, which particularly addresses the aforementioned issues[2].

7.1 Medical Background

When describing patient survival in a clinical environment, there are typically two main measures of interest, which are a) the overall patient survival (**OS**), describing the time span a patient survives, and b) the progression-free survival (**PFS**), describing the time-span until a significant disease progression can be observed [American Cancer Society 2017b; Cohen et al. 2008]. Depending on the used definition, both refer to the time span after the date of the first diagnosis (**DOFD**) or after the date of the first treatment (**DOFT**) [Cohen et al. 2008; NIH-NCI 2019].

While the PFS better describes the course of the disease and is therefore clinically more relevant, it typically suffers from a coarser granularity, as the monitoring needed for progression diagnosis is not frequent enough to allow for a determination to the day.

In contrast, the OS can typically be determined to the day due to the certifiable events which define it. OS, however, is less closely related to the actual disease than PFS, as various reasons other than described by the data can cause an event of death, amongst others, including comorbidities, medication, external factors, and non-disease-related organ failure, or simply dying of old age. Still, in the following the OS will be used, as only a few data were available and therefore no additional variance should be induced by using

[1] Concordance denotes the correctness of the order of the estimated survival times with respect to the ground truth. It is discussed in more detail in the Appendix C.

[2] An interesting alternative to the mentioned approaches are the so-called Random Survival Forests proposed by Ishwaran et. al [Ishwaran et al. 2008] working with handcrafted features, too. However, as this model is based on the fundamentally different concept of bootstrap aggregation, it will, later on, be covered in the metamethods part in Chapter 9.3.

coarse measures such as the PFS. For both, OS and PFS, usually, a cut-off value (often 5 years) is defined, which at best represents the time point at which the disease is likely not related to a potential event of death anymore. Often, however, practical considerations, such as the length of the observation period influence the choice of the cut-off value.

7.2 Related Work

Recent work has already addressed the use of deep neural networks for survival prediction, with the algorithms from Katzman et al. [2016] and Haarburger et al. [2018][3] providing well-founded deep learning-based variants of the CPH model. In contrast to the classical method, both approaches are usable with imaging data, and therefore do not require hand-crafted feature design. Both, however, are also built upon the CPH assumptions, and thus inherit their shortcomings regarding hazard probabilities (see above). As already pointed out in [Katzmann et al. 2019b], the definition of the hazard probability as an exponential term:

$$h(t|X) = h_0(t) \cdot \exp(\beta_1 x_1, ..., \beta_n x_n) = h_0(t) \cdot \exp(\beta \cdot X)$$

implies several limitations:

- The base hazard probability h_0 is shared across all individuals. Thus, it is independent of a-priori differences between subjects, e. g. different timepoints of therapy begin after the DOFD, etc.,
- The covariates $x_1, ..., x_n$ affect the base probability as additive exponential factors. Thus, they cannot have direct interactions,
- Covariates X are applied to the base hazard probability as time-independent, proportional exponential factors, meaning they affect the base hazard probability *equally at every timepoint*.

A similar algorithm from Lee et al. [2018] with a comparable intention is built upon a loss definition focussing on relative event-time order. While not inheriting the above-mentioned issues, it is therefore comparably prone to large errors when predicting survival times on an absolute time scale. In contrast, the following approach is explicitly modeled to take into account both, the event order as well as the absolute time scale, by directly predicting a patient-specific hazard ratio over time from image data. The predicted hazard ratio can subsequently be used to predict the expected OS, as will be demonstrated in Chapter 7.4.

7.3 Methods

The approach is motivated by the above-mentioned CPH. As pointed out in [Cox 1972], the CPH is fitted using a partial likelihood function $L_i(\beta)$ for individuals with index i, state vectors (i. e. features) z_i, and observation times o_i as:

[3] In their work, Haarburger et al. also proposed a method which is based on *over-median survival classification*, using output activations as an indicator of survival times. Using classificational output activations as regressional values, however, can be highly problematic, as will be discussed in more detail in Chapter 8. Although being an interesting contribution, the method was finally not able to outperform CoxPH-based prediction and will thus not further be taken into account within this chapter.

$$L_i(\beta) = \frac{h(o_i|z_i)}{\sum_{j:o_j \geq o_i} h(o_i|z_j)} = \frac{h_0(o_i)\exp(z_i\beta)}{\sum_{j:o_j \geq o_i} h_0(o_i)\exp(z_j)} = \frac{\exp(z_i\beta)}{\sum_{j:o_j \geq o_i} \exp(z_j\beta)} \tag{7.3}$$

Using this partial likelihood definition it is possible to derive feature weights β without taking into account the base hazard probability h_0. In a second step, h_0 can be estimated using standard optimization techniques to minimize the difference between observed and estimated survival times. In contrast to this, the approach is based on an end-to-end optimization process to estimate the patient-specific hazard function $h(t|z_i)$ for patient i, allowing for direct inference of features without hand-crafted feature design. No further assumptions are made regarding the distribution of the hazard function. Especially, the hazard proportionality assumption is dropped, as it is not capable of representing time-dependent, interacting, or non-proportional hazards which are often present in real-world data [Katzmann et al. 2019b; Schemper 1992].

7.3.1 Network Architecture

The framework does not depend on a specific architecture and can therefore be combined with various models like ConvNets [LeCun et al. 1989], ResNets [He et al. 2016] or even more complex architectures such as attention-gated networks [Schlemper et al. 2019]. However, as a necessary restriction the output layer has to consist of exactly K output neurons using sigmoidal activation, analogously to the work from Lee et. al [Lee et al. 2018], with each representing an equally sized discrete timestep. The number of outputs should be chosen on the basis of the observation period and the desired granularity. K is thus a metaparameter and specific to the target application scenario. Based on this, first the event observation times o_i with cut-off value v_{max}, $\forall i : o_i \leq v_{max}$ are normalized using

$$y_i = \frac{o_i}{v_{max}}(K-1) \tag{7.4}$$

Hazard probabilities at the discretized timepoint k are represented stationary, i. e. time-independent, assuming survival until k. It has to be noted that individuals might drop out of the data with no event observed for reasons such as movement, changes in the clinical institution, or simply due to the incompleteness of the available data. Therefore an observation variable $\omega_i \in \{0,1\}$ for patient i is introduced, indicating whether an event has ($\omega_i = 1$) or has not been observed ($\omega_i = 0$) within the observation period (so-called *right-censoring*).

7.3.2 Loss Definition

As pointed out in [Katzmann et al. 2019b], the model is trained using a combination of two loss functions L_1 and L_2. While L_1 minimizes the error between the estimated and the observed events of death, L_2 ensures that non-observed deaths can be reasonably explained with the estimated data distribution. Model parameters Θ are optimized according to

$$\arg\min_{\Theta} L_1^2 + L_2^2 \tag{7.5}$$

taking into account that stronger errors in each of the loss functions should be penalized stronger than smaller errors. First, the estimated complementary hazard function describing

the chance of *no* event at time k for patient i is defined as

$$\bar{h}(z_i) = 1 - \hat{h}(z_i) \tag{7.6}$$

based on the estimated hazard function $\hat{h}(z_i)$, i. e. the network output. The estimated *non-stationary* probability $\tilde{h}(k|z_i)$, i. e. the probability of observing an event *exactly* within timestep k (including the whole observation period until k) can now be derived as:

$$\tilde{h}(k|z_i) = \hat{h}(z_i) \prod_{j:0\leq j<k} \bar{h}(k|z_i) \tag{7.7}$$

Based on this, the expected survival $\hat{y}_i$ for patient i in case of an observed event ω_i can be derived as:

$$\hat{y}_i(z_i) = \frac{\sum_{k=0}^{K-1} k \cdot \tilde{h}(k|z_i)}{\sum_{k=0}^{K-1} \tilde{h}(k|z_i)} \tag{7.8}$$

Now, y_i and $\hat{y}_i$ are normalized by dividing by $K-1$, resulting in normalized values y_i^* and $\hat{y}_i^*$ which can be inserted into the binary cross entropy function multiplied by the observation indicator ω_i for deriving the expectation difference loss δ given by:

$$\delta(y_i^*, \hat{y}_i^* | \omega_i) = -\omega_i \left(y_i^* \log\left(\hat{y}_i^*\right) + (1 - y_i^*) \log\left(1 - \hat{y}_i^*\right)\right) \tag{7.9}$$

Concordance amongst samples is constrained by adding a concordance term $\mathscr{C}$:

$$\mathscr{C}(y_i^*, \hat{y}_i^* | \omega_i) = \sum_{i:0\leq i<n-1} \sum_{j:i\leq j<n} \left\{ \begin{array}{ll} (\omega_i \cdot \omega_j) \cdot \frac{max(\hat{y}_i^* - \hat{y}_j^*, 0)}{max(\hat{y}_i^*, \hat{y}_j^*)} & \text{if } y_i^* \leq y_j^* \\ 0 & \text{else} \end{array} \right\} \tag{7.10}$$

which penalizes non-concordant samples and finally leads to L_1 given by:

$$L_1 = \sum_{i:0\leq i<n} \delta(y_i^*, \hat{y}_i^* | \omega_i) + \mathscr{C}(y_i^*, \hat{y}_i^* | \omega_i) \tag{7.11}$$

For the definition of L_2, covering non-observed events, i. e. censored data, the estimation of observing an event within the observation period y_i of sample i can be treated as the loss itself, yielding:

$$L_2 = -(1 - \omega_i) \log(1 - \sum_{i:0\leq i\leq y_i} \tilde{h}(i|z_i)) \tag{7.12}$$

7.4 Experiments

In [Katzmann et al. 2019b], the model was evaluated using three different datasets, comprising:

1. The SEER Incidence database - an online database of cancer incidence data with more than 10 million cases in total. According to the focus of this work, the evaluation was conducted using the colorectal cancer subset, containing a total of 554,687 samples,

2. The Rossi Criminal Recidivism dataset - a dataset describing the criminal recidivism of individuals based on a small study population of 432 cases. The dataset is publicly available, often used as a benchmark dataset, and is employed as a proof-of-concept for small sample size applicability,
3. mCRC dataset - the dataset used in Chapters 5 and 6, including the patient metadata, such as patient demographics, histology and laboratory values (see below).

For the SEER incidence and the Criminal Recidivism datasets, the model was directly compared with a Cox proportional hazards model. The mCRC dataset, in turn, was used to assess the incremental value of using an image-based prediction in addition to the available clinical data. The CPH model was fit using the publicly available *lifelines* package [Davidson-Pilon 2019]. Each dataset was evaluated using the concordance index[4] (CI), the mean absolute error (MAE) and the median absolute error (MedAE). Confidence intervals were calculated using bootstrapping[5] until convergence with $\varepsilon < 10^{-3}$, as proposed in [Efron 1987].

7.4.1 Datasets

SEER incidence dataset

The Surveillance, Epidemiology, and End Results Program (SEER) is a study conducted by the U.S. National Cancer Institute of the National Institute of Health (NIH NCI), collecting cancer incidence data from population-based cancer registries, covering approximately 34.6 percent of the U.S. population. It covers patient demographics, e. g. gender and age, and disease data, such as primary location, morphology, and staging, and precisely determines the following patient survival. In total, it covers more than 10 million datasets with approximately 130 variables per patient [NIH-NCI 2017; Ries et al. 1999]. For matching the final application scenario, the colorectal cancer subset is used, containing 554,687 samples.

Notably, the SEER incidence dataset does not cover any imaging modality but is rather a collection of various demographic and disease-related information, i. e. consists of scalar data. While a variety of architectures can perform well in this scenario, it was decided to use an 8-layer feed-forward fully-connected network, adding residual connections to improve the gradient flow (cf. [He et al. 2016; Orhan et al. 2018]). The final input consisted of patient age, sex, year of birth, date of diagnosis, race, tumor laterality, and TNM-staging based on the TNM staging system [Brierley et al. 2016; Katzmann et al. 2019b] widely used in oncology.

The mean sojourn time[6] for colorectal cancer is approximated with 3 years [Zauber et al. 2012]. As 5-year survival is a widely employed measure in oncology, too, two experiments were conducted with 3-year and 5-year thresholds, respectively. In both cases, the method is compared to the Cox PH model given the same input.

Based on empirical analysis, the model output granularity was set to $k = 60$ neurons, as based on the loss formulation from Sec. 7.3.2, the overall performance is expected to only

[4] The formula for the concordance index as well as a discussion on its interpretation can be found in Appendix C.

[5] For a more detailed discussion on bootstrapping, please refer to Chapter 9.1

[6] The term “mean sojourn time” denotes the time span until a positive outcome can be expected, i. e. as other causes of death will be dominant again.

	CI	MAE	MedAE
Cox PH 3-y	**.689** [.689,.690]	11.8 [11.7,11.8]	9.00 [9.00,9.00]
Deep Model 3-y	.653 [.652, .654]	**8.49** [8.46,8.51]	**7.28** [7.25,7.31]
Cox PH 5-y	**.682** [.681,.684]	18.2 [18.1,18.3]	13.6 [13.0,14.0]
Deep Model 5-y	.660 [.659,.660]	**13.3** [13.3,13.4]	**11.8** [11.8,11.9]

Table 7.1: Results on SEER incidence dataset with 95% confidence intervals (errors in months)

slightly depend on k as long as it is chosen reasonably high to cover the desired granularity. k should generally be chosen in such a way that the expected prediction standard error is higher than the chosen output granularity so that no additional variance is introduced [7].

Rossi Criminal Recidivism dataset

As pointed out in Sec. 7.4, the criminal recidivism dataset describes the criminal recidivism of individuals based on a study population of 432 cases. As it is publicly available and often employed for event time prediction benchmarking (cf. [Fox et al. 2002]), it is used as a proof of concept for small sample size applicability. More than that, it is also a good example of sparse observations, as only 26.3% of the data have an observed event (i. e. criminal recidivism). Each sample has a total of 9 variables, including the observation interval, the event indicator (i. e. 0=no event; 1=event within the observation interval), as well as 7 additional covariates, including age, race, level of education, employment status, work experience, marriage status, and financial aid. All variables were z-normalized. Training and testing were done analogously to Sec. 7.4.1 and used the same network architecture.

mCRC dataset

Finally, the model was evaluated using the mCRC dataset C from Chapter 6[8]. As it was already demonstrated that the lesion phenotype can be predictive for patient survival [Aerts et al. 2014; Katzmann et al. 2018a,b], baseline/followup-pairs of single lesions have been used to predict the overall patient survival in days after the first diagnosis. An example of such a pair was depicted in Fig. 5.4. The lesions were extracted in accordance with the already presented work (see Chapters 5-6).

The used model was based on ResNet (cf. [He et al. 2016]). The clinical data, i. e. patient age, tumor grading, RECIST diameters at baseline and follow-up, and scan time after diagnosis were injected into the pre-output layer. Analogously the same data was used for training the CoxPH model, but, as already described above, lacked the imaging modality, for assessing its incremental value in this setup.

	CI	MAE	MedAE
Cox PH	**.594** [.539,.646]	28.4 [25.6,31.0]	29.1 [24.0,35.0]
Deep	**.587** [.530,.644]	**20.0** [17.6,22.4]	**17.6** [13.7,24.8]

Table 7.2: Rossi Criminal Recidivism dataset results with 95% confidence intervals (errors in weeks)

	CI	MAE	MedAE
Cox PH	**.536** [.508,.569]	423 [397,450]	384 [357,403]
Deep	**.510** [.481,.540]	**325** [301,349]	**262** [233,291]

Table 7.3: Results the mCRC dataset with 95% confidence intervals (errors in days)

7.4.2 Results

SEER incidence dataset

As depicted in Tab. 7.1, the use of the model for the SEER incidence dataset consistently reduced the absolute prediction errors while slightly decreasing concordance (3.6/2.2%) in comparison to the CoxPH model. The relative absolute error reduction was 13.2%, resp. 26.9% for 5-years, and 19.1%, resp. 28.1% for 3-years survival. All differences were highly significant with $p < .001$ (two-tailed t-test).

Rossi Criminal Recidivism dataset

The results for the criminal recidivism dataset are depicted in Tab. 7.2. Similar to the SEER incidence dataset, the mean and median absolute errors could again be significantly reduced at the cost of a small reduction of concordance. The relative improvements for MAE and MedAE were 29.6% and 39.5%. Both reductions were highly significant with $p < .001$, while the reduction of the CI was non-significant with $p = .849$ (two-tailed t-test).

mCRC dataset

Tab. 7.3 shows the results for the mCRC data. As for the previous datasets, mean and median absolute errors could be significantly reduced, indicating a superior fit of the overall distribution. However, both algorithms fail to reasonably reproduce the individual patient survival differences as shown by CIs of .536/.510 for CoxPH and deep survival prediction.

7.5 Discussion

With the method described in this chapter a framework for survival time prediction has been presented, fundamentally being inspired by the well-known proportional hazards model from Cox [1972]. In contrast to the CoxPH model, the presented framework is theoretically applicable to a wide range of applications without a need for prior data

[7] It should be noted at this point that a too high granularity might indeed lead to numerical instabilities, and might require the use of high-precision data types such as float64.

[8] As the study was conducted as part of an ongoing project, the dataset was meanwhile complemented by clinical data for two additional patients, so the final dataset consisted of data for 80 instead of 78 patients.

processing or hand-crafted feature engineering but instead infers meaningful features in a data-driven fashion.

Recent work such as the methods from Katzman et al. [2016] and Haarburger et al. [2018] have already demonstrated that deep learning can be applied for survival time prediction. However, both frameworks are based on the CoxPH model and thus inherit some of its limitations (cf. Sec. 7.2). In contrast, the herein presented method does not assume a particular overall or patient-specific hazard or survival time distribution, and thus provides greater flexibility for feature interaction.

Setting requirements only for output encoding and loss formulation, the method does not restrict the used network architecture, and thus can be easily applied to a wide variety of networks, such as ConvNets, ResNets [He et al. 2016], Attention-gated networks [Schlemper et al. 2019] or even graph-based neural networks [Scarselli et al. 2008] without requiring larger modifications to the model architecture. For both, the SEER incidence and the Rossi Criminal Recidivism dataset, the method was able to demonstrate a significantly reduced mean and median absolute error in direct comparison to the well-known Cox Proportional Hazards model at the cost of a slight drop in the concordance index.

Regarding the mCRC task, however, the applicability of the method remains unclear. As shown in the results section, neither the CoxPH nor the deep learning-based model were able to sufficiently well predict patient survival times. Given the low amount of patients ($N = 80$), this is expected to be a result of the high variance, a large number of parameters, and the various influences on patient survival, which may not be sufficiently well represented in the given data.

Further limitations of the study at hand comprise, that no systematic optimization of the meta parameter K has been conducted. While, as discussed in Sec. 7.3.1, the concrete choice might have only a minor influence, a systematic evaluation of the parameter in future studies would be desirable. Within this work, the applicability was mainly shown for ResNet-based architectures. Although due to its definition no major interactions with the above loss definition are expected, an evaluation of other architectures is desirable. While this study aimed to provide a proof-of-concept, most notably future studies should particularly set the method into direct comparison to previous work, such as the approaches from Katzman et al. [2016] and Haarburger et al. [2018].

7.6 Conclusion

With this contribution, a method for data-driven deep survival estimation was presented. Although superiority on the image-based task could not be shown, the method might be an interesting step toward a deep, assumption-free survival estimation, to the best of knowledge being the first to combine the benefits of deep learning and CoxPH regression without inheriting the above discussed methodological limitations, and being applicable to a wide variety of applications using a purely data-driven approach. Another promising approach with a similar aim, but based on the principles of random forest classification, will be thoroughly discussed and evaluated in Sec. 9.3.

IV Meta-Methods & Decision-Explanation

8 Deep Confidence Estimation ... 61

8.1 Introduction
8.2 Base approach
8.3 Methods
8.4 Experiments
8.5 Discussion
8.6 Conclusion

9 Bootstrapping Methods 73

9.1 Bootstrapping
9.2 Classifier Architectures
9.3 Deep Survival Forests
9.4 Conclusion

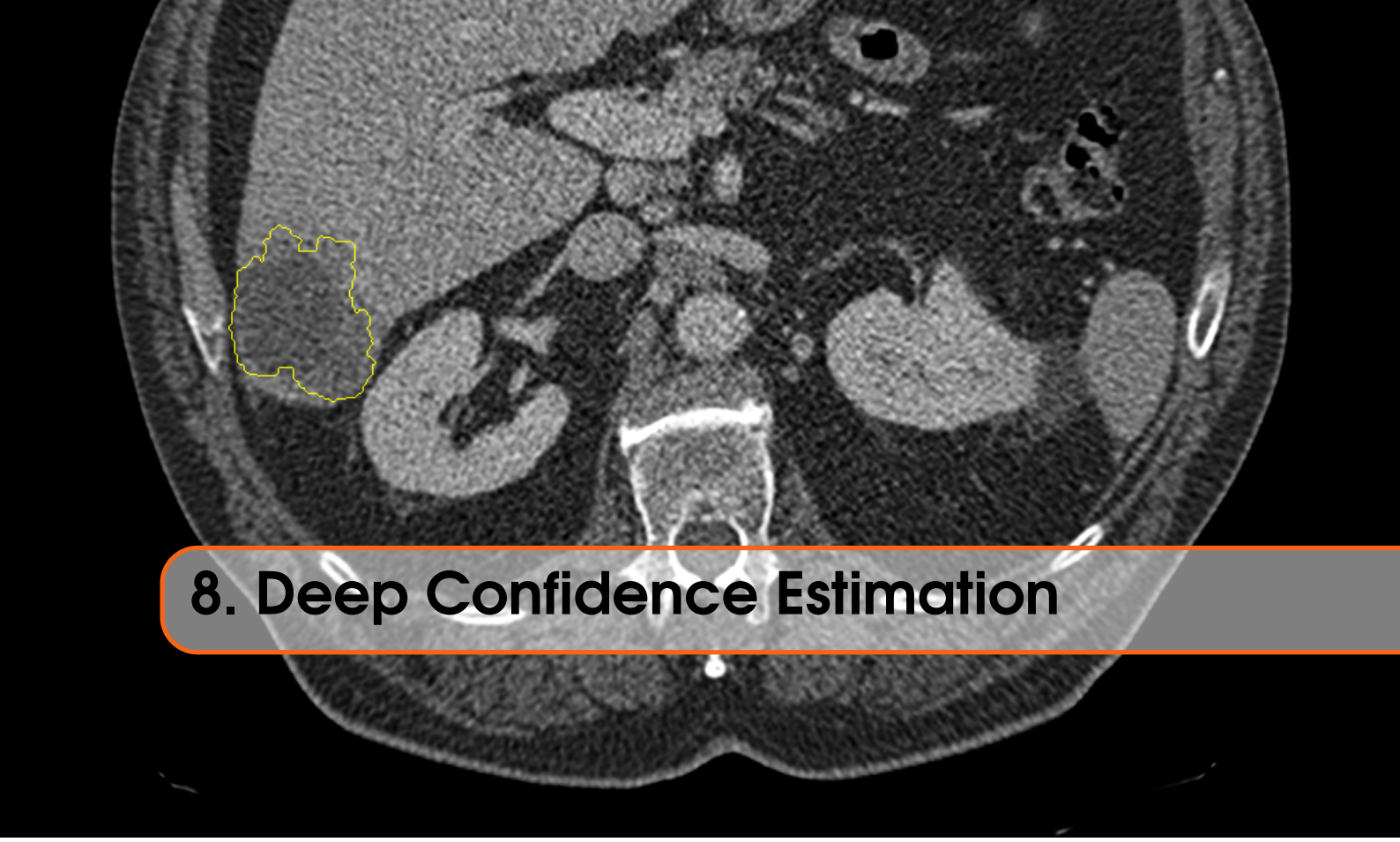
10 Deep Decision Explanation 93

10.1 Introduction
10.2 Deep Decision Explanation using Cycle-GANs
10.3 Experiments
10.4 Discussion
10.5 Conclusion

Within the last chapters a variety of algorithms for medical image classification have been discussed, including clinically highly-relevant tasks such as lesion growth (Chapter 5) and one-year survival prediction (Chapter 6), as well as hazard and survival regression (Chapter 7). In particular, we have seen that deep learning can successfully be applied to derive meaningful features from medical imaging data in a purely data-driven fashion, having the potential to lead to a significantly improved classificational performance. Specifically, it has been demonstrated, that lesion CT phenotype directly corresponds to a number of clinically highly relevant variables, such as future disease progression and overall patient survival, and moreover, this correspondence can be approximated using deep neural networks.

Although the presented approaches overall were highly promising, each of the previous methods to a certain degree has been specifically tailored in one or the other way to the characteristics of medical imaging data. Yet, no general framework has been proposed which is applicable to arbitrary architectures, while simultaneously solving the issues which arise from an application in the medical imaging domain, and which have been elaborated on multiple times within this work (cf. Chapter 4).

Addressing the shortcomings arising from an application of deep learning to medical image analysis, most notably the initially discussed issues of *unknown certainty* (Chapter 8), applicability in case of *small- or medium-sized* datasets (Chapter 9), and finally *deep decision explanation* (Chapter 10), the following chapters will present a number of novel, deep learning-based meta-architectures, significantly contributing towards a resolution of the aforementioned issues. Each of the following approaches is applicable to a wide variety of deep architectures, including, but not being limited to, the ones that have been discussed so far. Most notably, using multiple thorough evaluations and comparing to a variety of other approaches, it will be demonstrated that the use of the presented methods may yield superior performance in typical medical imaging scenarios without requiring additional steps, such as meta parameter optimization, and thus that they particularly can be used as an out-of-the-box toolset for medical image analysis.



8. Deep Confidence Estimation

In the previous chapters, multiple methods for estimating disease progression based on deep neural networks have been presented, which have demonstrated promising performance in a variety of applications. However, as exemplarily pointed out by Begoli et al. [2019], computer-assisted clinical decision support similarly requires novel, deep learning-tailored methods for uncertainty quantification and confidence estimation to fully leverage its potential. As they particularly emphasized, this becomes highly important due to the delegation of decisions to machines that are potentially life-threatening, going hand in hand with considerable psychological implications. Knowing the confidence of an automated medical assessment is crucial for any clinical application in order to avoid wrong assessment, mistreatment, or even death.

To better understand classifier decisions, a variety of decision explanation approaches has been proposed (cf. [Lundberg et al. 2017; Ribeiro et al. 2016; Simonyan et al. 2013]). However, these methods rely on visualization of decision-relevant regions in the input space and thus are typically better suited for single-case failure assessment than as a quantitative measure for failure identification. They will therefore be discussed in more detail in Chapter 10. Having a simple, quantitative measure of confidence, in contrast, can give the practitioner valuable information about *assessment reliability* without requiring a visual inspection of every single image. It, therefore, allows for a quick overview on whether additional cross-checks should be conducted *on top of* the usual workflow.

A naive, but often used approach for confidence estimation is realized by using the output distribution of a classifier as a measure of confidence itself, i. e. the *level of determination*, e. g. by taking the maximum output activation of a softmax classification output. In fact, higher output activations tend to be associated with higher probabilities of an accurate estimate [Hendrycks et al. 2016]. Still, however, the procedure has major shortcomings, as a) it does not allow for a quantification of certainty for out-of-distribution (OOD) samples[1],

[1] A borderline case for this are adversarial samples as proposed by Goodfellow et al. [2014], maximizing

and b) classification losses, such as categorical crossentropy, hinge loss or focal loss [Lin et al. 2017], are designed in such a way that already correct classifications contribute only marginally to the overall loss. As a result, estimated values are uncalibrated and concordance (cf. Chapter 7) amongst samples is not necessarily given[2].

In clinical practice, especially the first of the two factors becomes relevant, as in particular OOD samples, rather than the typical population, are problematic to assess for a model trained in a data-driven fashion. This, even more, becomes important as large and representative datasets are typically rare in the medical imaging domain, i. e. the known distribution is rather narrow, and thus the chances for a classification outside of the known distribution increases (cf. Chapter 2).

In their work, Begoli et al. [2019] explicitly point to the training of deep learning-based solutions to estimate their certainty themselves as being a highly effective way of addressing some of the mentioned uncertainty quantification problems. Notably, exactly such a method has been proposed by DeVries et al. [2018]. Within this chapter, an improved version of this algorithm will be discussed as presented at the International Symposium on Biomedical Imaging (ISBI) 2019 [Katzmann et al. 2019a]. As will be shown within the course of this part, it turns out that not only it is possible to estimate confidence directly, but also that this estimate can further be used to improve the training quality[3] in a psychologically motivated framework called **deep metamemory** by using confidence as a measure of difficulty in a curriculum learning-like approach as suggested by Bengio et al. [2009].

8.1 Introduction

The field of *uncertainty estimation* deals with the process of a substantiated estimate of the misclassification rate of a given estimator. Assuming a concrete classifier C, the confidence c_i for a sample i with $c_i \in \{0,1\} \in \mathbb{R}$ is the probability of the classifier giving a correct estimate with a known misclassification rate $m_i = (1 - c_i)$.

Within the last years, and as a result of the widespread use of deep learning-based techniques, uncertainty estimation for deep neural networks has recently been a highly active field of research. As discussed in [Katzmann et al. 2019a], much of this work is settled in the field of *variational inference*, including work on *Monte-Carlo Dropout* (**MCD**) [Gal et al. 2016], *Stochastic Batch Normalization* (**SBN**) [Atanov et al. 2018] and *model ensembling* [Lakshminarayanan et al. 2017]. The basic idea of variational inference is to use the model posterior distribution upon repeated modification of the network or input to estimate the parameters of a well-known probability distribution, e. g. normal or Bernoulli distribution, and to subsequently derive a statistically substantiated estimate of confidence. A downside of these approaches is their use of repeated inference, leading to higher computational complexity.

More specifically, SBN and MCD both use a very comparable mechanism, as they both rely on a stochastization of an originally deterministic network, leading to a variation in the

output activations by an introduction of small, additive OOD patterns.

[2] To refer back to Chapter 7, this is the main reason that the classification approach for survival time expectation from Haarburger et al. [2018], while being highly economical, does not necessarily provide a superior survival time regression.

[3] The idea of leveraging uncertainty for improved performance has later on been adopted by other authors, e. g. in [Ghesu et al. 2019]

output probability distributions. SBN does so by modifying the batch normalization layers [Ioffe et al. 2015] used throughout the network to *not learn* fixed values for mean and standard deviation, but rather to model mean and standard deviation as normal distributions themselves. As a result, the layer outputs are batch-normalized *differently in each iteration*, and thus the network output variation can be used to infer a probability distribution.

MCD works similarly but instead utilizes the dropout (rather than batch normalization) mechanism, which recently has often been used in deep neural networks, but is now more and more displaced by batch normalization. With MCD an additional dropout layer is introduced after each network layer but the output. The model is then trained as usual. Finally, in contrast to their usual behavior, the dropout layers stay active during inference, i. e. they randomly drop some of the activations. Just as with SBN, a repeated application of the network now allows for an estimate of the output distribution. As was shown by Gal et al. [2016], the repetition during inference can actually be omitted when trained using the dropout design, as the variation can then be directly approximated by using the learned weights. However, while this greatly reduces the *inference time* in comparison to a repeated application, it clearly increases the *training time*[4] and might even worsen the network performance, as it reduces the effective network capacity.

Lastly, confidence estimation through model ensembling as suggested by Lakshminarayanan et al. [2017] captivates by its impressive simplicity, training multiple estimators on randomized subsets of the data. In fact, this approach is directly related to the one presented in Chapter 9. While the authors can show that such an ensemble is effective and the results are comparable to MCD and SBN, the method comes at the cost of having a sufficiently high number of parallelized networks, and thus requires a significantly larger amount of system resources for training and inference.

Notably, the above-mentioned methods do not specifically take into account the difference between aleatoric and epistemic uncertainty. While epistemic uncertainty (Greek *episteme*: knowledge) describes the uncertainty which results from the limitedness of data, e. g. due to data amount or data quality, aleatoric uncertainty (Lat. *alea*: dice, chance) describes the uncertainty due to the specifically random nature of the data, as it may result from predicting (partially) random processes, such as quantum positions, with some recent research particularly accounting for a separate estimate of both. A good introduction to the topic is provided by Hüllermeier et al. [2021]. However, as within the given clinical scenario the source of inconfidence is not expected to be relevant for the treatment decision, the following analyses will not aim to explicitly differentiate between them.

Recent research has also taken into account *Bayesian Neural Networks* (**BNNs**) for uncertainty estimation (cf. [Kendall et al. 2017]). BNNs are modeled analogously to standard neural networks but realize weights as probability distributions rather than scalar values, e. g. as normal distributions $\mathcal{N}(\mu, \sigma^2)$, or by using Monte-Carlo approximations of more complex ones. As a consequence, BNNs natively support output variance. However, training and inference are highly costly, often making them computationally inefficient with respect to the final application.

In contrast to that, this work will aim for a *one-shot confidence estimation*, requiring neither relevant modifications of the network architecture nor significant additional resources, and thus being applicable to a wide variety of applications.

[4] While there are some approaches to greatly reduce the training time when using dropout (e. g. [Wang et al. 2013]), networks still train significantly faster without it.

8.1.1 Background

As pointed out in [Katzmann et al. 2019a], confidence estimation is basically a classifier's estimate of its own knowledge. In fact, this process is very comparable to the process of self-assessment in humans. Cognitive psychology deals with this issue under the term *metamemory* (i. e. knowledge about knowledge), being a subfield of metacognition (i. e. thinking about thinking) [Nelson 1990]. Although this type of meta-knowledge has been theorized about already by classical philosophers like Descartes, the first comprehensive theoretical concept of metamemory was proposed by Hart [1965][5], and later picked up under the name "metamemory" by influential psychologists like Tulving et al. [1970] and Flavell [1971]. Interestingly, humans seem to be able to reasonably well assess whether they can answer a question or not, although there might be large differences across individuals as pointed out by Kruger et al. [1999][6]. For example, humans are able to give a quite precise estimate of a problem's difficulty before they are given time to solve the problem itself, as shown by Kelemen [2000].

From a theoretical point of view, this is somewhat counter-intuitive, as such an assessment does not only require knowledge or non-knowledge about the actual answer but additionally knowledge about familiarity with the topic. The results, however, imply that the mechanisms for estimating confidence and the ones for giving a correct answer do not completely overlap. Instead, confidence estimation only requires a general estimate on:

1. whether the just faced data is sufficiently known, as it is similar to previous experiences (e. g. during training time), and
2. whether the subject can typically handle this type of data.

From an evolutionary point of view, confidence estimation in humans and other animals is of high importance for self-preservation. For low-confidence situations, a generally more reserved, cautious behavior might be advisable, while on the other hand collecting the possible benefit in certain or nearly-certain situations using an optimistic behavior might constitute a significant advantage. In contrast to this, a wrong decision in a dangerous situation might lead to significant damage or even death, in a Darwinian sense facilitating a natural selection of accurate confidence estimators. In the following sections, it will be shown how to induce such knowledge into a computational learning model as an inherent part of the training process, and subsequently use it for facilitating learning on medical imaging data with significantly improved test time performance.

8.2 Base approach

The basic confidence estimation framework stems from DeVries et al. [2018]. The approach is fundamentally based on the assumption that within training the network might learn a measure of confidence by allowing it to take hints, but has to quantify the measure of hints it receives. Assuming some arbitrary network architecture, this is done by introducing an additional output c which serves as a measure of confidence, as depicted in Fig. 8.1. As shown there, the basic idea is to calculate the network error not on the network output

[5] In his work, Hart used the term *feeling-of-knowing*

[6] The Dunning-Kruger effect describes the fact that high-performers typically underestimate their knowledge while low-performers typically overestimate it.

probability $\hat{y}$ itself, but rather to create an augmented output probability function $\hat{y}'$, utilizing the own confidence estimate c for taking exactly the amount of hints, i. e. ground truth y, that is necessary for minimizing the loss over the already existing estimate $\hat{y}$. Simultaneously, the network is trained to maximize its own confidence, i. e. to use as few hints as possible. The augmented output probability $\hat{y}'_i$ for a sample i is given by:

$$\hat{y}'_i = c_i \cdot \hat{y}_i + (1 - c_i) \cdot y_i \tag{8.1}$$

In their original work, DeVries et al. trained a classification network with m outputs using binary crossentropy, i. e. the prediction loss is calculated as:

$$\mathscr{L}_p = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m} \left(y_{i,m} \cdot \log(\hat{y}'_{i,m}) + (1 - y_{i,m}) \cdot \log(1 - \hat{y}'_{i,m}) \right) \tag{8.2}$$

However, the method does not depend on any specific loss definition and could be applied similarly using other classification or regression losses, such as the mean squared or mean absolute error. The only requirement is that the loss function is monotonously decreasing with respect to the result direction, i. e. $\mathscr{L}_p(\hat{y}'_i) \leq \mathscr{L}_p(\hat{y}_i)$. The confidence is maximized by introducing a binary crossentropy loss $\mathscr{L}_c$ against a steady target output of 1, reading:

$$\mathscr{L}_c = -\frac{1}{n} \sum_{i=1}^{n} \log(c_i) \cdot \lambda \tag{8.3}$$

The overall loss is formed as $\mathscr{L} = \mathscr{L}_p + \mathscr{L}_c$. Notably, Eq. 8.3 contains a scaling parameter λ. The reason for this is as follows:

Eq. 8.1 has two trivial solutions, being $c \to 0$ and $c \to 1$:

- The first case ($c \to 0$) results in $\hat{y}' = y$, meaning that the actual network prediction $\hat{y}$ is not taken into account at all, and thus, that in combination with Eq. 8.2 no loss is forwarded to the network for learning. Due to Eq. 8.3, however, this case will typically be avoided, as $\mathscr{L}_c \to \infty$ for $c \to 0$.

- The second case ($c \to 1$) results in $\mathscr{L}_c \to 0$, and thus in the optimal solution for the latter of the loss terms of $\mathscr{L}$. As a result, Eq. 8.1 takes the other trivial solution (i. e. $\hat{y}' = \hat{y}$). In fact, in this case, the network will learn the problem but cannot produce a confidence estimate. If now an additional parameter λ is introduced and dynamically adapted to satisfy $\mathscr{L}_c \approx \beta$ for any chosen $\beta > 0$, the second trivial solution is ruled out, too.

Within their work, DeVries and Taylor stated that the concrete choice of β is not too important for the learning process within a range of $[0.1, 1] \in \mathbb{R}$ [DeVries et al. 2018]. In a later presentation at the International Conference on Medical Imaging and Deep Learning (MIDL) 2018, however, they demonstrated that the sharpness of the confidence estimate can largely depend on the choice of β [Taylor 2018]. Therefore, this work aims to improve upon the original approach by estimating the optimal β as a component of the network and a result of the optimization process itself.

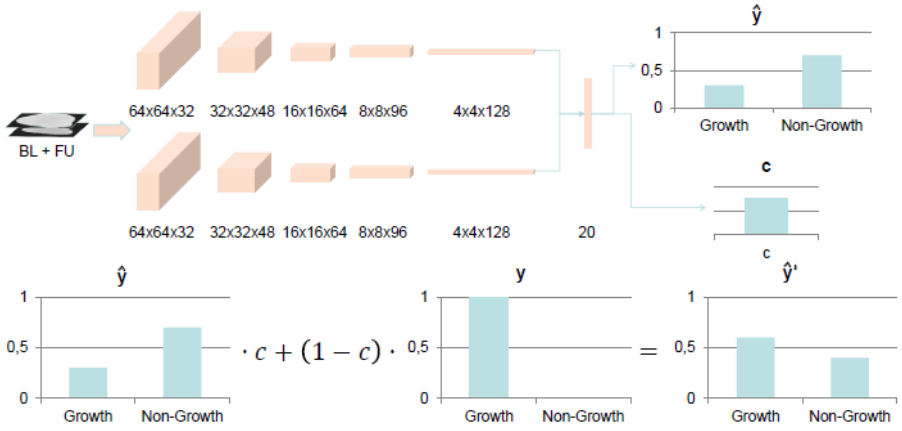


Figure 8.1: The basic principle of the confidence learning approach exemplarily demonstrated with the lesion growth prediction architecture from Chapter 6 (top-left). Confidence is modeled additionally to the output probability $\hat{y}$ as a separate one-neuron sigmoid output c (top-right). During training, the confidence c is used for weighting the network output $\hat{y}$ and the ground truth y to yield an augmented probability estimate $\hat{y}'$ which is subsequently used to calculate the training loss. Simultaneously c is tried to be maximized.

8.3 Methods

The above work shall now be utilized to a) give a confidence estimate for lesion growth prediction based on the approaches from Chapters 5 and 6, as well as the approach from DeVries et al. [2018], b) to further improve the confidence estimation by automatically find an optimal parameter β as a result of the training procedure which is evaluated against the original approach, and c) to improve the overall training by employing a curriculum learning-like approach, providing the network with the most valuable samples using an approach denoted as *metamemory importance sampling*.

8.3.1 Metamemory Importance Sampling

As pointed out in Sec. 8.1.1, humans are reasonably well able to assess whether their skill set is adequate to solve a specific task. Moreover, in fact, humans use active rehearsal if they are unconfident, i. e. explicitly repeating examples that they feel unconfident about (cf. [Waters 1982]). This procedure can easily be transferred into the deep learning setting and well aligns with the curriculum learning idea from Bengio et al. [2009]. Therefore, the following procedure is proposed:

1. Estimate confidences c_i for samples i,
2. Calculate confidence-based sampling probabilities π_i,
3. Train network for one epoch on training pairs (x_i, y_i), randomly sampled from (X, Y) according to sampling probabilities π_i.

For determining the sampling probabilities π_i, it is started with a naive approach, setting π_i to the inconfidence (i. e. $1 - c_i$) and subsequently normalizing across samples, yielding:

$$\pi_i = \frac{1 - c_i}{\sum_{k=0,..,|X|} (1 - c_i)} \tag{8.4}$$

As a side effect of this definition, samples the classifier assumes itself confident with – mistakenly or not – are not taken into further consideration, as their sampling probability approaches zero. If the confidence estimate is incorrect, however, this could reduce the effective training set size and even prohibit successful training. As reweighting is done only after full epochs, the confidence values computed at the beginning of an epoch can become less precise for later samples within the same epoch, thus a naive application of the above sampling probabilities might lead to loss oscillation. Similarly, continuously estimating confidence before every single batch is computationally inefficient.

Both these issues can be greatly reduced by transforming the resulting probability distribution $\Pi = \pi_0, ..., \pi_{n-1}$ into a reciprocal (i. e. log-uniform) distribution by using an inverse transform sampling $RS_{log_{eq}}(\pi_i)$ (e. g. described in [Devroye 1986]), and clipping the minimum and maximum sampling probabilities to an interval of $[0.1, 1]$, yielding modified sampling probabilities π'_i:

$$\pi'_i = RS_{log_{eq}} \left(\frac{1 - c_i}{\sum_{k=0,..,|X|} (1 - c_i)} \right) \quad : \quad 0.1 \leq \pi'_i \leq 1 \tag{8.5}$$

While the concrete interval borders did not seem to have too much influence, empirically the sampled probability distribution had. Logarithmic equal distribution clearly outperformed linear resampling, as it better reduced loss oscillation. Further, it was closer to the error distributions and overall performed more stable. As a thorough evaluation has not been conducted, however, even better-performing distributions might exist.

8.3.2 Model augmentation

As in the approach from DeVries et al. [2018], first, the network gets augmented by an additional confidence estimation lane having a single sigmoid output neuron yielding the confidence estimate, which is appended to the pre-output fully-connected layer (see Fig. 8.1).

In contrast to the original approach, the choice of β is integrated into the network. For this purpose, an additional variable $\hat{\beta}$ is introduced, being subject to the network weight optimization process, which is instead used for the confidence loss weighting from Eq. 8.3. $\hat{\beta}$ is adapted with respect to the confidence estimation error with an additional loss term using the mean squared error between the estimated and a target β parameter:

$$\mathscr{L}_\beta = ||\hat{y}_\beta - \beta||^2 \tag{8.6}$$

starting with a target value of $\beta = 1$ and adjusting it after each epoch according to:

$$\beta' = \mathscr{L}_p + \mathscr{L}_c \quad , \quad \beta \leftarrow \beta' \tag{8.7}$$

Notably, adjusting β to be correlated to the task loss adjusts the inter-class-margin width to be correlated to the actual classifier performance, better-aligning confidence estimates to classification accuracy (cf. [Katzmann et al. 2019a]).

8.4 Experiments

Three different datasets have been used for evaluation, namely the CIFAR-10 and CIFAR-100 datasets for demonstrating the applicability for large RGB image datasets with few and high numbers of classes, as well as the radiological mCRC data which already has been used in the previous chapters 5–7. The model is trained using Adam [Kingma et al. 2014] with an initial learning rate of $lr = 1 \cdot 10^{-3}$. lr is reduced by half at plateaus of at least 15 epochs with no improvement. The training was stopped after 35 epochs without improvement. As already pointed out by DeVries et al. [2018], modeling confidence estimation as part of the network itself might lead to interference with the actual training task. To this end, an initial cooldown period of 20 epochs for confidence estimation is introduced, as the largest effects are expected within the early training.

Finally, the influence of the confidence estimation on the test time performance has been quantified using an ablation study, comparing the proposed method (*metamemory*) to training without metamemory importance sampling (*conf. only*) and a baseline classifier with no confidence estimation at all (*BL*). Each method was evaluated with respect to accuracy (ACC), F1 score, positive and negative predictive value (PPV/NPV), matthews correlation coefficient (MCC) and area under the ROC curve (AUC). For multiclass data, 1-vs-all micro-averaging was used, dropping the binary PPV and NPV metrics. Further, the alignment of confidence and target class probability were assessed using Spearman's rank correlation coefficient r_s:

$$r_s = \frac{cov(R(c), R(p_y))}{\sigma_{R(c)} \sigma_{R(p_y)}} \tag{8.8}$$

with $R(v)$ denoting the rank transformation of v (cf. Appendix C). Significance was tested using two-tailed z-tests and 10,000 iterations of bootstrapping[7] [Efron 1982].

8.4.1 CIFAR-10

As mentioned above, first an ablation study was conducted using the CIFAR-10 dataset [Krizhevsky et al. 2009]. Therefore, a ConvNet architecture (cf. Chapter 9 Fig. 9.2) has been used, consisting of four blocks of 3x3 convolutions (32, 64, 128 and 196 filters), batch normalization, leaky ReLU activation and 2x2 max-pooling, followed by two additional fully-connected layers with 128 neurons each, and finally a 10-neuron softmax output layer. First, a baseline classifier was trained without additional augmentation, such as confidence estimation or importance sampling. Then the same network was trained using confidence augmentation but not metamemory importance sampling (*conf. only*) for an approximation of the error induced by an additional confidence estimation task. Finally, both were compared to a model which was trained as described above (*metamemory*), including confidence estimation and metamemory importance sampling, to demonstrate that the importance sampling effectively reduced the errors induced by an additional confidence estimation task and yields to significantly higher classification performance.

As depicted in Tab. 8.1, the baseline and metamemory approach significantly outperformed the confidence-augmented classifier with respect to the tested metrics with all $p < .001$. There were no significant differences between the baseline and the pro-

[7] The bootstrapping methodology will be discussed in more detail in Chapter 9.1.

	BL	conf. only	metamemory	sig.
ACC	.952	.917	**.955**	***
F1	.763	.576	**.774**	***
MCC	.737	.533	**.746**	***
AUC	**.971**	.909	.969	***

Table 8.1: Results for multiclass classification on the CIFAR-10 dataset using a baseline classifier *BL*, an augmented version including confidence estimation without (*conf. only*) and with metamemory importance sampling (*metamemory*). Significance of metamemory importance sampling over pure confidence estimation is denoted in the last column (***=$p < .001$, two-tailed z-test).

	BL	metamemory	sig.
ACC	.986	**.988**	
F1	.334	**.385**	*
MCC	.332	**.380**	*
AUC	.568	**.592**	*

Table 8.2: Results for multiclass classification on the CIFAR-100 dataset using a baseline classifier *BL* and metamemory importance sampling (*metamemory*). Significance of metamemory importance over baseline classification is denoted in the last column (*=$p <$ .05, two-tailed z-test).

posed metamemory approach ($\alpha = .05$). Spearman's rank correlation coefficient between confidence and target class probability r_s was measured with $r_s = .508$.

8.4.2 CIFAR-100

Secondly, applicability to problems with a high number of classes was evaluated. For this, the CIFAR-100 dataset was used. As depicted in Tab. 8.2, metamemory importance sampling significantly outperformed the baseline approach with respect to F1, MCC, and AUC ($p < .05$). Despite the high number of classes, target class output probability was still correlated to confidence with $r_s = .225$.

8.4.3 Radiological image data

As the method is expected to be especially beneficial when training data is rare, as it is typical for medical imaging scenarios, it also has been applied to an extended[8] version of the dataset A.2 for liver lesion growth prediction from [Katzmann et al. 2018a]. An overview of the data can be found in Tab. 8.3. Each approach was trained using 4-fold stratified grouped cross validation. The validation set was randomly split from the training set (2/3 for training, 1/3 for validation).

In contrast to the previous studies, it was aimed at an assessment of the incremental value rather than the absolute performance, the architecture from Chapter 6 was re-used

[8]The dataset used within this study has been part of the PANTHER project. It was incrementally acquired and annotated, thus during the later stages of the project the amount of available data increased.

samples	lesions	scans	patients
592	320	138	75

Table 8.3: Overview on the used mCRC dataset.

	BL	Metamemory	sig.
F1	$.298 \pm .040$	$\mathbf{.341 \pm .037}$	
PPV	$.234 \pm .036$	$\mathbf{.248 \pm .032}$	
NPV	$.893 \pm .014$	$\mathbf{.911 \pm .014}$	
MCC	$.157 \pm .047$	$\mathbf{.213 \pm .045}$	
AUC	$.604 \pm .037$	$\mathbf{.675 \pm .036}$	*

Table 8.4: Results and bootstrapped standard deviations for liver lesion growth prediction on the mCRC dataset using the baseline classifier (*BL*) and metamemory importance sampling (*metamemory*). Significant differences between both approaches are denoted in the last column (*=$p < .05$, two-tailed z-test).

with a reduced number of convolutional filters by a factor of 4, to make the training less costly. Subsequently, the training procedure was followed as described in [Katzmann et al. 2018a], first training a sparse autoencoder which then is used to train the final classifier. Augmenting the model architecture for confidence estimation was done after autoencoder pre-training.

As shown in Tab. 8.4, using metamemory importance sampling significantly improved the classification performance of mCRC liver lesion growth prediction with respect to the AUC ($p < .05$). The metamemory approach also consistently yielded higher values for F1, PPV, NPV, and MCC, although these were non-significant. Averaged Spearman's rank correlation between confidence and target class probability across all folds was $r_s = .251$.

Finally, it was assessed whether high confidence goes hand in hand with a strong determination of the classifier class output probabilities, as is assumed by the naive approach. For this purpose, the class determination was measured by using the Gini coefficient [Gini 1912][9] given by:

$$\frac{\sum_i^m \sum_k^m |p_i - p_k|}{2n \sum_i^m p_i} \tag{8.9}$$

and visualized for samples with "highest" (0.9-1.0), "high" (0.7-0.9), "medium" (0.3-0.7), "low" (0.1-0.3) and "lowest" (0.0-0.1) confidence. The results are depicted in Fig. 8.2, showing average Gini coefficients of 0.88, 0.72, 0.57, 0.46 and 0.33, respectively. Notably, only a small fraction of the predictions showed a Gini coefficient below 0.2, although the

[9]The Gini coefficient is a measure of (in-)equality in distributions, and has for example been used in the CART algorithm from Breiman et al. [1984]. When applied to discrete probability distributions, a Gini coefficient of zero indicates a uniform distribution while a coefficient of 1 means that exactly one of the possible results has a probability of 1 and all others of 0. Other measures of impurity would have been possible, too, such as the entropy, which has been used in the C4.5 algorithm. However, the Gini coefficient largely resembles the entropy measure and is computationally more efficient. Overall, the concrete choice is expected to only have a minor influence on the final result.

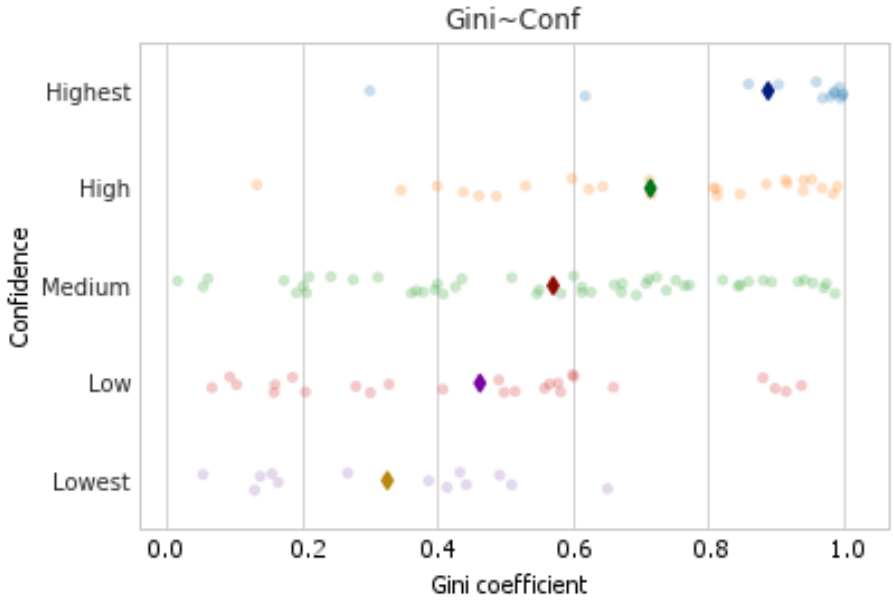


Figure 8.2: Visualization of the classifier output determination on the mCRC dataset as measured by the Gini coefficient for different confidence levels (see Sec. 8.4.3). Notably, while high confidence levels show a clear output determination, medium-confidence samples are hardly identifiable by output determination.

AUC was at only .675, implying that although the output determination correlates with the classifier confidence, it is not well aligned to it. Medium confidence samples showed output determinations over the full spectrum.

8.5 Discussion

As implied by the above results, the proposed metamemory importance sampling effectively reduces interferences of confidence estimation and classification tasks in the original confidence estimation framework from DeVries et al. [2018]. In each of the tested scenarios, it performed at least on par (CIFAR-10) or superior (CIFAR-100/mCRC) to the baseline classifier approach, while preserving a significant correlation of confidence estimate and test time classification performance. Notably, the largest improvements were observed when classes were represented by fewer samples, either due to a larger number of classes (CIFAR-100), or due to a lower number of samples in total (mCRC), which for clinical decision support is particularly beneficial, as data is typically scarce and difficult to acquire. Further, The assessment of the correlation between output determination and classifier confidence demonstrated the difficulties of interpreting output determination as a measure of confidence.

The process of metamemory importance sampling shows strong similarities to the idea of **curriculum learning** [Bengio et al. 2009]. However, there are also major differences. First, sample difficulty is determined not by a human expert or heuristically, but deduced as an inherent part of the training process itself by using the classifier confidence. As a result, the difficulty might vary with respect to the *current* learning state of the network, just as the difficulty of tasks in human learning strongly depends on prior knowledge

and understanding, rather than a *global problem difficulty*. While curriculum learning is not bound to a specific schedule, typically the problem difficulty of drawn samples is monotonically increased over the course of the training. In contrast, problem difficulty is inherently temporary in the presented approach and is defined by the current network state rather than a prior estimate. As a result, samples can get omitted or re-introduced into training as needed to maximize confidence. The approach is thus better comparable to the work from Houthooft et al. [2016] or Graves et al. [2017], as both suggest choosing samples according to the expected gain in knowledge. Both methods, however, do not take into account confidence, but rather task loss, and are therefore subject to the shortcomings initially discussed.

Major limitations of the study at hand arise from the confidence estimation model augmentation. In their original work, DeVries et al. did not conduct a systematic analysis of the optimality of this positioning, as was not done within this study. However, while using the latent space representation well corresponds to the intuition that it best represents the problem specifics, it might omit relevant information for OOD detection, and thus other positionings could be reasonable, too. Similarly, the study does not contain a systematic evaluation of various confidence estimation networks, which should be included in future work.

While the achieved confidence estimates within this study were shown to be directly related to the classifier performance, this study did not conduct a direct comparison of this particular property to the one achieved by the original approach from DeVries et al. [2018]. As no major modifications were introduced to the confidence estimation part of this method, however, the method is expected to perform mostly similar to it. Moreover, it would be desirable to see an evaluation of the approach on additional medical imaging datasets and additional architectures, such as ResNet [He et al. 2016] and its variants.

8.6 Conclusion

As pointed out earlier, knowing about classifier confidence is crucial for medical image classification. Within this chapter, the application of a method for classifier confidence estimation on medical imaging data was shown, without requiring larger modifications to the model, and thus being applicable to various architectures. Using a psychologically motivated framework called *deep metamemory*, it was demonstrated that these confidence estimates can successfully be used to improve classifier test time performance when training data is scarce, being a typical scenario in medical image analysis. With this, it is hoped that this contribution will be a step toward the transformation of medical image classification systems into clinical practice. While this chapter covered the estimation of confidence as a scalar measure and thus might allow the practitioner to identify situations of classifier uncertainty, the method at hand does not allow for an inspection of the reasons for this. To this end, Chapter 10 will discuss a method for classifier decision explanation, allowing the practitioner to visually inspect classifier decisions in the input space.

9. Bootstrapping Methods

In the previous chapters, methods for lesion growth prediction, survival estimation, hazard regression, and confidence estimation were analyzed. Common to all of them was that they each introduced ways to combat the overfitting problem which may occur when training complex classifiers in a small data setup, such as is often faced in the medical imaging domain. While having seen sparse representations, autoencoder-based approaches, dimensionality reduction, and manual architecture tuning, so far a general, widely applicable mechanism for training deep networks in a small data environment is missing. This chapter, therefore, introduces a mechanism based on *bootstrapping*, being applicable at multiple granularities in the model design, as well as to a wide variety of different deep architectures, which is demonstrated to show clear advantages when training deep neural networks on small datasets, and contributing towards their successful application in these scenarios.

9.1 Bootstrapping

The term bootstrapping refers to the process of estimating *distributions of statistics* from only one set of samples by repeatedly computing the statistics of interest on randomly generated subsets of the original data. It thus belongs to the broader class of statistical resampling methods and was originally introduced by Efron [1979]. It is typically applied if the theoretical distribution F of a statistic is unknown, replacing it with the empiric distribution $\hat{F}$.

Let us assume an m-class classifier $f(x) : X \rightarrow C$ with classes C and $m = |C|$, producing n estimates $\hat{Y} = \{\hat{y}_1, ...\hat{y}_n\} \in C^n$. Given ground truth labels Y, the accuracy across these estimates can be calculated as:

$$acc(Y, \hat{Y}) = \frac{1}{n} \sum_{i}^{n} [y_i = \hat{y}_i] \tag{9.1}$$

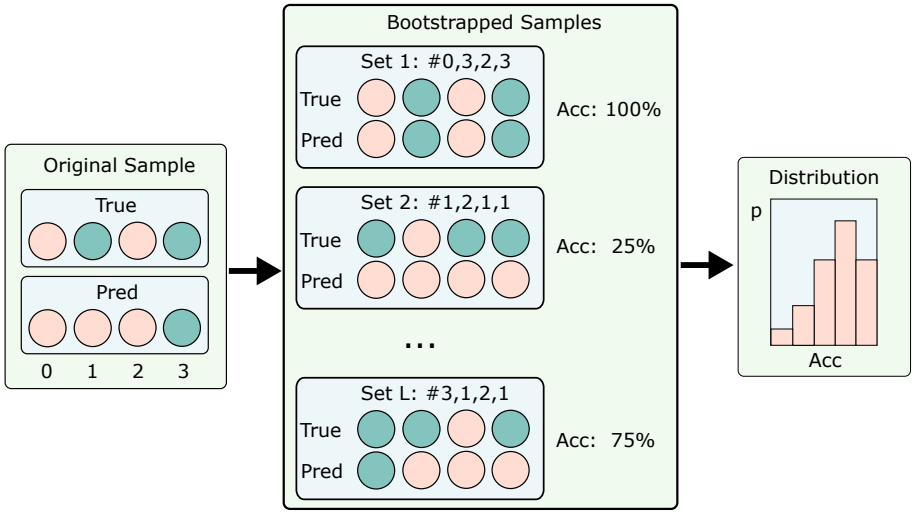


Figure 9.1: Visualization of the estimation of the accuracy distribution given a paired sample of ground truth and classifier predictions. The bootstrapping method draws L sample sets from the original paired data and calculates the accuracy of each. As a result, a distribution of the probability of different accuracies is generated which can further be used for the estimation of confidence intervals, statistical testing, etc.

As we do only have one specific realization of the pair $(Y, \hat{Y})$ it is not possible to calculate derived statistics for *acc* without experimental repetitions. The idea of bootstrapping is to repeatedly resample the data by drawing n samples from the original data with replacement and to subsequently calculate the statistic of interest, e. g. *acc*, on each of these realizations. A visual example of the process is depicted in Fig. 9.1. By repeating this process until convergence, bootstrapping is capable of generating an empirical distribution for *acc*, which can further be used to calculate additional, derived statistics, such as the mean, standard deviation, standard error, percentiles, confidence intervals, and others, proverbially pulling itself up by its own bootstraps. As a result, bootstrapping allows to generate measures of confidence by employing only a single prediction set, and for this purpose was shown to even outperform statistics from approximate parameterized distributions (cf. [Efron 1987]). Although having its roots in the field of statistical resampling, a larger number of later works in the machine learning domain is fundamentally built upon the bootstrapping concept, with random forest classifiers [Breiman 2001] and their derivatives probably being the most well-known example.

9.2 Classifier Architectures

Bootstrapping can be applied to deep neural architectures in a variety of ways, all sharing the property that the bootstrapping mechanism serves as a constraint for the constructed latent space, in particular only allowing those optimizations which perform well not only on the overall sample set on average but over its whole estimated distribution. It can thus be used as a regularization of model parts, as a major determinant of the model architecture, or as a meta-model itself. In the following, several architectures will be highlighted which have successfully been applied to small data for medical image classification, yielding an overall improved performance, and thoroughly discuss their benefits and disadvantages.

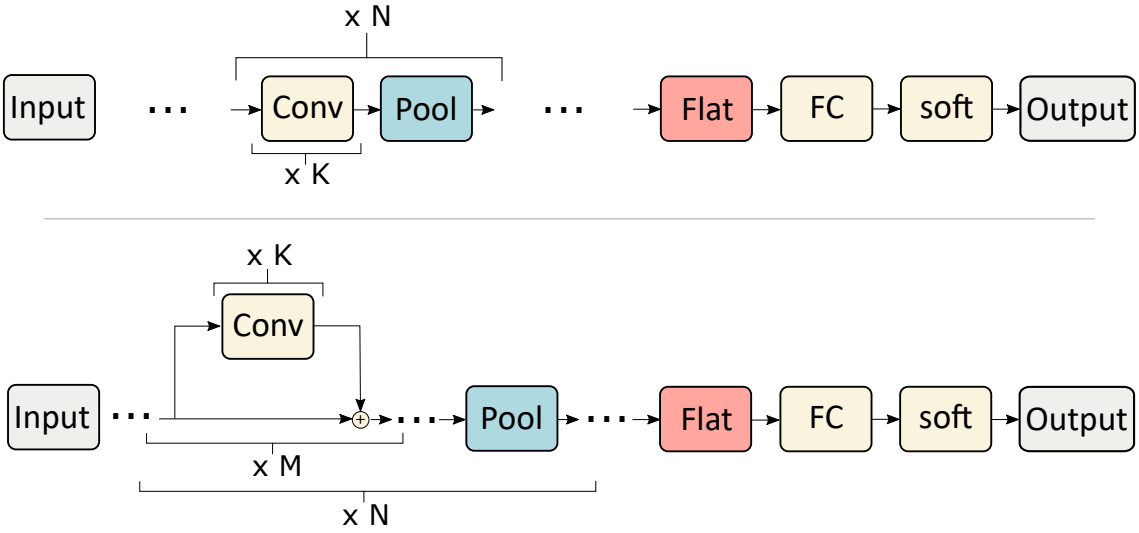


Figure 9.2: Comparison of the basic ConvNet (top) and its respective ResNet architecture (bottom). Both networks consist of N stages of convolutional blocks, with each stage being followed by a pooling operation (e. g. maximum pooling). While in the ConvNet approach, each stage contains K convolutional blocks, the residual network employs M steps of K convolutional blocks (i. e. *residual blocks*). Most importantly, the residual network bypasses former activations using an additive term to enforce a better gradient flow for deep networks (cf. [He et al. 2016]).

9.2.1 Bootstrapped Path Shaking

First, it will be shown that the bootstrapping procedure can be used as a regularization mechanism directly within the model architecture, which will be referred to as *Bootstrapped Path Shaking* (**BPS**).

The BPS concept is based on a similar idea like the *Shake-Shake* and *ShakeDrop* regularization approaches from Gastaldi [2017] and Yamada et al. [2018] which will be briefly discussed in the following. Both Shake-Shake and ShakeDrop are constructed as an augmentation of residual architectures, such as ResNet [He et al. 2016], Wide ResNet [Zagoruyko et al. 2016] or ResNeXt [Xie et al. 2017]. The basic concept of ResNets is explained in Fig. 9.2, the Shake-Shake and ShakeDrop regularizations are depicted in Fig. 9.3 (top and middle).

Generally, most neural architectures can be easily amended to contain residual connections by simply replacing convolutional with residual blocks. As a result, the Shake-Shake and Shake-Drop approach are widely applicable. In the original ResNet (and its derivatives), convolutions are applied deterministically and equally in each training and/or prediction step, i. e.:

$$f(x) = x + c(x) \tag{9.2}$$

with f denoting the output of a single ResNet stage for input x and *one* convolutional lane c (e. g. BN+Conv+ReLU). The Shake-Shake approach modifies this behavior by introducing *two* convolutional lanes c_0, c_1 and outputting their mixture. During training, this mixture is given by:

$$f(x) = x + \alpha \cdot c_0(x) + (1-\alpha) \cdot c_1(x) \tag{9.3}$$

and in the gradient backpropagation step is exchanged by:

$$f(x) = x + \beta \cdot c_0(x) + (1-\beta) \cdot c_1(x) \tag{9.4}$$

with random factors $0 \leq \alpha, \beta \leq 1$. As pointed out by Yamada et al. [2018], the above step effectively results in a latent-space interpolation and thus an augmentation of the samples as proposed by DeVries et al. [2017], well-aligned with other recent work (cf. [Verma et al. 2019]). Yamada et al. [2018] improve on the Shake-Shake method by omitting the necessity of introducing multiple lanes, proposing a novel stabilization mechanism by adding a random gating variable g, determining whether the latent-space augmentation is applied or skipped (cf. Fig. 9.3).

As pointed out by Yamada et al. [2018], Shake-Shake performs best when the same lane is dominant in both the forward and backward pass, i. e. when $(\alpha < .5 \wedge \beta < .5) \vee (\alpha > .5 \wedge \beta > .5)$, and otherwise might yield only low accuracy. As Shake-Drop lacks a second, error-correcting branch, it addresses this issue by randomly dropping the augmentation using the gating parameter g. Both approaches yield good results if the latent space is well-constrained, being the case if the problem is not too complex and/or a large amount of training data is available. However, if only a few data are available, i. e. if the latent space is more rugged, the proposed interpolation can become ineffective, as interpolated samples may lie outside of the desired latent space distribution[1] (cf. Fig. 9.4). Note, that this issue is thus strongly related to the problem of catastrophic forgetting such as addressed by Kirkpatrick et al. [2017].

In turn, one might suggest addressing this problem by using a method which is denoted as *Bootstrapped Path Shaking*. Instead of randomly dropping or using residual lanes as in the Shake-Shake and ShakeDrop concepts, the key idea is to use bootstrapped subsets of the original training set to train k residual pathways with $k \geq 1$, following the notion of bootstrapping as an approximation of the underlying distribution of the available data. In contrast to the Shake-Shake and similar to the ShakeDrop regularization, BPS does not require multiple processing lanes (i. e. $k \geq 2$), although being combinable to increase the effective network capacity, which may result in a better test time performance. Each of the k BPS lanes is trained only on a bootstrapped subset of the original data. This can be realized by keeping track of the sample indices during training, only forwarding activations and gradients to each residual path if the sample is contained in its respective bootstrapped subset. During the test phase, all paths are averaged similarly to the original bootstrapping idea and as also done in other bootstrapped approaches, such as random forest classifiers (see below). A comparison to Shake-Shake and Shake-Drop is depicted in Fig. 9.3. In Sec. 9.2.5, it will be demonstrated that BPS can effectively be used to constrain the latent space if only a few data are available in order to avoid the above-discussed issues arising from an application of latent space augmentation methods such as Shake-Shake and ShakeDrop.

9.2.2 Deep Random Forests

In the previous section, it was shown that bootstrapping can be applied as part of the model architecture, serving as an effective regularization, and being applicable to a wide range of

[1] This issue can vividly be expressed in an analogy to an autonomous robotic system driving towards a wall. While approaching, the robot's activations for an evasion maneuver towards both left and right would be increasing. Averaging the activations would be fatal as it results in driving straight into the wall.

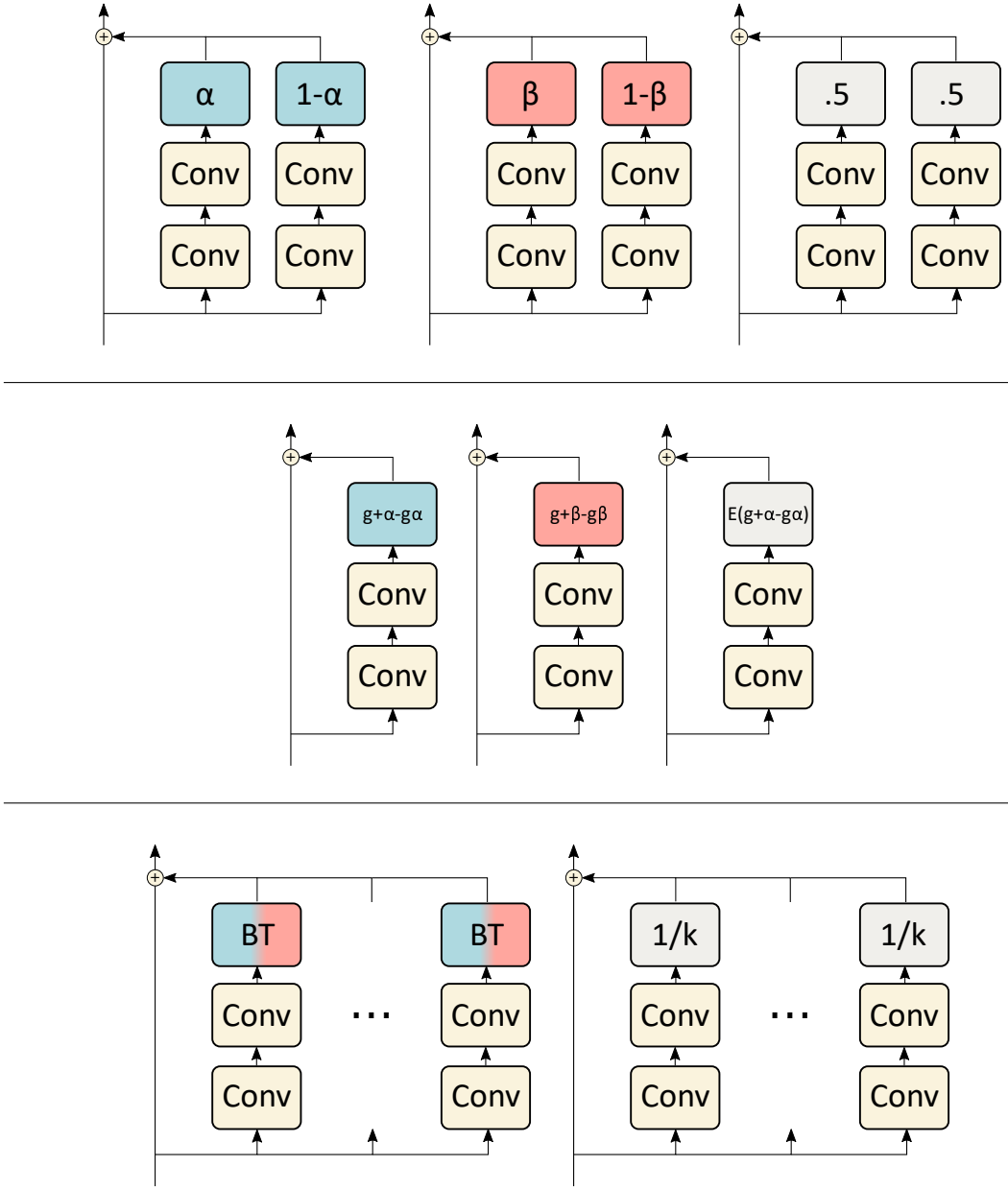


Figure 9.3: Comparison of the Shake-Shake (top) and ShakeDrop (middle) concepts from Gastaldi [2017] and Yamada et al. [2018], as well as the proposed Bootstrapped Path Shaking (BPS) algorithm (bottom) for residual convolutions. Blue and red modules indicate an activation factor applied during the forward or backward pass, respectively, while grey indicates the weighting during the test phase. Shake-Shake combines two residual pathways by randomly sampling a factor α in the forward, and a different factor β in the backward pass of the training. ShakeDrop realizes a comparable behavior but utilizes only a single pathway, which instead is both trained randomized and non-randomized, depending on an additional gate variable g. BPS follows a similar approach to Shake-Shake but randomizes by feeding bootstrapped subsets to the different convolutional lanes.

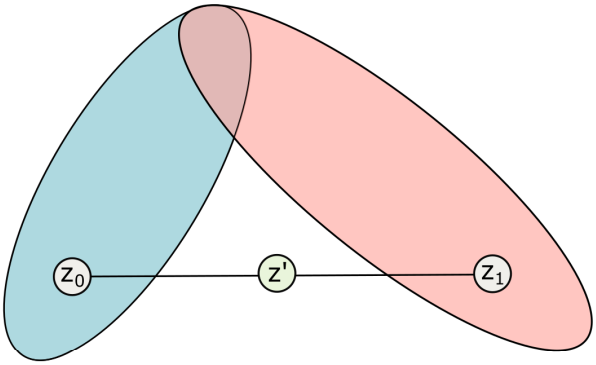


Figure 9.4: Visualization of the averaging problem resulting from latent-space augmentation for methods such as Shake-Shake and similar. Assuming two processing lanes producing activations z_0, z_1, both lying in a well-constrained latent-space region, euclidean interpolation of a sample $z' = (z_0 + z_1)/2$ might yield a sample outside of the well-defined region, and thus might be detrimental to the training process.

architectures. However, as discussed there, introducing changes to the model architecture might not always be desirable or feasible, e. g. when using pre-trained networks or when a specific architecture is needed due to task requirements. For this purpose, with [Katzmann et al. 2020] a method for bootstrapped meta-classification was proposed, called *Deep Random Forests* (**DRF**), using the well-known *random forest classifier* (**RF**, see below) concept from Breiman [2001].

RFs are known to be generally easy to use, require few optimizations and provide a remarkably robust classification performance. This was for example emphasized by Parmar et al. [2015], assessing various feature selection and classification methods for NSCLC radiomics prediction, and including a total of 168 pairings, which concluded: "*[The] majority of feature selection methods gave highest predictive performance when used with the random forest (RF) classifier.*".

Applying RFs requires the computation of meaningful, hand-crafted features prior to classification, which, as pointed out in [Katzmann et al. 2020], is often diametrically opposed to the goals of the application of deep neural networks, explicitly focussing on the leveraging of information which has previously been unused, and thus is not yet covered. A combination of both approaches might therefore lead to a synergistic model, benefitting from both the data analysis capabilities of deep neural networks as well as the remarkable robustness of random forest classifiers. Again, the presented approach will follow the core concepts of the bootstrapping methodology.

Random Forest Classifiers

RFs are built as an ensemble of *randomized decision trees* (**RDTs**). Similar to classical *decision trees* (**DT**), RDT nodes chose optimal split features in order to partition samples in such a way that the impurity of the child nodes (e. g. using the Gini coefficient, cf. Eq. 8.9) is minimized, i. e. at best contain samples of only a single class. In contrast to DTs, RDTs take into account only a random feature subset. This process is repeated until all nodes are pure, i. e. consist of samples of only one class, or cannot be further split, e. g. as samples are identical but have different labels. In the test phase, samples are classified by walking along the tree until one of the terminal nodes is reached and returning its class.

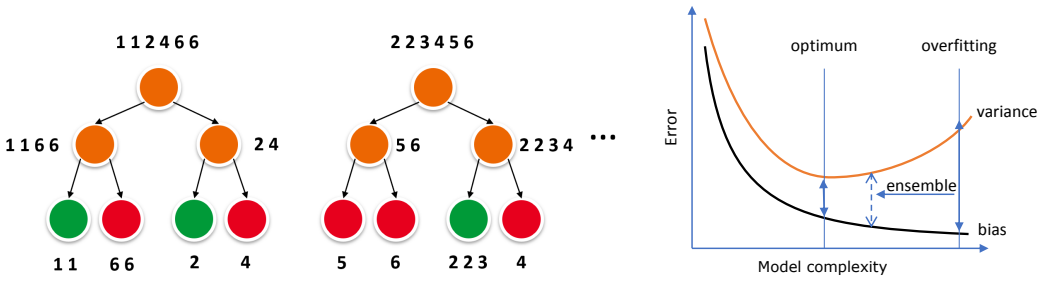


Figure 9.5: **Left:** Example of a simple RF architecture built from samples 1-6, with samples 1-3 being positive (green) and 4-6 being negative (red). Multiple randomized trees are grown, each using a bootstrapped subset as depicted above each tree. **Right:** Visualization of the bias-variance dilemma. As depicted, RFs basically ensemble multiple low-bias/high-variance classifiers (i. e. RDTs), leading to approximations nearer to the classification optimum. Source: [Katzmann et al. 2020] © 2020 IEEE

RFs are created by repeating this procedure for K RDTs, each of them independently grown from a bootstrapped subset of the original data (cf. Sec. 9.1). During the test phase, a sample is classified by maximum voting across all RDTs, furthermore providing the opportunity to express class probabilities as the portion of RDTs choosing this class. In summary, the idea of RFs is to use bootstrapped ensembles of low-bias/high-variance estimators to yield estimates nearer to the classification optimum (cf. Fig. 9.5).

Deep Architecture

Analogously to the above RF scheme, DRFs consist of K decision trees, each being trained on a bootstrapped subset of the original data. Each DT again is trained analogously to the RF concept. However, different from the RF concept no pre-computed features are available, but only raw inputs. This is where the deep neural network comes into play:

For each node, a deep model is trained to solve the actual classification task using the samples within this node. For this, any architecture which is adequate for the task at hand can be used. While the root node classifier in each tree is randomly initialized, child nodes get initialized with the weights of their parents, being inspired by the methodology used in *neural architecture search* (cf. [Elsken et al. 2019]).

As a result of this training, the pre-output layer now contains a latent-space representation which is meaningful with respect to the classification goal, and its activations can further be used as features. Now again analogously to the RF, the samples within each node are split according to a minimization of the resulting node impurity (see above). To summarize, the process consists of the following steps:

1. Create K root nodes
2. For each tree, draw N samples with replacement (bootstrapping)
3. Train root nodes T_k to split samples (X_k, Y_k) into two sets $(X_{k,0}, Y_{k,0})$ and $(X_{k,1}, Y_{k,1})$
4. Train subnodes $T_{k,j}$ with sets $(X_{k,j}, Y_{k,j})$

Analogously to the construction of RFs, step 4 is applied repeatedly, i. e. subtrees are created until the node impurity cannot further be reduced. As the node classifiers have significantly larger computational requirements than the RDT step functions, a maximum

tree depth d is defined. If a node at the maximum tree depth is not pure, it returns the result of the classification at this level instead.

An interesting side effect of the RF-like structure of the proposed approach is the ability to calculate an out-of-bag (**OOB**) error. This is achieved by evaluating each tree on its respective unused part of the training data and averaging across trees, effectively eliminating the need for an additional validation set by yielding a good estimate of the test time performance.

9.2.3 Bootstrapping Ensembles

When comparing the RF and DRF architectures, the major difference lies in the complexity of the node classifiers. The RF sample partitioning is conducted using a step function on a single feature dimension, and thus allows for a binary partitioning in each stage, being equally true for the DRF concept. In contrast to the features in the RF approach, however, DRF features are particularly trained to split the samples along their final classification label, rather than occasionally doing so. As a result, the node impurity reduces much faster than in RFs, diminishing the value of the tree structure. In fact, in [Katzmann et al. 2020] it was shown that the test time performance improves only marginally with increasing tree depth, implying that most of the observed performance boost can be attributed to the bootstrapped ensembling process, as will be discussed in Secs. 9.2.4 and 9.2.5.

Following this argumentation, it might be promising to reduce the proposed DRF model to a bootstrapped ensemble, being equivalent to setting a maximum tree depth of $d = 0$. However, as now the model can be easily composed into a single, piecewise differentiable architecture, it can be trained in an end-to-end fashion, solving some of the issues which have been specifically addressed by the DRF concept, such as the derivation of sub-classifiers, the optimality of the used split, and finally the optimal initialization of submodels. Notably, using a single model allows for a continuous evaluation of the whole model during training, rather than separately for each component, ensuring *overall* optimality with respect to the OOB loss. Altogether, using a single model might result in a significantly better runtime and model efficiency as well as an improved performance. In the following, this approach will be denoted as *Bootstrapped Network*, or **BTNet**.

BTNets are modelled as a *meta-model*, i. e. being built upon K parallel models m_k, all having the same architecture. Each model m_k is assigned a bootstrapped subset (X_k, Y_k) of the original training data with $(X_k, Y_k) \in (X^N, Y^N)$ and $|X_k| = |Y_k| = |X| = |Y| = N$. The sample indices are kept track of via an additional input I. The overall concept is depicted in Fig. 9.6. In each training step, the meta-model passes each submodel its respective subset by applying a *filtering step* using the sample indices I. Afterwards, the resulting class probability estimates $\hat{Y}_k$ are recombined and averaged in a *recombination step* to form an overall estimate

$$\hat{Y} = \frac{1}{K} \sum_{k=1}^{K} \hat{Y}_k \tag{9.5}$$

The losses $\mathscr{L}_k$ are computed for each model seperately on its bootstrapped subset only, resulting in an overall loss function:

$$\mathscr{L}_{\text{total}} = \sum_{k=1}^{K} \mathscr{L}(Y_k, \hat{Y}_k) \tag{9.6}$$

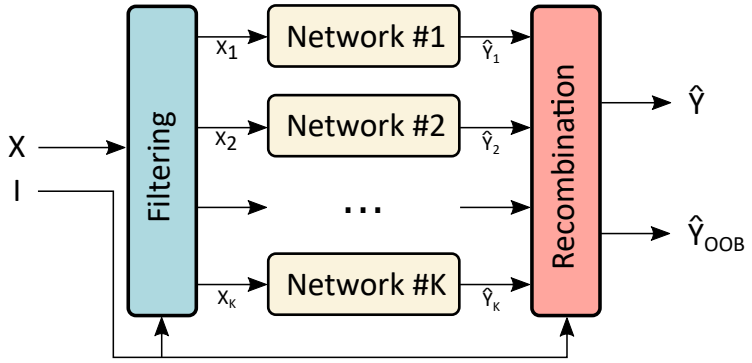


Figure 9.6: Basic idea of the Bootstrapped Network (BTNet) described in Sec. 9.2.3. An arbitrary network architecture is replicated K times and composed in a single model. Using a filtering layer (blue), each subnetwork is trained using a bootstrapped subset (X_k, Y_k) of the original data by keeping track of indices I. Finally, in a recombination step (red), the network estimates are recombined to yield a per-sample estimate Y as well as an OOB-classification Y_{OOB}, which can be used for internal validation.

The used loss function $\mathscr{L}$ is mostly arbitrary and can be chosen according to the classification task. However, using a loss function that can handle label imbalance is recommended, such as correlational loss or focal loss (cf. [Lin et al. 2017]), as the bootstrapping process does not give any guarantees on the label distribution, which especially can become important for smaller datasets.

9.2.4 Experiments

Each of the described architectures is evaluated in comparison with a baseline *ConvNet* and its respective residual version (*ResNet*, see Fig. 9.2). The original ConvNet consisted of an initial convolutional block with 64 3x3 filters, followed by subsequent blocks of 64, 128, and 256 filters. Each block consisted of a convolutional layer, followed by leaky ReLU activation [Maas et al. 2013]. Each but the initial block was batch normalized [Ioffe et al. 2015], and after each but the initial and the last block a 2x2 maximum pooling has been applied. Finally, a global maximum pooling[2] has been applied to the resulting feature matrices, followed by a m-neuron softmax output layer for the final classification, with m representing the number of classes of the dataset.

The residual network was built analogously, with the convolutional blocks of the ConvNet being replaced by stages of residual blocks like in the original ResNet implementation from He et al. [2016] (see Fig. 9.2). Each stage consisted of 3 residual blocks, with each residual block consisting of two subsequent convolutional blocks, which were built analogously to the above description (Conv+LeakyReLU+BN). For the Shake-Shake regularization, all residual blocks were duplicated as depicted in Fig. 9.3. The BPS and ShakeDrop regularization used no replications, i. e. had the same architecture as the ResNet, and only introduced their respective regularization mechanism.

The BTNet and the DRF, both being based on the original ConvNet model, were built using a 30-times replication of their respective architecture.

[2]Global maximum pooling serves a similar need as maximum pooling but returns the *global* maximum for each filter, mapping images to scalar values. It is preferable if the concrete location of classification is irrelevant.

All methods were evaluated on a variety of datasets[3], including:

1. The **CIFAR-10 and CIFAR-100 datasets** from [Krizhevsky et al. 2009] (N=50,000/10,000 samples for train/test, each) – Both are widely used and publicly available benchmark datasets, thus allowing for easily reproducible results. Additionally, the CIFAR-10 cats-vs-dogs and cars-vs-trucks binary subsets are used[4].

2. The publicly available **BreastMNIST**[5] (N=624/156), **ChestMNIST** (N=89,687/22,433) and **PneumoniaMNIST** (N=5,232/624) datasets from the MedMNIST collection [Yang et al. 2021] for breast lesion malignancy, chest disease and pneumonia classification – These datasets significantly better represent a medical imaging scenario than the above-used RGB image data. The datasets are based on ultrasound and x-ray imaging, and show a considerable amount of label imbalance.

3. The **LIDC-IDRI** dataset (N=772) from Armato et al. [2011], Armato III et al. [2015], and Clark et al. [2013] – As a high-quality and widely-used benchmark dataset in the medical imaging domain, the LIDC-IDRI dataset for lung lesion malignancy classification serves as a comparison for the final application domain[6].

Where available (CIFAR-10 and subsets, CIFAR-100, BreastMNIST, PneumoniaMNIST, ChestMNIST), the dataset's original train/test split has been used. If not, it has been chosen randomly using an 80/20 split. For the non-bootstrapped networks, the validation set was randomly split from the training data using another 80/20 split. This could be omitted for the bootstrapped networks due to their inherent capability to validate on OOB samples (cf. Sec. 9.2.2 and Fig. 9.6). Datasets with less than 10,000 samples (BreastMNIST, PneumoniaMNIST, LIDC-IDRI) were trained using 5-fold repetition to increase the result reliability[7]. Random seeds were identically varied for each method and repetition to ensure comparability across results.

As an increase in performance is expected by using ensembling methods already due to the increased network capacity, finally an additional experiment on the CIFAR-10 dataset is conducted, comparing the BTNet approach with a non-bootstrapped classifier ensemble of the same capacity. The bootstrapped network is expected to particularly benefit from its ability to better account for label noise and small to medium-sized training data. Thus, both the BTNet as well as a non-bootstrapped ensemble are trained using 30 replications of the original network (a ConvNet architecture as described above), systematically applying label noise to the training data (i. e. changed labels) with an increasing rate between 0 and 90 %.

[3] In contrast to the previous studies, the mCRC dataset could *not* be used in this study due to the data use policy, unfortunately prohibiting further use of the data. An evaluation of the Deep Random Forest concept on the mCRC dataset, however, already has been conducted in [Katzmann et al. 2020].

[4] As will be discussed more thoroughly in Chapter 10, these two represent the most difficult 1-vs-1 classifications within the CIFAR-10 dataset and thus serve as a binarization of the CIFAR problem.

[5] The BreastMNIST dataset is also used in Chapter 10.

[6] The preprocessing of the LIDC-IDRI data was conducted analogously to [Nibali et al. 2017], and will be in the focus of the deep decision explanation in Chapter 10. It will be discussed there in more detail.

[7] This step was omitted for the ChestMNIST, CIFAR-10, subsets, and CIFAR-100 dataset, as the bootstrapped estimated standard errors were below 10^{-3} using a single repetition [Efron 1982].

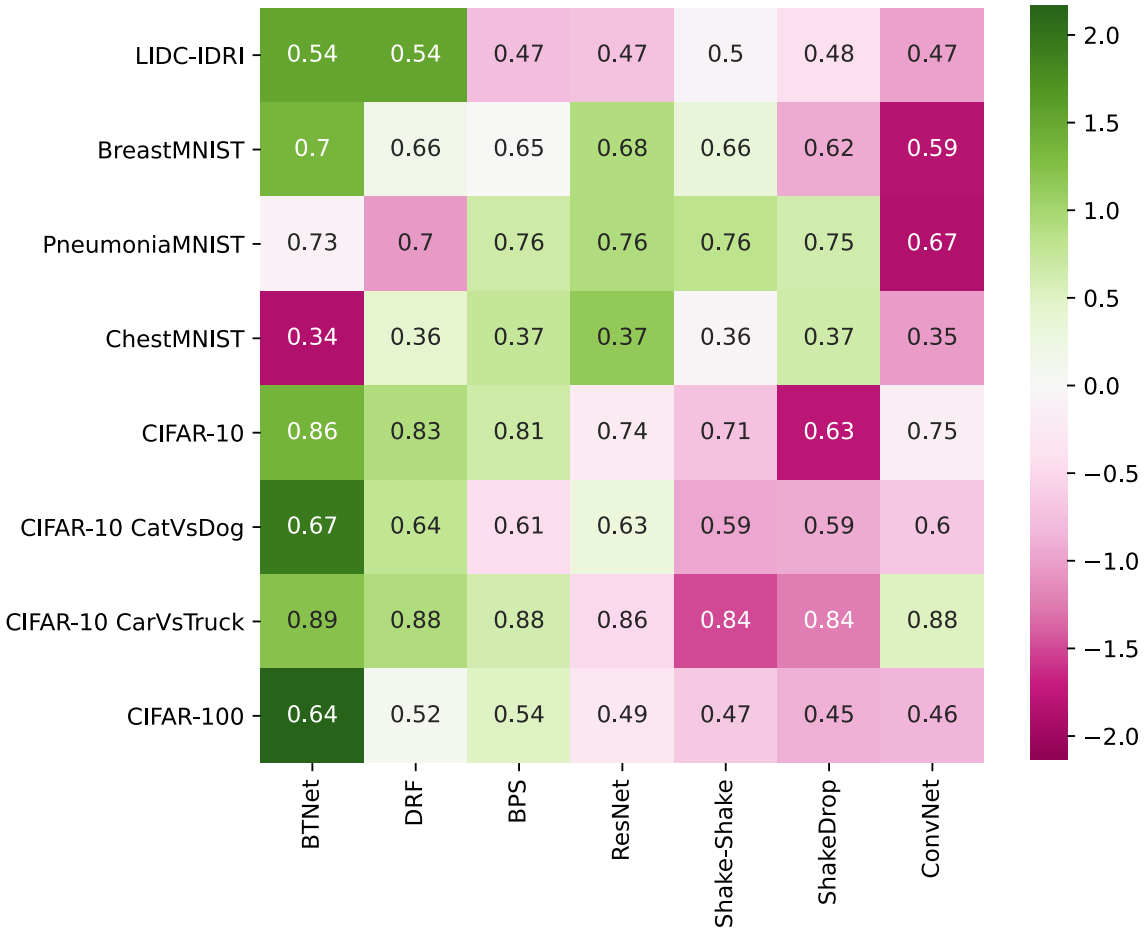


Figure 9.7: Comparison of bootstrapped and non-bootstrapped classifiers on various medical (LIDC-IDRI, BreastMNIST, PneumoniaMNIST, ChestMNIST) and non-medical (CIFAR-10, its subsets and CIFAR-100) imaging datasets with respect to the matthews correlation coefficient (MCC) (green is higher). For the color encoding, the results were z-normalized per dataset to account for varying dataset difficulty, with the original values being annotated per cell. The classifiers are sorted by their average z-value across all datasets (decending f.l.t.r).

Additionally, the training set size was varied with $N \in \{50, 100, 150, 200, 300, 500, 1000\}$, and the performances of both networks were evaluated against each other using the full CIFAR-10 test set. Each experimental pairing (noise/training set size) was repeated 10-fold with identically varied seeds across experiments. Finally, direction and significance were assessed using a two-tailed t-test with $t(N-2) = t(8)$.

9.2.5 Results

A comparison of the results for all evaluated approaches and datasets can be found in Fig. 9.7. Across datasets, the BootstrapNet (BTNet) overall performed best, on average yielding the highest MCC values and achieving superiority over all other tested approaches in 6 out of 8 datasets. It clearly outperformed the other tested approaches on the LIDC-IDRI, BreastMNIST, Cat-vs-Dog, Car-vs-Truck, CIFAR-10, and CIFAR-100 dataset. For PneumoniaMNIST it performed average, and only for ChestMNIST provided a slightly

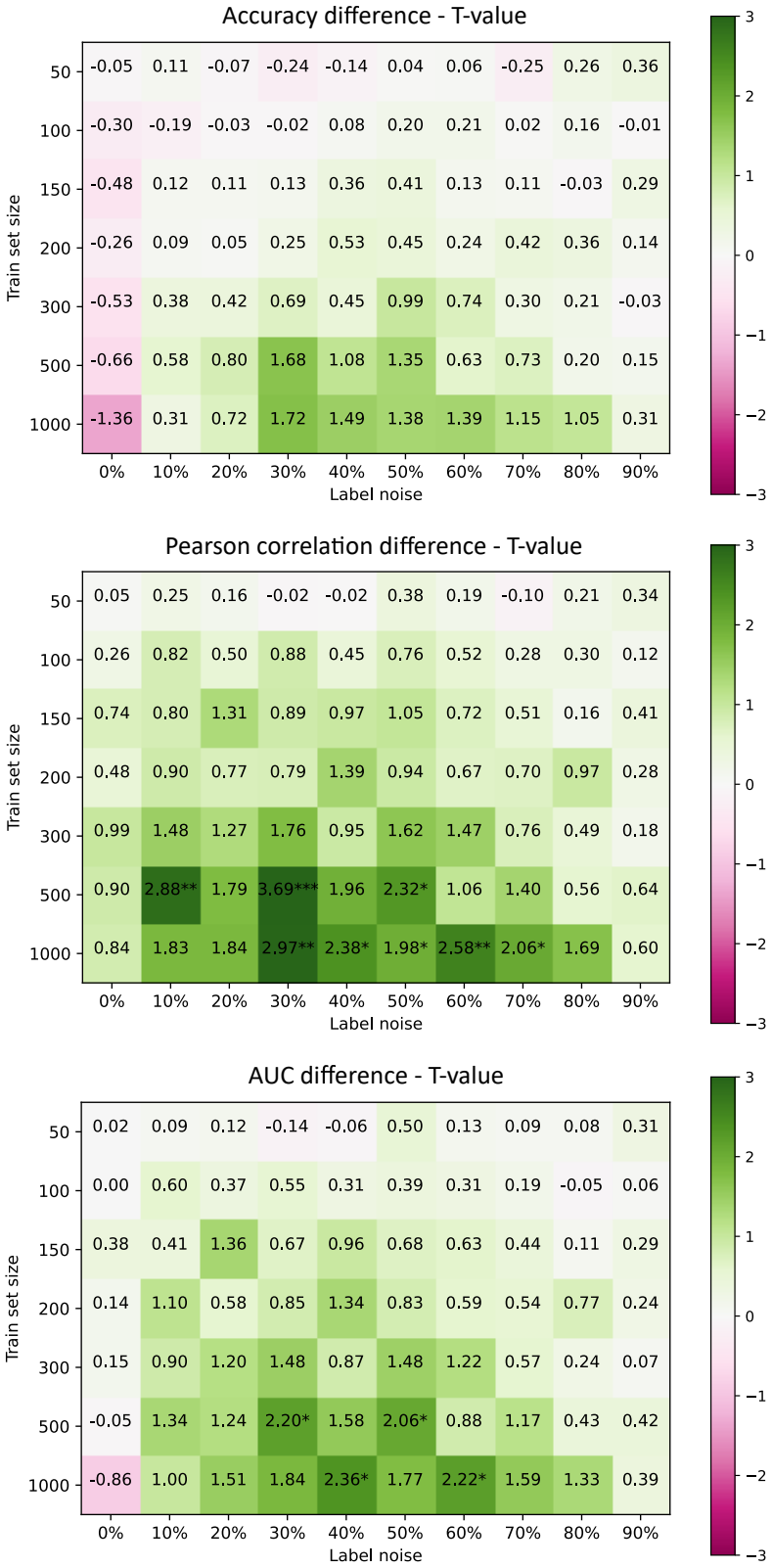


Figure 9.8: Comparison of the BTNet and plain classifier ensemble results on the CIFAR-10 dataset for systematically varied label noise and training set sizes as measured by accuracy (top), pearson correlation (middle) and AUC (bottom). T-values are indicated as green (higher value for BTNet) and red (higher value for ensemble) and are annotated each cell. Significance for $t(8)$ is indicated as $p < .05$*/.01**/.001***

lower MCC than the other tested approaches, overall resulting in an average improvement of 0.95σ as measured by dataset-wise z-normalization[8].

The overall second-best approach was the Deep Random Forest (DRF), performing better than average in 7 out of 8 datasets and yielding a mean improvement of 0.47σ, followed by Bootstrapped Path Shaking (BPS; 0.24σ), being the best non-ensembled method, and performing better than average in 5 out of 8 datasets. In particular, BPS clearly outperformed the Shake-Shake and ShakeDrop regularization, with both having a comparable aim. In this order, it follows ResNet with better-than-average results on 4 datasets, Shake-Shake (2), ShakeDrop (2), and finally ConvNet (1).

Regarding the label noise experiment, Fig. 9.8 depicts a comparison of the results achieved by the BTNet and the plain ensemble approach[9]. As assumed, using the BTNet approach particularly resulted in significantly superior performance over plain ensembling when label noise was present. For medium training set sizes, the BTNet outperformed the plain ensemble with T-values of up to 3.52 (Pearson correlation, $p < .001$) and 2.36 (AUC, $p < .05$). In contrast, if no label-noise was present, although slightly better aligning to the distribution as measured by the Pearson correlation, the BTNet was gradually outperformed by the plain ensemble with increasing training set sizes as measured by accuracy and AUC.

9.2.6 Discussion

As demonstrated by the results, using bootstrapped ensembling as part of the model architecture can greatly increase classification performance. The largest improvements were observed when using ensembling techniques. In fact, this was expected, due to the resulting higher network capacity. However, as was shown by both, the superiority of the bootstrapped network over plain ensembling, and the superiority of the Bootstrapped Path Shaking over other methods with similar (ResNet/ShakeDrop) or even higher (Shake-Shake) network capacity, this superiority can not only be attributed to network capacity, but rather to a more robust approximation of the underlying latent space distribution.

Notably, as was shown with the label noise experiment, this effect was strongest, when label noise was present and only small- to medium-sized datasets were available, which well aligns with the initial assumptions. In contrast, when compared to non-bootstrapped ensembling, the effect gradually decreased, and is finally inverted, the larger the available dataset gets. This well fits with intuition, as for large datasets, due to their high parameter amount, complex neural networks are well-known to successfully handle noisy or incomplete data, but tend to over-specialize if only a few data are available. In contrast, bootstrapped ensembling enforces a regularization which is beneficial in small-data settings, but simultaneously reduces the statistical independence of the data given to each of its components [Efron et al. 1997].

Interestingly, as demonstrated by the experimental results, bootstrapping can successively be integrated into the network structure, e. g. by using Bootstrapped Path Shaking (BPS), and may similarly result in superior performance. In the experiments, it outperformed other regularization methods, such as Shake-Shake and ShakeDrop, in particular being noteworthy as a) BPS reduces the statistical independence of the training data (see above), and b) especially the Shake-Shake regularization used replications, and thus had a higher network capacity.

[8] Similarities between datasets and approaches are depicted in a dendrogram in the Appendix Fig. B.4

[9] The complete results for all metrics and each of the classifiers can be found in the Appendix Fig. B.2–B.3.

In fact, both approaches failed to provide a sufficient regularization in our scenario. There are multiple possible reasons for this: First, the network architecture was rather small. More complex architectures, and thus more possible paths for the Shake-Shake and ShakeDrop regularization, might result in a better performance for both methods. Second, when proposed, both methods were evaluated using data augmentation. As the inner-network regularization of both methods is undirected, it might be possible that input space augmentations are important in order to avoid ambiguities. Finally and most importantly, in their original publications, both methods were shown to be sensitive to the concrete parameterization and can become unstable if the combination of parameterization and architecture does not fit. As within this study, no systematic meta-optimization was conducted, the parameterization might have been suboptimal for the problem at hand.

The study has some further limitations which have to be noted at this point: First, while being a good baseline, the MedMNIST datasets provide only a limited insight into algorithmic capabilities within medical image analysis, as with 28x28 pixels they have a very small resolution, just like the MNIST and CIFAR-10/100 datasets. In this regard, especially the LIDC-IDRI dataset is much more realistic, as it consists of high-resolution CT images, with its processing well-aligning with typical clinical algorithms. It is thus encouraging that the BTNet and DRF approach also performed best on the LIDC-IDRI data.

The second limitation arises from the invasiveness of methods such as BPS, Shake-Shake, and Shake-Drop. In a practical setting, there might be situations in which a change of architecture is impossible or infeasible. This, for example, might happen if the network architecture itself is a result either of convention or of optimization, such as the EfficientNet architecture from Yakubovskiy [2020]. Similarly when using transfer learning, no augmentation of the architecture is possible. However, both the BTNet and DRF architecture may be applied in these scenarios.

Another limitation of the bootstrapped methods arises from their requirement to keep track of the training indices. If the training set size is not known in advance, bootstrapped methods require additional extensions to be applicable. In highly-distributed training settings, in which the training data is not centrally known, such as in Federated Learning environments [Rieke et al. 2020], the application of bootstrapped methods might become infeasible.

Finally, time and computational resources can become an issue. Especially if only a few computational resources are available (e. g. small GPU and system memory, low CPU performance, etc.), model ensembling techniques such as the BTNet or DRF approach might become infeasible, although the BTNet architecture already greatly reduced the computational effort in direct comparison to the DRF method. Notably, this limitation similarly applies to any other kind of ensembling technique. However, in these situations, especially the use of Bootstrapped Path Shaking might be a promising alternative.

Altogether, Bootstrapped Networks and Deep Random Forests can be seen as an out-of-the-box framework, allowing for an easy-to-use environment without extensive architecture tweaking, providing reasonable results in most scenarios. Due to the improvements with respect to computational effort and final classification performance, whenever applicable, the use of the BTNet approach is recommended.

9.3 Deep Survival Forests

After demonstrating that bootstrapping can be used as part of the model architecture for classification, this part should briefly revisit the topic of survival estimation, which was already discussed in Chapter 7. In the following, a bootstrapped architecture for survival estimation is proposed, well aligning with recent research from other authors.

The *Deep Random Forest* (DRF) framework, which has been discussed in the previous section, is fundamentally based on the random forest (RF) classifier approach. With their so-called *Random Survival Forests* (**RSF**), Ishwaran et al. [2008] have demonstrated that, by introducing slight modifications to the RF approach, bootstrapping can be used for survival prediction. To this end, the randomized decision trees (RDTs) in the RF approach are replaced by so-called *randomized survival trees* (**RST**) as following:

First, similar to the RF classifier, for each of the K survival trees, N samples are drawn from the original data with replacement. Now, each tree is grown by splitting the samples at each node to maximize the survival difference (rather than node purity) of the child nodes using a randomly selected feature. It should be noted, that similarly to the RF approach it is aimed at a binary classification at each node. Moreover, with the exception of the feature selection criterion, the concept is yet identical to it.

As a second modification, the stopping criterion (i. e. reduced node impurity) is replaced by the constraint that each terminal node, i. e. leaf, must contain at least one unique event, i. e. an event of death in case of survival estimation.

The survival trees are trained analogously to the RF concept. Now, for each leaf a *cumulative hazard function* (**CHF**, cf. Chapter 7) $\hat{H}$ can be computed by using the non-parametric Nelson-Aalen estimator [Nelson 1972], reading:

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{\sum_{i:y_i=t_i} \omega_i}{N - \sum_k^N [y_k < t_i]} \tag{9.7}$$

and assuming N individuals $i \in [1,...,N]$ with observation times y_i and event indicators ω_i (i. e. death/no death). More intuitively, the equation can be interpreted as the discrete integral over the hazard function $\hat{h}(t)$, i. e. the chance to die at a given time t_i if an individual is still under consideration to that time.

During inference, given a concrete sample x, the CHF is estimated with each survival tree by walking through the tree analogously to the RF classification process and returning the CHF of the terminal node, i. e. the leaf CHF[10]. Finally, the CHFs are ensembled across trees[11], returning a single CHF function that can be used to calculate a survival expectation.

In order to create *Deep Survival Forests* (**DSFs**), the modifications to the DRF framework are mostly straightforward and correspond to the above modifications of the RF framework. First, the clustering criterion is changed to maximize the survival difference between the two resulting subsets. This can be achieved by two ways: First, by a reformulation as a binary classification problem with newly assigned labels $y' \in \{0,1\}$ derived from survival times $y \in Y$ as:

$$y'_i = \begin{cases} 0 & \text{if } y_i < T \\ 1 & \text{else} \end{cases} \tag{9.8}$$

[10] Notably, this CHF must exist due to the above-set constraint

[11] The CHF ensembling process is equal to the calculation of the CHF on the concatenated original observations of each of the reached leaves. The process is described in more detail in [Ishwaran et al. 2008].

and threshold $T \in Y$ chosen in order to maximize:

$$\Delta = |Y_0| \cdot |Y_1| \cdot \left| \frac{1}{|Y_0|} \sum_{y \in Y_0} y - \frac{1}{|Y_1|} \sum_{y \in Y_1} y \right| \qquad , Y_k = \left\{ y_i \in Y | y'_i = k \right\} \tag{9.9}$$

Using this definition, the node classifiers can be trained using binary crossentropy:

$$\mathscr{L}(y', \hat{y}') = -(y' \log \hat{y}' + (1 - y') \log(1 - \hat{y}')) \tag{9.10}$$

A second option, which follows the notion of the original RF approach more tightly, is to use a feature extraction network such as the sparse autoencoder architecture described in Chapters 5 and 6 to replace the original input at each node by a latent feature representation, which is conditioned on the samples within this node. This allows leaving the maximization of Δ to the RF feature selection. While directly predicting the split better aligns with the presented DRF approach, using the RF feature selection comes with easier implementation and, although yielding comparable results, empirically resulted in higher stability if only a few samples were available[12]. Thus, the second approach will be used in the experimental section.

9.3.1 Experiments

For the evaluation, three different approaches were trained, namely the *Cox proportional hazards* model (CPH, [Cox 1972]), the *Random Survival Forest* (RSF, [Ishwaran et al. 2008]) and the above-proposed *Deep Survival Forest* (DSF). The algorithms were evaluated using the Rossi Criminal Recidivism ($N = 432$) and SEER incidence colorectal cancer datasets ($N = 548,496$) from Chapter 7, the GBSG2 dataset[13] ($N = 686$) from Sauerbrei et al. [1999] and Schumacher et al. [1994], and finally the NSCLC dataset[14][15] ($N = 416$) from Aerts et al. [2015, 2014] and Clark et al. [2013]. While the former three datasets consist of scalar features, the NSCLC dataset is built of lung cancer CT imaging data, and thus needs further processing. For this, again the preprocessing protocol for the LIDC-IDRI data was followed which was already used in Sec. 9.2.4. As the CoxPH and RSF models require scalar input data, radiomics features were extracted by using the pyradiomics-library analogously to Chapter 6, feeding the scalar data to the CoxPH and RSF models, and the raw image data to the DSF approach.

None of the datasets provided a train/test split. Thus, train/test splits have been chosen randomly using an 80/20 split[16]. Each experiment, but on the SEER dataset, has been repeated 5-fold. Repetition was omitted for the SEER dataset, as due to a large number of samples the bootstrapped estimated standard errors were below 10^{-3} when using a single fold.

[12] Using such a feature extraction would be similarly applicable to the DRF approach. However, this was not systematically evaluated, as the framework was reasonably stable in all tested scenarios.

[13] The German Breast Cancer Study Group 2 (GBSG2) dataset consists of 8 covariate columns (age, tumor size, number of breast nodes, ...) and measures the time until reccurence in breast cancer patients.

[14] The NSCLC dataset is one of the original datasets which has been used in the field-defining radiomics study from Aerts et al. [2014]. It contains high.quality CT imaging data from non-small-cell lung cancer (NSCLC) cases and further documents patient survival times.

[15] As briefly described in Sec. 9.2.2, unfortunately the DSF framework could not be evaluated using the mCRC datasets from the previous chapters due to its data use policies.

[16] No validation set was split off as none of the methods requires external validation. While the CoxPH model uses a training-time regularization, the survival forests use an internal validation by OOB-data.

The models were evaluated using Harrel's concordance index, the area under the AUC[17] (AUAUC) and finally mean absolute error (MAE, cf. Appendix C). For the concordance index as well as for fitting the CPH model, the publicly available *lifelines* package [Davidson-Pilon 2019] has been used. In analogy to the discussion in Sec. 9.2.3, the DSF tree depth was limited to 3, accounting for the higher computational effort in comparison to the RSF. Both the DSF and RSF were trained with $K = 10$ trees. Additionally, an RSF with $K = 100$ trees was trained to assess the value of the number of trees.

9.3.2 Results

A comparison of the results for each method and dataset can be found in Fig. 9.9. Overall, the DSF was able to successfully yield survival estimates on the non-imaging datasets (SEER, Rossi, GBSG2), too, being in the same magnitude as the reference approaches, and, in particular, achieving values of up to 0.74 and 0.67 for AUAUC and concordance index. Moreover, it yielded the lowest mean absolute error on the GBSG2 dataset (0.73σ). Overall, however, it was clearly inferior on these datasets, especially when compared with the CoxPH and RSF-100 models. The RSF model overall provided reasonable results. Notably, the number of trees seems to be a highly important determinant of its final performance, as can be seen by the large differences between the results achieved by the RSF-10 and RSF-100 estimators. Regarding the CoxPH model, as initially assumed, it performed clearly best on the largest dataset (SEER). Expectedly, while overall yielding excellent values for concordance, it, however, provided only mediocre absolute survival estimates as measured by the MAE.

In contrast, regarding the medical imaging dataset, the DSF model performed considerably better than the radiomics-based approaches, clearly yielding the highest values for concordance and AUAUC, and similarly providing better than average MAE values. While the RSF-100 model performed second-best with respect to concordance and AUAUC, it provided only a poor fit with respect to the absolute survival time expectation as measured by the MAE. The Cox model provided mediocre results across all metrics on the NSCLC dataset.

9.3.3 Discussion

With the results from the previous section, it was demonstrated that, by using *Deep Survival Forests* (DSF), bootstrapped ensembling can successfully be used for deep survival estimation. In contrast to the CoxPH model or the RSF approach from Ishwaran et al. [2008], DSFs are trained in a purely data-driven fashion and do not require any handcrafted feature engineering. Thus, using DSFs may especially be beneficial if the problem space can only poorly be modeled, or if the major survival determinants are yet unknown.

In direct comparison to the approach from Chapter 7, the DSF provides a variety of benefits. First, following the notion of the well-known and widely applied random forest classifier, the DSF concept is rather easy to understand and involves no complex algebra. Secondly, it provides a clearly better numerical stability, resulting from its RF-based definition, using a threshold function, rather than a multiplicative term. Finally, as the method is based on the bootstrapping methodology, it is inherently able to provide an

[17] The area under the AUC is defined as the integral over the time-continuous binary application of the AUC and is described in more detail in Appendix C.

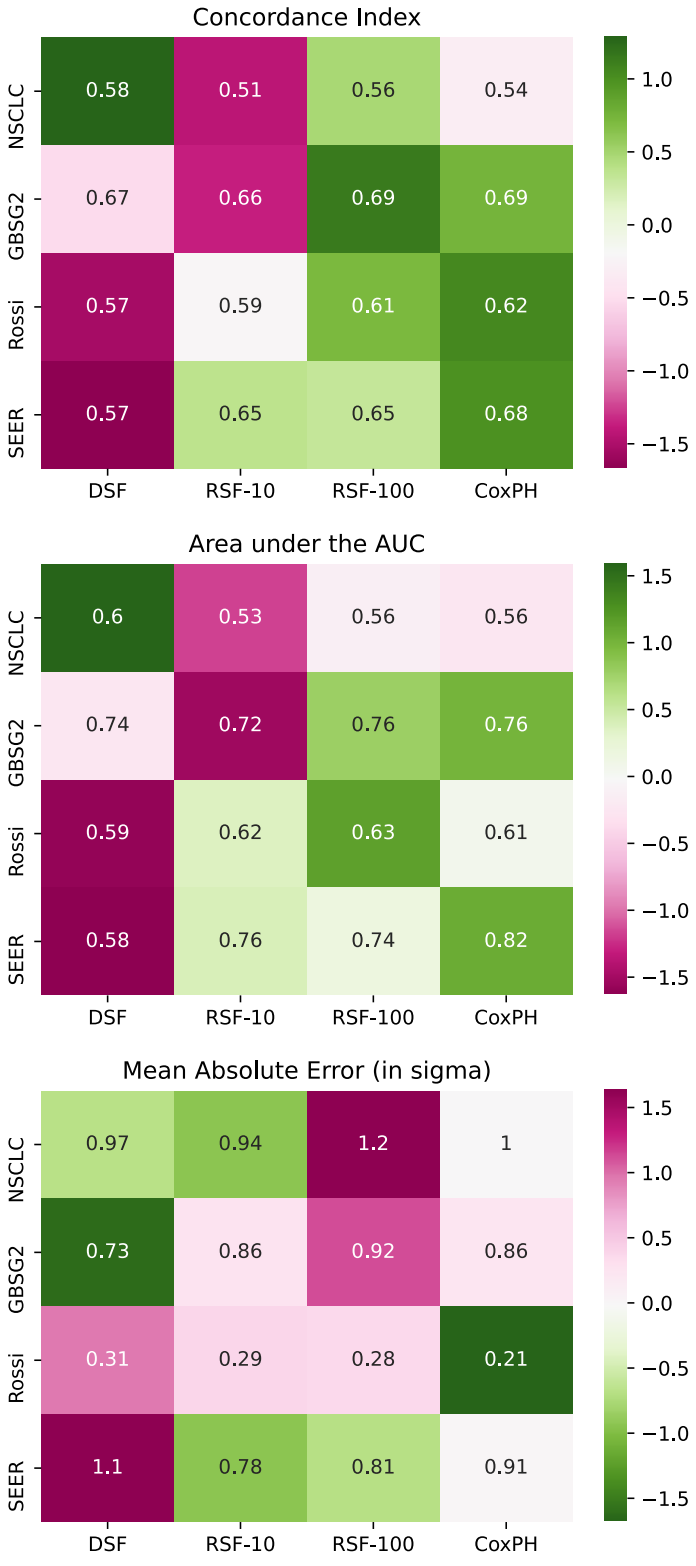


Figure 9.9: Comparison of the Cox proportional hazards model (CoxPH), the Random Survival Forest with 10/100 trees (RSF-10/RSF-100) and the Deep Survival Forest (DSF) approach on the Rossi Criminal Recidivism (Rossi), GBSG2, SEER incidence colorectal cancer (SEER) and NSCLC dataset with respect to the concordance index (top), the area under the AUC (middle) and the mean absolute error (in standard deviations, bottom). For the color encoding (green is better), the results were z-normalized per dataset to account for varying dataset difficulty. The original values are annotated per cell.

estimate of its performance without the need for an additional validation set by using OOB samples. Thus, a larger amount of data can be used for training, which can ultimately result in a superior performance itself.

In comparison to the CoxPH model, it has to be noted that the DSF results were inferior on the non-imaging datasets. To a significant degree, this can likely be attributed to the number of trained survival trees, as on average the DSFs performed on par with RSFs when trained with the same number of trees. In contrast, when increasing the number of trees, the RSF showed major improvements, which is similarly expected for the DSF approach, as a higher number of trees allows for a better-adapted cumulative hazard function, and thus an improved survival estimate. Therefore, future work should in particular address the influence of the number of trees on the final performance of the DSF approach.

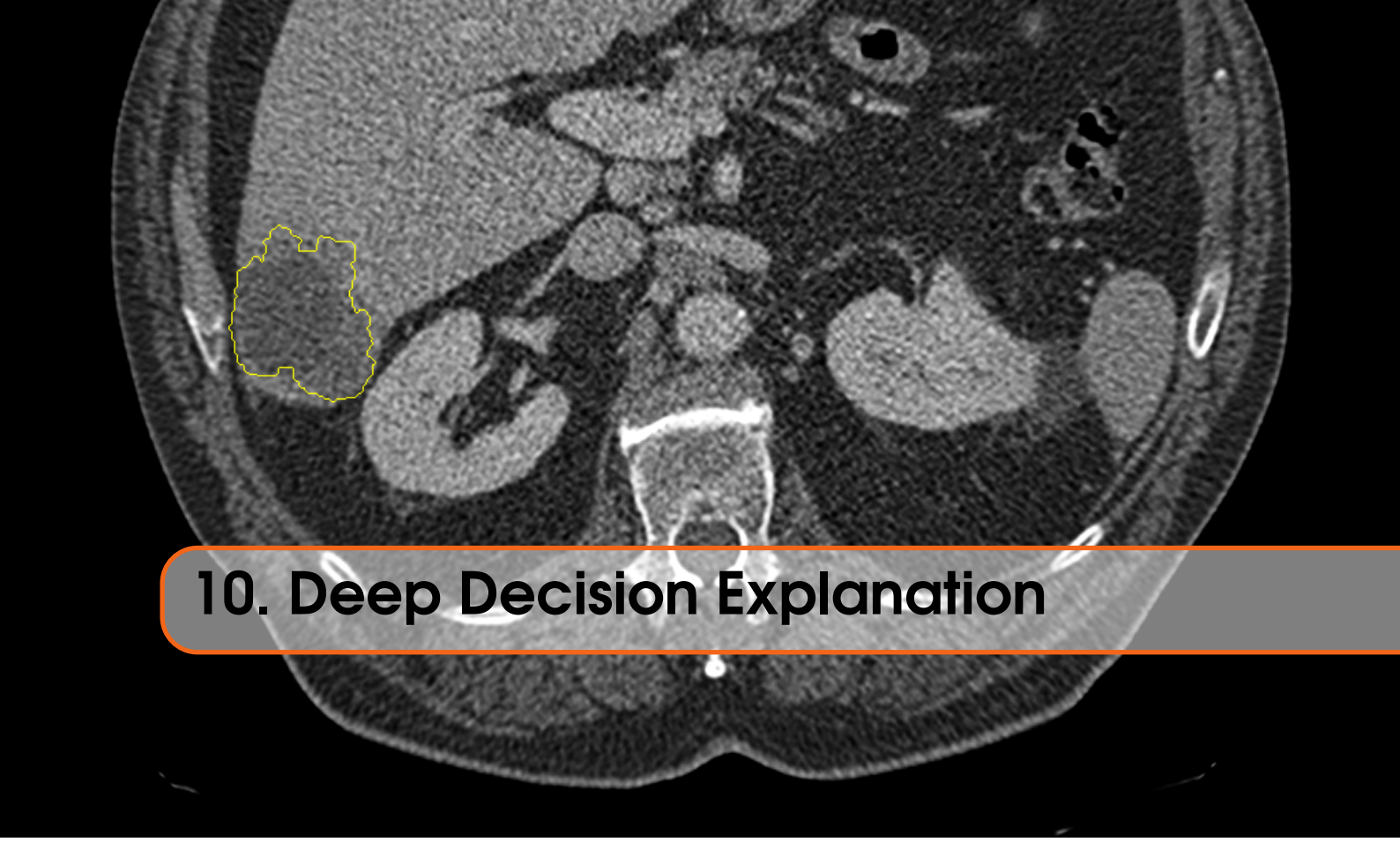
A major limitation of this study is the choice of evaluated datasets. While all of them are publicly available and widely used, additional radiological datasets would have been highly interesting for assessing the capabilities of the DSF approach. Future work should therefore in particular conduct additional experiments using further radiological datasets.

Finally, it has to be noted that, as an ensembling technique, DSFs underlie the same limitations regarding computational complexity and memory requirements as the Deep Random Forest approach (cf. Sec. 9.2.6). Their use is therefore especially encouraged for small- to medium-sized datasets, assuming a reasonably high computational capability for the problem at hand. Although yet no such extension has been proposed, similar to the reduction of computational complexity using BTNets rather than DRFs, the method is expected to have a strong potential for further optimizations.

9.4 Conclusion

Within this chapter, a variety of methods have been presented which combine deep neural networks with the statistical method of bootstrapping and the well-known principles of the widely used random forest classifier. As demonstrated, bootstrapping methods may serve as an excellent regularization in deep neural networks. They are widely applicable, can be implemented at multiple granularities, and can finally especially improve performance on small- and medium-sized datasets while showing strong robustness against label noise.

The bootstrapping methodology has been shown to be applicable to multiple problems, such as image classification and survival time estimation, and, in particular, in the classification setting clearly outperformed other regularization methods, providing a significantly improved test time performance. With this, bootstrapped methods address a variety of obstacles that arise from an application of deep neural networks in medical image analysis, often suffering from small and sparsely annotated imaging data. It is hoped that with this contribution the scientific community is successfully supported on its way towards a better and easier development and deployment of algorithms for medical image analysis, and thus finally for improved overall healthcare.



10. Deep Decision Explanation

As already pointed out in Chapter 4, the two major shortcomings of deep neural networks for medical image analysis are their unsure accuracy when applied to small datasets, as well as their generally low comprehensibility. While the previous chapters have intensively discussed the former point, this chapter is specifically dedicated to the comprehensibility of deep neural networks. For understandable reasons, practitioners are hesitant to adopt methods whose decision process is rather complicated to comprehend, as they impede the identification of error cases, and therefore the understanding of whether the method is applicable to the case at hand, or not. A variety of potential applications of deep neural networks go along with a manageable risk in case of algorithmic mistakes. However, wrong decisions in the medical sector, especially when employed for therapy recommendations, can have disastrous consequences, possibly even leading to the patient's death. It is therefore of crucial importance to create reliable, as well as traceable algorithms, which are not only able to identify error cases (cf. Chapter 8), but also know and actively communicate about their own reliability given a specific case. A variety of methods for deep decision explanation, therefore, have been proposed in the past, with each of them having its own pros and cons. Therefore, this section covers their application to deep decision explanation in clinical decision support, and furthermore, based on the work from [Katzmann et al. 2021], proposes a novel method with significantly improved explanatory quality, as will be demonstrated both quantitatively and qualitatively using a comprehensive user study.

10.1 Introduction

The relevance of deep decision explanation for medical image analysis becomes clear when having a look at a concrete example. In [Katzmann et al. 2018a] (see Chapter 6), the authors proposed the application of the guided backpropagation algorithm from [Simonyan et al. 2013] for the visualization of regions indicative of lesion growth and expected death within one year from scan date.

Given a sufficiently well-trained classifier, such as the one proposed in Chapter 6, it is possible to identify decision-relevant input regions to facilitate an understanding of whether the network focuses on the right image regions, which in turn allows concluding whether the estimator's prediction can be reliable given a concrete case. Moreover, as the estimator is purely based on data-driven training, the visualization might even allow gaining greater insight into typical disease phenotypes, and might ultimately lead to the identification of novel imaging biomarkers and thus to improved patient healthcare. An example visualization is depicted in Fig. 10.1, highlighting relevant image regions of segmented mCRC liver metastases.

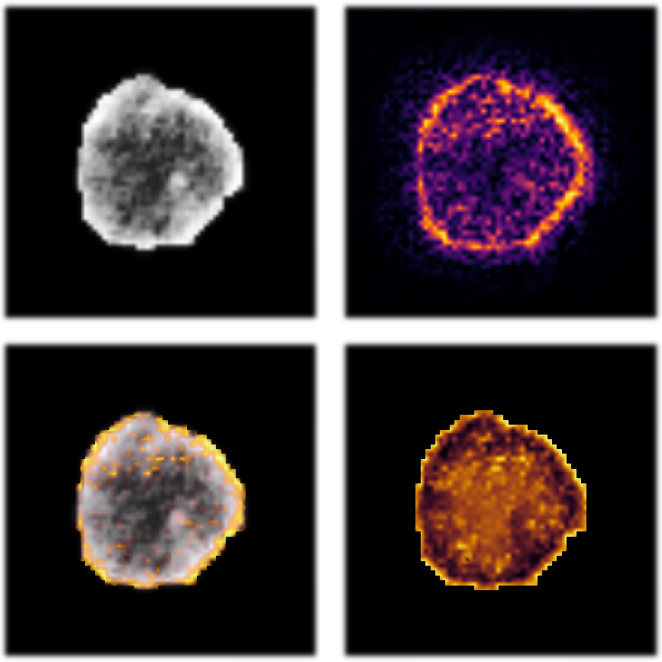


Figure 10.1: Saliency map-based reasoning of classifier decisions from [Katzmann et al. 2018a] using the decision visualization algorithm from [Simonyan et al. 2013]. Lesion image (top-left) and overlay visualization (bottom-left) of input regions which are predictive for future tumor growth. Expectedly, lesion marginalization and contrast enhancement are important predictors for future progress. However, inner structure is marked as predictive, too. The raw saliency map (top-right) highlights the importance of inner structure, with only the most inner necrosis being non-predictive for growth, as well as (bottom-right) the saliency map adjusted by the influence of contrast enhancement, again highlighting the value of inner structure for prediction. Source: [Katzmann et al. 2018a]

10.1.1 Theoretical Considerations

Despite the deterministic nature of deep neural networks, decision explanation is not only non-trivial but actually already theoretically limited. Let us consider a deep neural network which realizes a function $y = f(x) : \mathbb{R}^n \rightarrow \mathbb{R}^m$, with n being the input dimensionality of images X, and m representing the number of output neurons for outputs Y, typically following the relationship $n \gg m$. The task of decision explanation $g(y)$ thus is an inversion of this function: $g = f^{-1} : \mathbb{R}^m \rightarrow \mathbb{R}^n$. Unambiguous decision explanations can exactly

Figure 10.2: Left: An image of the Uros House Armory Show by Grimanesa Amoros, 2011, at the Times Square in New York City. Right: Possible explanations for a classifier deciding whether an image shows a pedestrian. A decision explanation might either highlight the pedestrians (top), their shadows (middle), or even the Times Square itself (bottom). Source: https://commons.wikimedia.org/wiki/File:Grimanesa-amoros-uros-house-day-1.jpg ; Licensed under CC BY-SA 4.0, https://creativecommons.org/licenses/by-sa/4.0/

then be given if f and g both realize an injective projection $f : X \to Y, g : Y \to X$, and thus together allow for a bijective projection $X \leftrightarrow Y$ (cf. Schröder-Bernstein theorem).

In the discrete case, i. e. the computationally representable, $|L^n| > |L^m|$ holds true for any $n > m$ and every discrete number of representable states L, thus enforcing collisions, or explanatory ambiguities, in any real-world scenario.

In fact, this is a desirable property: First, it should be noted that *the reduction process itself is an immediate classification goal*, i. e. samples of similar labels should be represented in nearby or identical latent space positions. Secondly, only a subset of possible explanations is desirable from a human point of view. Fig. 10.2 (left) shows an image of the Times Square in 2011. Let us assume a classifier C taking an image X as input and returning a label Y, stating whether an image shows pedestrians or not. If given the image in Fig. 10.2, a classifier that is trained based on associative rather than causal learning, e. g. by employing stochastic gradient descent, will typically utilize a combination of various classification strategies to find a solution, e. g.:

A The image contains typical shapes of pedestrians, including silhouette, face, and clothes. $\to$ The image contains pedestrians.

B The image contains shadows with typical silhouettes, implying two legs and an appended upper body. → The image contains pedestrians.

C The image shows the Times Square, which is typically highly frequented. Taking into account the color of the sky, the image was taken during the daytime. → The image contains pedestrians.

Each of these strategies in itself *is valid*, and without further constraints, none of them can be expected to be ruled out by the classifier training process. However, while neither of these strategies is wrong, not all of them are similarly preferable with respect to the final application scenario.

Decision visualization is often conducted as a guided back-projection of the classifier decision towards the previous layers. As intermediate layers in deep learning do not necessarily have an intuitive representation due to the distributed computing scheme and the activation nonlinearities, the process is repeated layer-by-layer until the input layer is reached, which in turn can be understood intuitively. Now taking into account the above scenario, the resulting visualization might emphasize strongly different image components (cf. Fig. 10.2):

- In the case of A, the result would look rather intuitive. The decision explanation would highlight all visible pedestrians, well-aligning with the human intuition.

- For B, the algorithm would highlight human shadows. Although not necessarily matching a typical human-kind explanation, humans in fact tend to have a characteristic shadow, with studies demonstrating that unconscious shadow perception provides a major clue for scene understanding in humans, too [Mamassian et al. 1998].

- Finally, C is mostly based on a logical argument: Assuming a recognition of the Times Square, highlighting the sky and the place in fact makes sense for deriving the true answer. Obviously, this is not the decision strategy the classifier is expected to have. However, in fact, this behavior does well align with subconscious human reasoning when given only a limited amount of information, as humans then tend to employ a significant amount of heuristics, based on their previous experiences (cf. [Johnson-Laird et al. 1999; Kelley 1973]).

Therefore, it becomes clear that an adequate visualization is one that **a)** matches the actual image semantics, **b)** aligns to the intuitive understanding of a human observer, and **c)** should naturally provide an image quality which allows for a visual inspection. In a nutshell, ***the perfect explanation is the one which is best understood by a human observer***, highlighting the outstanding importance of human perception when evaluating decision explanation algorithms.

10.1.2 Related Work

Various methods have been developed to support human insight into deep decision-making, and multiple categorizations for them have been proposed. In [Linardatos et al. 2021]

a categorization based on four key dimensions is suggested, including a) globality vs. locality, b) explanation purpose (e. g. black-box explanation), c) model-specificity and d) input modality. Within this categorization, we will focus on local black-box explanations for image input. Amongst these, a method is considered *model-specific* if it requires a particular model architecture or is part of the architecture itself, and *model-agnostic* if it is applicable to any network architecture.

The former of these groups consist of algorithms such as *(Grad)CAM(++)* [Chattopadhay et al. 2018; Selvaraju et al. 2017; Zhou et al. 2016], *DeconvNet* [Noh et al. 2015], or *Attention-gated networks* [Schlemper et al. 2019]. Their common idea is to modify the network architecture in such a way that a successful classification requires a localization of the object, e. g. by predicting additional maps which are used to mask the pre-output activations. However, these methods typically result in blob-like highlightings, which were shown to not neccessarily reflect the relevant image structures [Chattopadhay et al. 2018; Schlemper et al. 2019]. With *PAWS* and *DINO*, Assran et al. [2021] and Caron et al. [2021] have impressively shown that a combination of multiple attention layers in fact can lead to an impressive explanatory performance. However, both were trained with datasets of sizes far beyond typical data amounts in medical imaging scenarios.

Similarly, a variety of *model-agnostic* approaches has been proposed, comprising methods such as *LRP* [Bach et al. 2015], *DeepLIFT* [Shrikumar et al. 2017], *DeepTaylor* [Montavon et al. 2017], and *SHAP* [Lundberg et al. 2017]. As pointed out by Lundberg et al. [2017], these methods utilize a common mechanism called *additive feature attribution*, starting with a fixed amount of *decision relevance* at the model output, and splitting it by layer-wise propagation through the model until the input layer is reached. As each of the methods comprises different relevance splitting constraints, they may result in different visualization qualities, with user studies indicating the most intuitive results for *DeepSHAP*, a deep learning-inspired approximation of SHAP [Lundberg et al. 2017]. Notably, all these methods precisely follow the weights of the network. As a result of this, the image quality might become poor and noisy if the classifier was trained on only a few data (cf. [Adebayo et al. 2018; Kim et al. 2019]). Furthermore, the approaches were shown to behave sensitively against minor input modifications and adversarial attacks [Ghorbani et al. 2019; Kindermans et al. 2019]. Other approaches from this domain include *LIME* [Ribeiro et al. 2016], *Anchors* [Ribeiro et al. 2018] and *LORE* [Guidotti et al. 2018], which employ local proxy models for which the adequacy for the problem and data at hand is not guaranteed (cf. [Li et al. 2020]).

Notably, there has been recent research in the field of "visual analytics", such as the work from Pezzotti et al. [2017] and Choo et al. [2018]. Visual analytics can be understood as a framework for network design and processing understanding. Visual analytics especially targets network designers, allowing them to visually inspect the activity within the model by depicting network flow, latent space embeddings, samples of highest activation, convolutional filters, or gradients, and aim to make the training process better understandable. Additionally, with the work from Ming et al. [2018] an approach exists that targets the derivation of understandable rules from a fully trained model through an interactive process, yielding a highly-intuitive ruleset, which can then further be used as an understandable and easily comprehensible model itself. Each of these frameworks, however, requires active user interaction and, in turn, cannot readily be used as an out-of-the-box explanation approach.

A last relevant group can be considered a sub-group of model-agnostic methods but differs in that way that it employs an inversion of the problem by creating (semi-)synthetic images, which both highlight the relevant regions, simultaneously trying to maximize image realism and quality, as well as align to the given input image. A first approach into this direction was *activation maximization* (**AM**), which maximized class activations by an optimization on the input image [Erhan et al. 2009; Simonyan et al. 2013]. As AM tends to create unrealistic images, later approaches such as the work from Nguyen et al. [2016], Liu et al. [2019] and Singla et al. [2019] have combined AM with *generative adversarial networks* (**GANs**, [Goodfellow et al. 2020]) to improve image realism, but were trained on significantly larger datasets for achieving a good explanatory quality.

Thus, while recent work has given an enormous impetus to the field of *explainable artificial intelligence* (XAI), and strongly contributed to the understandability of deep neural networks, it was not yet possible to satisfactorily address the particular issues which arise from an application of deep neural networks in medical imaging scenarios.

10.2 Deep Decision Explanation using Cycle-GANs

To specifically address the aforementioned issues, with [Katzmann et al. 2021] a novel model architecture for deep decision explanation has been proposed, based on a combination of activation maximization and *Cycle-consistent Generative Adversarial Networks* (**CycleGANs**, [Zhu et al. 2017]), and is capable of visualizing intuitive, high-quality decision explanations for trained neural classifiers. The following sections will give an overview of the method, its components, and their interactions. It will show up the benefits which arise from this combination, and demonstrate its value both quantitatively and qualitatively. Finally, a thorough discussion on the limitations of the study at hand is given.

10.2.1 Method

The proposed method is fundamentally based on the idea of combining a GAN-based architecture with AM in order to yield a higher image quality. It thus well aligns with the aforementioned work from Liu et al. [2019] and Singla et al. [2019], and following the notion above can be categorized as an image synthesis-based approach. In both mentioned publications, GANs have been used to improve the result of AM by increasing the image realism, thus addressing its most relevant shortcoming. GANs, however, tend to produce low-quality results themselves if trained on only a few data (cf. [Gurumurthy et al. 2017]), failing to sufficiently address the medical imaging scenario. Thus, as described above, our method utilizes a derived architecture called CycleGANs. CycleGANs have lately caught attention for unpaired high-quality image-to-image translations (see Fig. 10.3), a mechanism which we will make use of in our scenario.

Architecture

As mentioned above, the decision explanation architecture is based on CycleGANs, which have derived from GANs. Fig. 10.4 (left) depicts the general architecture of a GAN. GANs work by employing a combination of two neural networks, a generator and a discriminator, which are trained alternately. The generator is trained to create synthetic images of a target domain X. The discriminator in turn is trained to distinguish *fake*, i. e. synthesized, from *real* images. Both networks are appended in such a way that ultimately the discriminator

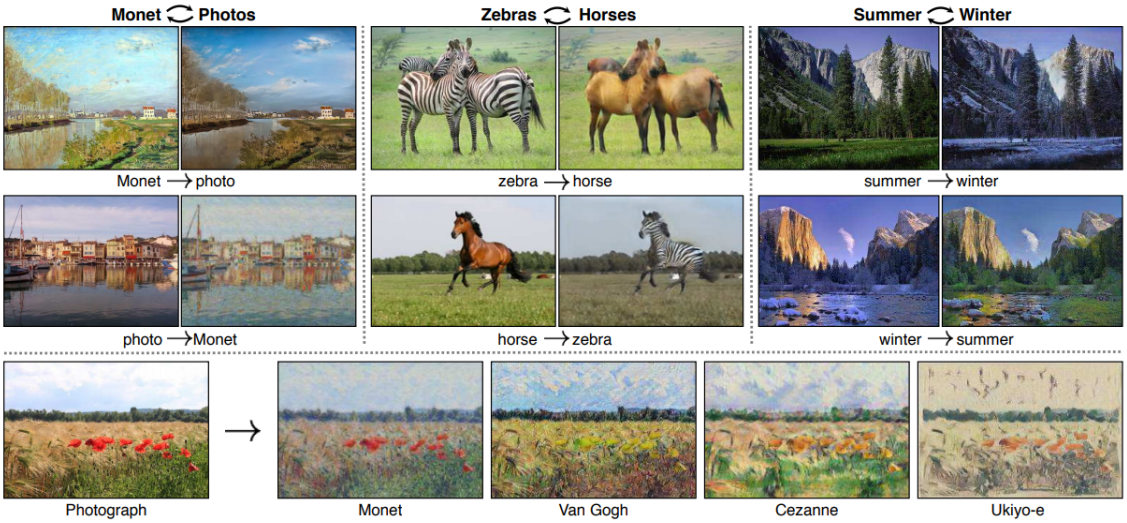


Figure 10.3: CycleGANs are capable of creating high-quality image-domain transfers using only unpaired image data, as exemplarly shown for conversions between photo and art, horses and zebras, and summer and winter images, as well as for multiple painting style transfers. Original source: [Zhu et al. 2017] © 2017 IEEE

yields a gradient for the generator to optimize the synthetic image generation process towards a higher image realism, resulting in a zero-sum game-like training behavior with continuously improving image quality. GANs, however, tend to show specific problems when trained on only small datasets, such as mode collapse[1] and catastrophic forgetting[2] [Thanh-Tung et al. 2020].

In contrast, CycleGANs are based on a combination of two GANs which are cyclically connected (see Fig. 10.4, right). Each training step always involves a full cycle through the network. As a result, each generator/discriminator pair serves as a regularization for the respective other, using a mechanism called *cycle-consistency* (see below). As a result, CycleGANs can be trained on smaller datasets, avoiding the shortcomings of catastrophic forgetting and mode collapse addressed above, while yielding a reasonable performance.

For the method at hand, the idea is now to train a CycleGAN to generate two slightly modified versions of the original image, each of them maximizing one of the output class probabilities of the classifier to be explained (see Fig. 10.5), followed by a visualization of the difference of these modified images. As will be shown, this difference can be used as a direct measure of decision relevance, and well-fitting human intuition.

Cycle-consistency

A major benefit of CycleGANs lies in their generator's behavior to constrain each other, achieved by enforcing the eponymous cycle-consistency. Assuming generators G^+ and G^-, cycle-consistency denotes the property that a subsequent application of both generators on any given image $x \in X$ should yield the original image, i. e.: $G^+(G^-(x)) \approx G^-(G^+(x)) \approx x$, which can be reached by introducing an additional loss term. While various choices would

[1] Mode collapse is defined as being a failure state of the generator network in which it produces one or only a few different images, so-called *modes*.

[2] Catastrophic forgetting denotes a situation in which the quality of the generated solutions becomes worse in the later training, with the network effectively "forgetting" the previous learning progress.

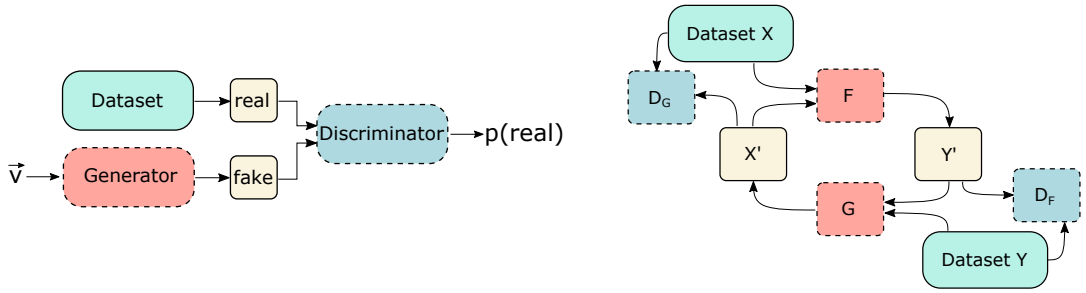


Figure 10.4: **Left:** A visualization of a GAN, transforming random vectors $\vec{v}$ into *fake* samples using a generator. A discriminator tries to identify *fake* vs. *real* samples. The networks are trained alternately, with the generator improving by using the gradient from the discriminator with inversed labels. **Right:** The CycleGAN architecture from [Zhu et al. 2017], consisting of two generator-discriminator-pairs $(F, D_F), (G, D_G)$. Generators F, G are trained to mimic samples X, Y by creating fake samples X', Y'. Analogously, discriminators D_F, D_G are trained to recognize fake samples, effectively leading to generators F, G realizing a domain transfer between domains X and Y. Source: [Katzmann et al. 2021]

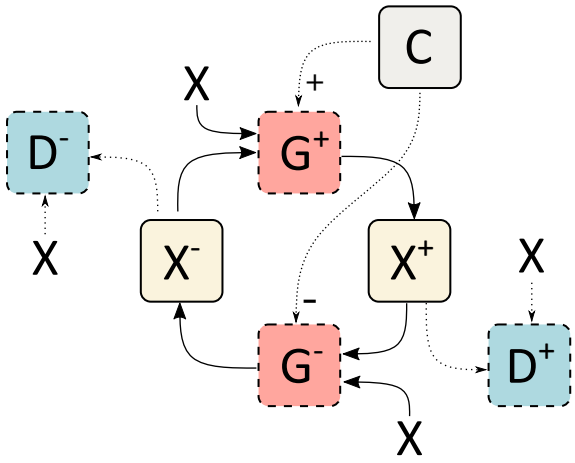


Figure 10.5: Overview on the architecture of the proposed decision explanation, consisting of two generator/discriminator pairs, for consistency with their aim being denoted as (G^+, D^+) and (G^-, D^-), each maximizing one of the output class probabilities of the classifier to be explained C by creating slightly modified versions X^+, X^- of the original images X. In contrast to the original CycleGAN approach (see Fig. 10.4), only a single image domain is used. Notably, the process is fully guided by the response of the classifier C. No ground truth labels are used at any time. After training, the differences between X^+ and X^- can be used as a measure of decision relevance. Source: [Katzmann et al. 2021]

be reasonable for this, in [Katzmann et al. 2021] it was decided to use the multiscale structural dissimilarity loss from Isola et al. [2017], being an extension of the structural similarity, which takes into account luminance, contrast and image structure, and thus well fits human perception [Wang et al. 2003]:

$$\mathscr{L}_{\text{DSSIM}}(x,y) = 1 - \frac{(2\mu_x\mu_y + c_1)\cdot(2\sigma_{xy} + c_2)}{2\cdot(\mu_x^2 + \mu_y^2 + c_1)\cdot(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{10.1}$$

with means μ_x, μ_y, standard deviations σ_x, σ_y, covariance σ_{xy}, stabilizing parameters $c_i = (k_i \cdot L)^2$, $k_1 = .01$, $k_2 = .03$ and a dynamic range parameter of $L = 1$ as suggested by the original work. Moreover, an L1-term is added to take into account the average image intensity, yielding:

$$\mathscr{L}_{\text{cycle}}(x,x') = \left|x - x'\right| + \mathscr{L}_{\text{DSSIM}}(x,x') \tag{10.2}$$

Patch-GAN architecture

The architecture of the generators G^+ and G^- was based on the U-Net from Ronneberger et al. [2015]. For the discriminators it was decided to use the PatchGAN approach from [Isola et al. 2017]. PatchGANs differ from normal GANs in the way that the discriminators do not yield a single binary classification, i. e. real vs. false, as it would result from a ResNet [He et al. 2016] or ConvNet [LeCun et al. 1989] approach, but that the final classification layers are "cut off" from the network, with the classification being done per patch on the 2-dimensional likelihood maps. The full architectures can be found in the Appendix Fig. B.5 and B.6. Using PatchGAN-generators prevents typical issues of GANs, such as the aforementioned mode-collapse problem, and especially yields a higher image quality [Isola et al. 2017].

To distinguish *false* from *real* images, the discriminators are trained by applying patch-wise binary crossentropy, reading:

$$\mathscr{L}_{\text{CE}}(y,\hat{y}) = y \cdot \log(\hat{y}) + (1-y)\cdot\log(1-\hat{y}) \tag{10.3}$$

with y and $\hat{y}$ indicating the label whether a sample is real and the discriminator's estimate of it, respectively.

Activation Maximization

As pointed out above, the concept is based on the maximization of the output activations of the classifier C to be explained. To this end, each generator is trained including a loss term for activation maximization on the classifier. For this, the cross-entropy equation from above is reused:

$$\mathscr{L}_{\text{AM}}(l,\hat{l}) = \mathscr{L}_{\text{CE}}(l,\hat{l}) \tag{10.4}$$

with l indicating the class to be maximized by the generator, i. e. $l = 1$ for G^+, and $l = 0$ for G^-, and $\hat{l}$ being the estimate of the classifier for sample x^+/x^-. Notably, *the ground truth label of the sample is used at no time*. To enforce that changes are only introduced if needed for modifying the classifier's class activation, an additional loss term $\mathscr{L}_{\text{sim}}$ is added to the optimization for the generators G^+ and G^-, following the notion of $\mathscr{L}_{\text{DSSIM}}$, i. e. $\mathscr{L}_{\text{sim}}(x, G^+(x))$ for G^+ and $\mathscr{L}_{\text{sim}}(x, G^-(x))$ for G^-.

Overview

Combining the definitions from above, the loss function reads:

$$\begin{aligned}\mathscr{L}_{\text{gen}}(x_a,x_b,l,G^+,G^-,D^+,C) = &\mathscr{L}_{\text{cycle}}(x_a,G^-(G^+(x_a))) \quad + \mathscr{L}_{\text{sim.}}(x_a,G^+(x_a)) \; + \\ &\mathscr{L}_{\text{CE}}(D^+(G^+(x_a),x_b),0) \quad + \mathscr{L}_{\text{AM}}(l,C(x_a))\end{aligned} \tag{10.5}$$

with images $x_a, x_b \in X$ from classifier image domain X, label l to be maximized by the generator G^+, its discriminator D^+, and its counterpart generator G^-. Both generator/discriminator pairs are trained alternately, inverting $+$ and $-$ for G^-, D^-. The process is repeated until convergence. As pointed out in [Katzmann et al. 2021], this definition trains the generators to:

1. maximize the class activation of the class assigned to each generator (**AM loss**) by introducing only slight changes (**similarity loss**),
2. generate images which are indistinguishable from real images (**discriminator loss**),
3. be cycle-consistent (**cycle-consistency loss**)

10.2.2 Relevancy Visualization

Taking an input image $x \in X$ and following the procedure above, the generators G^+, G^- create modified images x^+, x^-, each maximizing one of the classifier output class activations, by introducing as few changes as possible (cf. similarity loss). We define the differences between the original and the generated images as:

$$\Delta^+ = \Delta(G^+(x),x), \qquad \Delta^- = \Delta(G^-(x),x) \tag{10.6}$$

in which Δ denotes a local difference metric, such as pixel-wise L1-difference or patch-wise structural dissimilarity. The choice of the metric may depend on the problem at hand, e. g. whether there should be a focus on intensity or structural modifications. The resulting difference images Δ^+, Δ^- now reflect changes that are needed to maximize the positive or negative class. Thus, within the given context, they indicate decision relevance *for the respective other class*. Now, their difference R as given by:

$$\begin{aligned}R &= \Delta(G^-(x),x) - \Delta(G^+(x),x) \\ &= \Delta^- - \Delta^+\end{aligned} \tag{10.7}$$

fulfills the properties of overall decision relevance: It shows low magnitudes in areas with few introduced changes, as indicated by Δ^+ and Δ^-, i. e. being neither decision-relevant for one nor the other class. Similarly, it shows low magnitudes in areas in which equally strong modifications would be required to change the class activations of the classifier to be explained within the given context. In contrast, if a change only has to be introduced to maximize one of the classes but not the other, this results in high values for R, indicating decision relevance for the respective other class.

10.3 Experiments

For the evaluation, three different publicly available datasets have been employed, namely LIDC-IDRI [Armato et al. 2011; Armato III et al. 2015; Clark et al. 2013], BreastMNIST

[Yang et al. 2021] and CIFAR-10 [Krizhevsky et al. 2009], each representing a different imaging modality. For multiple classifier architectures, the decision explanation approach was compared to three state-of-the-art (SoA) algorithms, namely DeepSHAP [Lundberg et al. 2017], DeepTaylor [Montavon et al. 2017] and LRP [Bach et al. 2015]. For each of these methods, the publicly available reference implementations from Lundberg et al. [2017] (DeepSHAP) and Alber et al. [2019] (DeepTaylor, LRP) were used.

The evaluation was done in two steps: First, a quantitative evaluation is conducted, assessing the performance of the classifier to be explained using various classification metrics, and testing whether the introduced model architecture actually modifies the classifier's class output probabilities. Secondly, a double-blind, thorough user study is conducted, comparing the method to the results of the SoA algorithms, and evaluating a total of 9,792 answers, including 34 different participants. Statistical tests were conducted using two-tailed t-tests with $t(N-2)$, confidence intervals were computed using bootstrapping [Efron 1987].

10.3.1 Datasets

LIDC-IDRI

For the first study, the well-known LIDC-IDRI dataset has been used, containing a total of 1,018 low-dose lung screening CT images, in combination with a ConvNet as proposed in Chapter 9. The data was preprocessed according to the protocol from [Nibali et al. 2017], yielding 772 lung lesion images (348 malignant/423 benign), each being extracted at a resolution of 64x64 pixels representing a 45x45 mm window centered around the lesion. A 70/30 split was used, dividing the data into 537 samples (236/301) for training and validation, and 135 samples (112/123) for testing.

BreastMNIST

The second study was conducted on the BreastMNIST dataset, containing 780 breast ultrasound images with or without malignant lesions. The predefined train/test splits were used, resulting in 624 samples for training and validation (456 malignant/168 benign) and 156 samples for testing (114/42). The images were nearest neighbor-padded to 32x32 pixels. As a classifier, the ResNet reference implementation from [Chollet et al. 2015] was used.

CIFAR-10

Finally, the CIFAR-10 dataset is used to show that the method can also be applied to large RGB imaging datasets. The dataset consists of 50,000 training samples of 32x32 pixel patches from 10 classes (5,000 per class) and an additional test set of 10,000 images (1,000 per class). As the method is restricted to binary classification tasks, the two most difficult one-vs.-one tasks were identified, being cat-vs.-dog and car-vs.-truck (cf. [Katzmann et al. 2021] for more details). For the classification, we used the EfficientNet-B0 implementation from Yakubovskiy [2020] which is based on the work from Tan et al. [2019], and fine-tuned it on the CIFAR-10 dataset.

10.3.2 Quantiative Evaluation

First, each classifier was assessed to ensure a sufficiently high classification performance. Therefore, a variety of metrics was evaluated, comprising accuracy, sensitivity, specificity,

positive and negative predictive value (PPV/NPV), informedness, markedness, Matthews correlation coefficient (MCC), and the area under the curve (AUC)[3]. Secondly, it was evaluated whether an application of the method actually leads to modified class output probabilities for each classifier. Therefore, the class activation shift was assessed for each classifier with samples $G^{-}(x)$, x and $G^{+}(x)$.

10.3.3 User Study

In the user study, the method was double-blindly compared to the three SoA approaches (see Sec. 10.3). For each dataset, 24 samples (12 positive/12 negative) were randomly drawn from the data, and the classifier decisions were visualized by each method. The different explanations were shuffled by a computer program so that no identification of the generating method was possible. The resulting questionnaire was given to multiple participants (8/6/9/11 for LIDC-IDRI, CIFAR-10 trucks-vs.-cars, CIFAR-10 cats-vs.-dogs, and BreastMNIST). For the medical imaging datasets, it was ensured that each participant had working experience with medical imaging data (6.5 years on average). All participants were asked to evaluate three criteria for each sample and method, namely:

a) *intuitive validity* ("Does it look reasonable at first glance?")
b) *semantic meaningfulness* ("Does it make sense with respect to the image context?")
c) *image quality* ("Does it look good?")

yielding 288 questions per questionnaire. Each criterion was assessed by assigning an integral score between -4 and 4. After collecting and reordering using the above-mentioned computer program, the scores were analyzed both raw as well as z-normalized for each sample and criterion (see 10.3.4). Further, the inter-observer reliability was assessed using pairwise Pearson correlation (cf. Appendix C). Finally, the factor structure of the criteria was analyzed using principal component analysis.

10.3.4 Results

Classifier performance

The quantitative evaluation demonstrated a reasonable classification performance for each of the tasks and classifiers. For the medical imaging datasets, the overall accuracy was at .809/.802 for the LIDC-IDRI and BreastMNIST datasets, respectively. The MCC and AUC values were at .617/.809 for LIDC-IDRI, and at .453/.822 for the BreastMNIST classifiers, respectively (see Tab. 10.1). The CIFAR-10 classifiers, based on the EfficientNet architecture, expectedly performed even better, yielding a 10-class MCC of .956, and binary MCCs of .874 and .960 for the cats-vs.-dogs and cars-vs.-trucks subsets, respectively (Tab. 10.2). Thus, all classifiers provided a reasonable performance for training the proposed decision explanation framework.

Domain transfer

After training the architecture, it was assessed whether it successfully realize a domain transfer as was theoretically assumed. As discussed in [Katzmann et al. 2021], this assessment yielded positive results:

[3] All metrics are described in detail in Appendix C

	LIDC-IDRI	**BreastMNIST**
Accuracy	.809 [.757,.860]	.802 [.737,.859]
F1	.801 [.737,.856]	.871 [.824,.913]
Sensitivity	.813 [.737,.884]	.921 [.869,.966]
Specificity	.805 [.729,.866]	.477 [.326,.629]
PPV	.790 [.715,.861]	.827 [.760,.890]
NPV	.826 [.757,.894]	.690 [.515,.851]
Informedness	.618 [.514,.719]	.398 [.239,.557]
Markedness	.616 [.511,.718]	.517 [.329,.691]
MCC	.617 [.511,.718]	.453 [.285,.607]
AUC	.809 [.757,.860]	.822 [.740,.895]

Table 10.1: Classification performance on test set for malignancy classification networks on LIDC-IDRI and BreastMNIST with 95 % CI. Source: [Katzmann et al. 2021]

	CIFAR-10	**Cats-vs.-Dogs**	**Trucks-vs.-Cars**
Accuracy	.960 [.959,.963]	.937 [.926,.947]	.980 [.974,.985]
F1		.937 [.926,.947]	.979 [.974,.985]
Sensitivity		.943 [.929,.957]	.970 [.959,.980]
Specificity		.931 [.914,.946]	.988 [.983,.994]
PPV		.932 [.917,.947]	.988 [.982,.994]
NPV		.942 [.927,.957]	.971 [.960,.980]
Informedness		.874 [.853,.895]	.959 [.948,.970]
Markedness		.874 [.852,.894]	.960 [.948,.970]
MCC	.956 [.954,.959]	.874 [.853,.894]	.960 [.948,.970]
AUC	.978 [.977,.979]	.987 [.984,.991]	.997 [.995,.999]

Table 10.2: Test set performance on CIFAR-10 (overall) as well as on the cats-vs.-dogs and trucks-vs.-cars subsets. For the overall set, binary metrics have been omitted. For the subsets, the first class was encoded as positive. Source: [Katzmann et al. 2021]

	Intuitivity	Semantics	Quality
	LIDC-IDRI		
Ours	**2.84** [1.50, 3.98]	**2.78** [1.39, 4.10]	**2.04** [0.50, 3.41]
DeepSHAP	-0.64* [-2.00, 0.86]	-0.52* [-1.97, 1.05]	-1.02* [-2.12, 0.16]
DeepTaylor	-0.60* [-2.16, 1.10]	-0.71* [-2.23, 1.00]	0.14 [-1.32, 1.57]
LRP	-1.61** [-2.72,-0.33]	-1.55** [-2.73,-0.11]	-1.16* [-2.24,-0.02]
	BreastMNIST		
Ours	**1.66** [0.65, 2.57]	**1.67** [0.63, 2.62]	**1.76** [0.70, 2.71]
DeepSHAP	-1.56** [-2.63, 0.39]	-1.62** [-2.70,-0.43]	-1.38** [-2.47,-0.20]
DeepTaylor	1.11 [-0.26, 2.37]	1.24 [-0.12, 2.49]	0.78 [-0.53, 1.99]
LRP	-1.21** [-2.33, 0.02]	-1.29** [-2.41,-0.07]	-1.15** [-2.29, 0.04]
	CIFAR-10 - cats-vs.-dogs		
Ours	0.83 [-0.47, 2.03]	0.85 [-0.58, 2.13]	0.08 [-1.18, 1.36]
DeepSHAP	-1.25 [-2.60, 0.16]	-1.35 [-2.71, 0.07]	-0.03 [-1.83, 1.69]
DeepTaylor	**2.07** [0.75, 3.12]	**2.29** [0.96, 3.32]	**0.90** [-0.58, 2.28]
LRP	-1.65* [-2.88,-0.32]	-1.79* [-3.07,-0.36]	-0.95 [-2.39, 0.48]
	CIFAR-10 - cars-vs.-trucks		
Ours	1.10 [-0.28, 2.30]	1.55 [0.33, 2.54]	0.77 [-0.67, 2.10]
DeepSHAP	-1.64* [-3.30, 0.18]	-1.91* [-3.54,-0.04]	-1.22* [-3.01, 0.67]
DeepTaylor	**2.58** [1.25, 3.58]	**2.76** [1.53, 3.60]	**2.03** [0.31, 3.51]
LRP	-2.05* [-3.65,-0.18]	-2.40** [-3.78,-0.72]	-1.58* [-3.27, 0.30]

Table 10.3: Average z-adjusted questionnaire results (higher is better) per algorithm on the LIDC-IDRI, BreastMNIST and CIFAR-10 cats-vs.-dogs and cars-vs.-trucks datasets with 95 % CI. p-values for two-tailed t-test are indicated as $p < .05$*, $p < .01$**, $p < .001$***. The best result is marked bold. Modified tables based on [Katzmann et al. 2021]. The full tables are found in the Appendix B.5–B.6.

For the LIDC-IDRI dataset, the average malignancy rating was at .470, 95 % CI [.418,.522] on the original data x. After the positive transfer $G^+(x)$ it increased to .820 [.788, .848], and after negative transfer $G^-(x)$ it reduced to .289 [.250,.330]. Similarly, for the BreastMNIST dataset, the average malignancy rating changed from .656 [.604,.707] to .871 [.835,.904] and .337 [.294,.383], and for the CIFAR-10 datasets from .504 [.482,.525]/.491 [.469,.513] before, to .920 [.911/.930]/.863 [.850,.876] after postive and .152 [.138,.165]/.179 [.165,.194] after negative transfer, for cats-vs-dogs and cars-vs-trucks, respectively.

By using a paired t-test, each of these transfers was shown to be highly significant with $10.2 < t(|X|-2) < 33.3$, and all $p \ll 10^{-5}$. Thus, the domain transfer as proposed in the framework in fact led to a significant maximization of the classifier class output probabilities.

User Study

The z-adjusted questionnaire results of the user study can be found in Tab. 10.3. The detailed results including raw questionnaire results and rank evaluation can be found in the Appendix Tabs. B.5–B.8. Figs. 10.6–10.8 depict qualitative comparisons between our method, DeepSHAP, DeepTaylor and LRP for the LIDC-IDRI, CIFAR-10 and BreastMNIST datasets, respectively. Notably, on both medical imaging datasets the approach outperformed the SoA approaches, clearly yielding the highest z-adjusted (and raw, see Appendix) scores. It ranked best across all tested criteria with average ranks of 1.28, 1.36, and 1.52 for intuitive validity, semantical meaningfulness, and image quality, respectively, on the LIDC-IDRI, and 1.75, 1.77, and 1.64 on the BreastMNIST dataset (cf. Appendix Tabs. B.5–B.6). For the significantly larger CIFAR-10 datasets, leading to a better-adapted classifier, however, the DeepTaylor algorithm was superior (n.s.), with the presented algorithm placing second.

The inter-observer reliabilities on the medical datasets were in the range $.386 \leq \rho \leq .639$, implying a fair to substantial agreement (cf. [Landis et al. 1977]). For the CIFAR-10 subsets, although the reliabilities for intuitivity and semantic meaningfulness were in the range $.61 - .85$, they were unexpectedly low for image quality with $.09 - .20$, forming two clusters with one preferring color overlays and the other preferring rather parsimonious explanations. Depending on the dataset, the criteria showed correlations between $r = .51 - .90$ (cf. Appendix Fig. B.1), with the strongest correlations forming between intuitive validity and semantic meaningfulness. Using PCA, it was thus possible to extract an overall preferability-factor Φ which accounted for 77-88 % of the observed variance (Fig. 10.9), showing a comparable structure across all datasets and highlighting the importance of each of the measured criteria.

10.4 Discussion

As discussed in previous chapters, visualizing network decisions is a crucial step toward a clinical application of deep neural networks, as it allows both the clinician as well as the developer to gain insight into the strategies and shortcomings of trained networks. As was successfully reproduced with the study at hand, previous methods already perform reasonably well when applied to classifiers that were trained on large and comprehensive datasets. Similarly, it could be shown that SoA methods fail to provide a reasonable explanation if data is scarce or if the decision process contains more ambiguities.

In turn, it was possible to demonstrate that an architecture based on cycle-consistent activation maximization is able to produce decision explanations of much higher quality, clearly outperforming recent SoA approaches. With this, it is hoped that a significant contribution towards the understanding of deep learning-based clinical decision support could be made.

10.4.1 Limitations

First, it should be noted that as a result of the model architecture, a successful decision explanation requires an evenly successful training of the CycleGAN architecture behind it. Although CycleGANs are less prone to some of the issues faced with GANs (cf. Sec. 10.2.1), their training still requires a careful parameterization, and due to their iterative behavior, their training requires a considerable amount of time. Although in this

study the CycleGAN architecture was not systematically varied to assess its influence, there might be scenarios in which other generator architectures could be required to introduce decision-relevant changes.

Secondly, the method specifically aims for scenarios in which a decision visualization is needed for classifiers trained on small- to medium-sized imaging datasets. As was shown on the CIFAR-10 dataset, consisting of 10,000 training samples on both of the conducted experiments, classifiers that are trained on very large datasets are typically easier to explain using SoA methods, which in turn, however, often cannot be applied to medical imaging datasets, as they yield dissatisfying quality.

Finally, an even larger user study would be desirable. While the results contained nearly 10,000 answers and many of the comparisons yielded significant results despite the low number of degrees of freedom, a user study with a significantly larger amount of participants, especially when conducted in a radiological setting, would clearly help to further improve the method. Future studies should especially cover unguided assessments in which the alignment of decision-relevant regions as defined by the classifier, as well as decision-relevant image positions as manually set by a radiologist is compared.

10.5 Conclusion

Clinical decision support using deep neural networks requires new ways of decision explanation. While large and publicly available datasets are accessible for typical image vision tasks, medical imaging data is typically sparse and underlies strict regulations. As a result, classifiers which are trained on medical imaging data are often not satisfactorily explainable by SoA decision explanation methods.

As was shown in this chapter, CycleGANs can successfully be used as an effective method for medical decision explanation, yielding results that outperform various SoA methods in terms of intutivity, semantic meaningfulness, and image quality, to the best of knowledge being the first method to employ such a combination. Notably, the method does not require any specific classifier architecture, and thus serves as a black-box explanation module, which can easily be combined with any other of the approaches presented within this work. With decision explanation being a highly-relevant step towards the application of deep learning-based clinical decision support in the daily routine, this architecture in the future may thus contribute to improved patient healthcare.

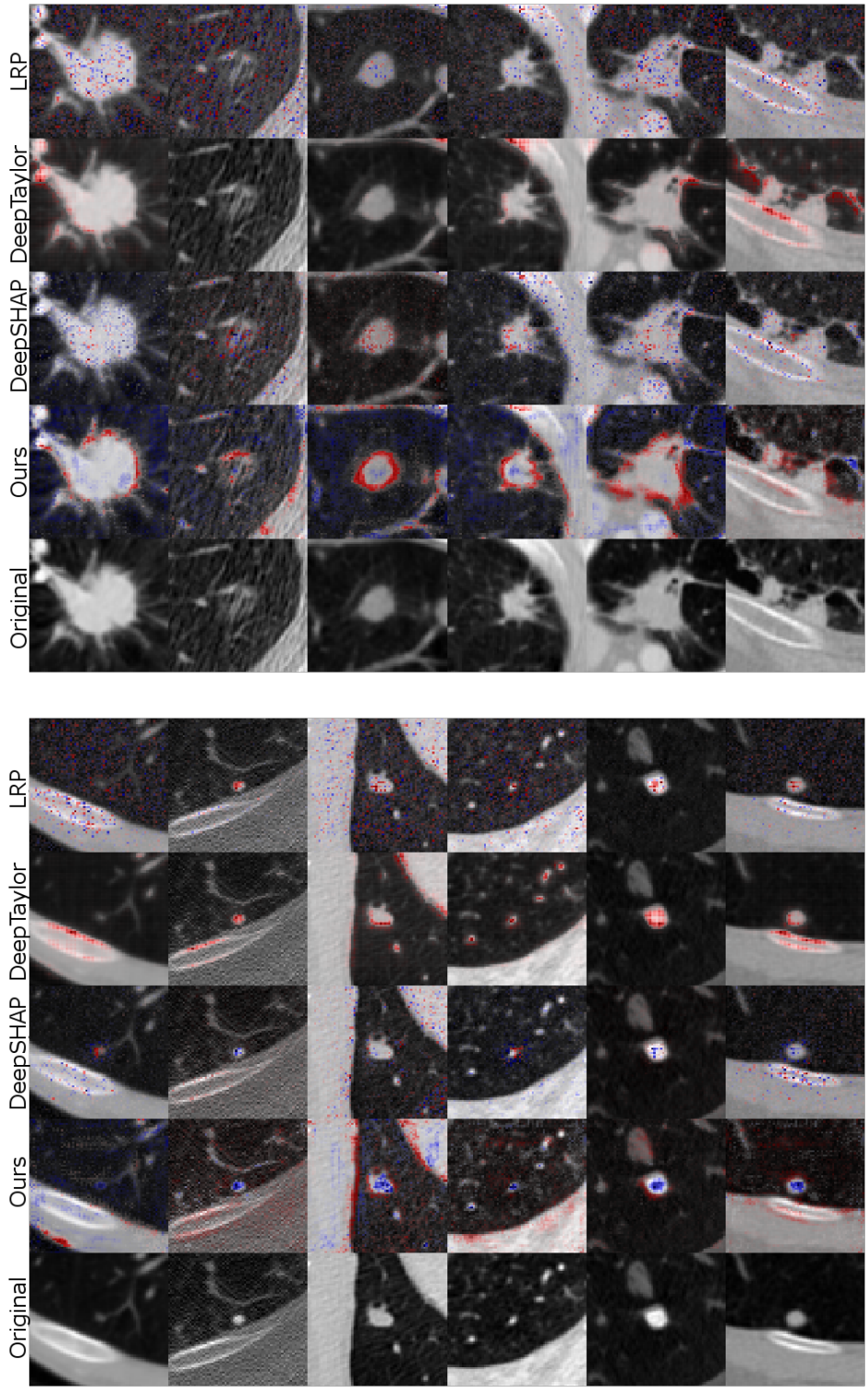


Figure 10.6: Exemplary decision explanations for the LIDC-IDRI classifier for lung lesion malignancy classification, comparing the proposed decision explanation method to the DeepSHAP, DeepTaylor and LRP approaches for benign (left) and malignant (right) lesions. Signs of malignancy (red) and benignity (blue), as quantified by each algorithm, are depicted as a colored overlay. Source: [Katzmann et al. 2021]

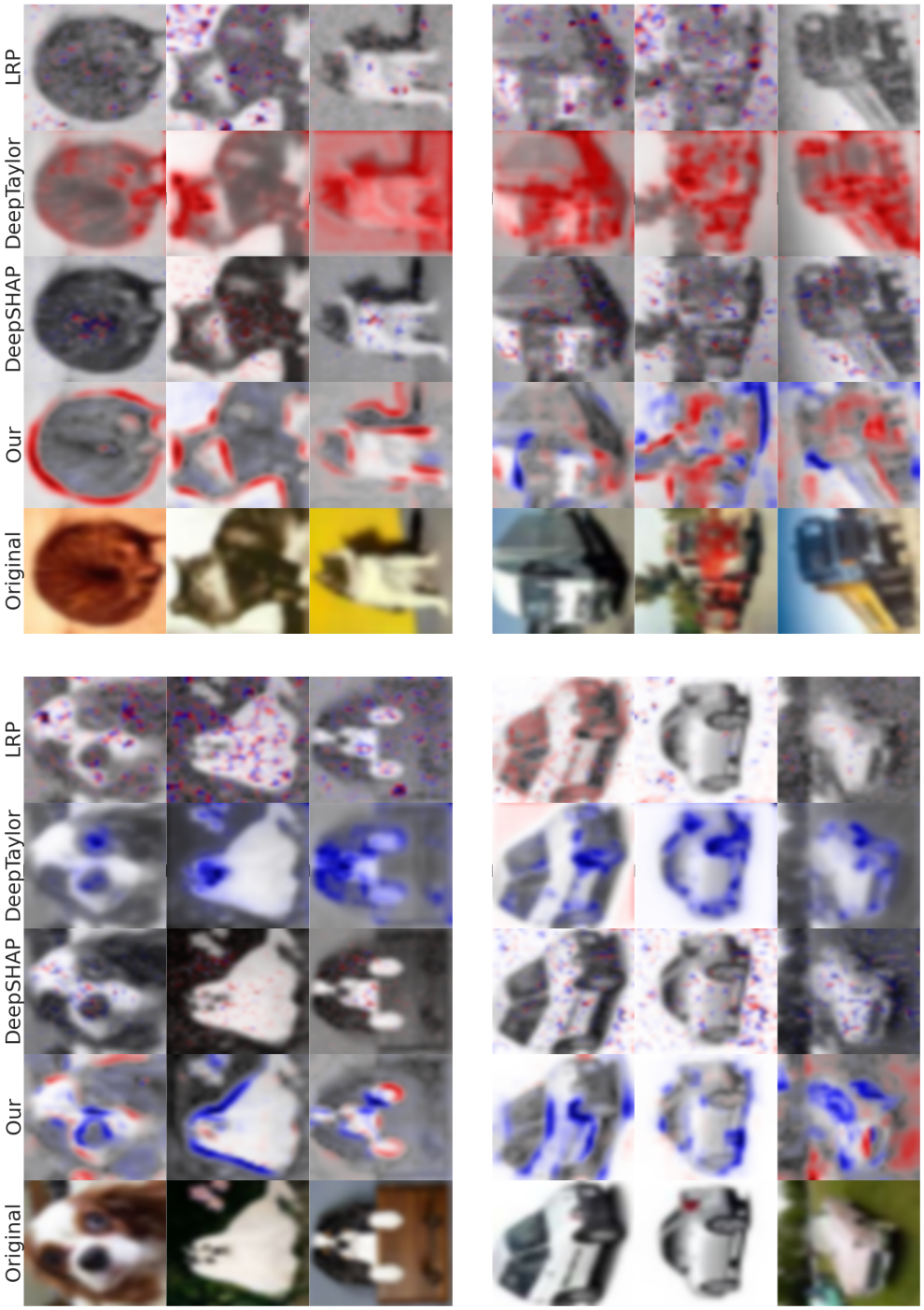


Figure 10.7: Exemplary decision explanations for the CIFAR-10 cats-vs-dogs (top) and cars-vs-trucks (bottom) classifiers, comparing the proposed method to the DeepSHAP, DeepTaylor and LRP approaches. Typical indicators for cats (red) and dogs (blue), or trucks (red) and cars (blue), respectively, are depicted as a colored overlay over the grayscale original image. Source: [Katzmann et al. 2021].

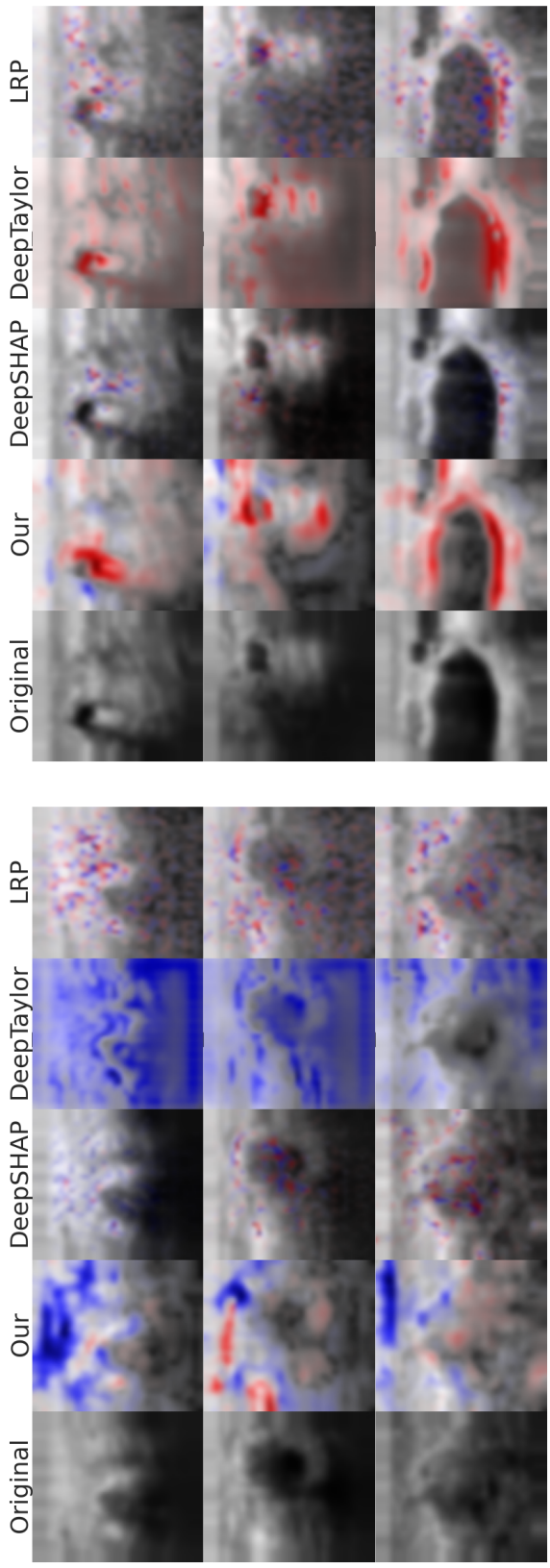
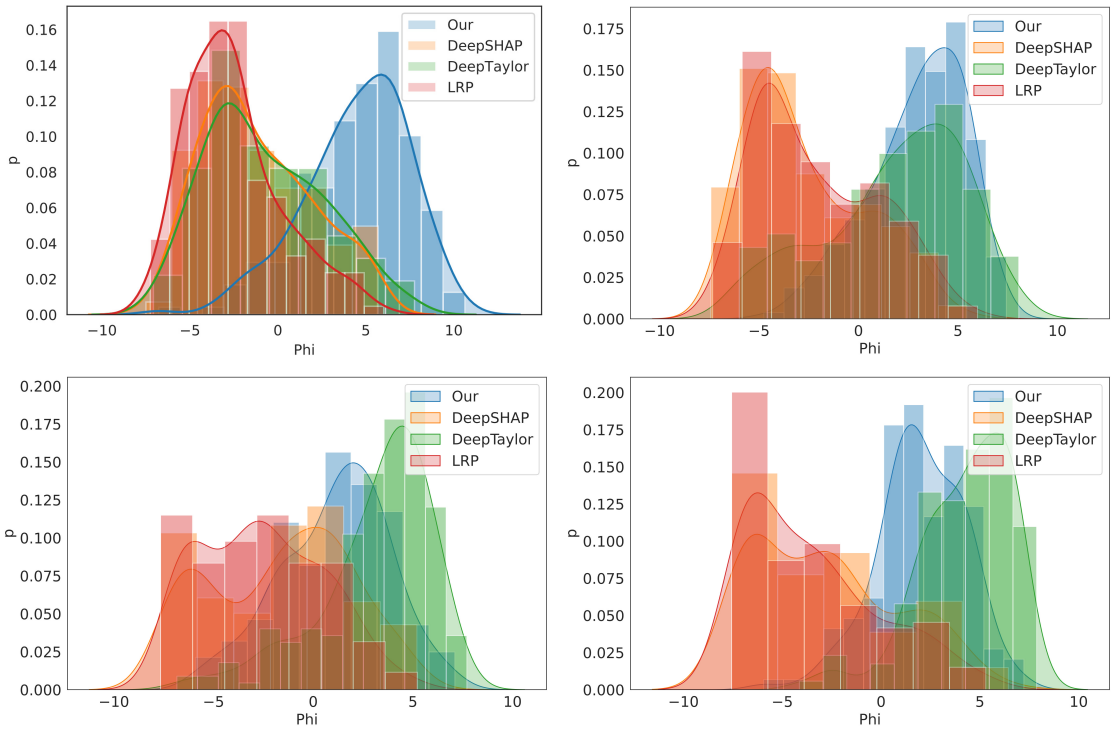


Figure 10.8: Exemplary decision explanations for the BreastMNIST classifier for breast lesion malignancy classification, comparing the proposed method to the DeepSHAP, DeepTaylor and LRP approaches for benign (left) and malignant (right) lesions. Signs of malignancy (red) and benignity (blue), as quantified by each algorithm, are depicted as a colored overlay. Source: [Katzmann et al. 2021]



	LIDC-IDRI	BreastMNIST	Cats-vs.-Dogs	Trucks-vs.-Cars
		Φ coefficients		
Intuitivity	.358	.341	.356	.340
Semantics	.367	.343	.377	.357
Quality	.274	.316	.267	.303

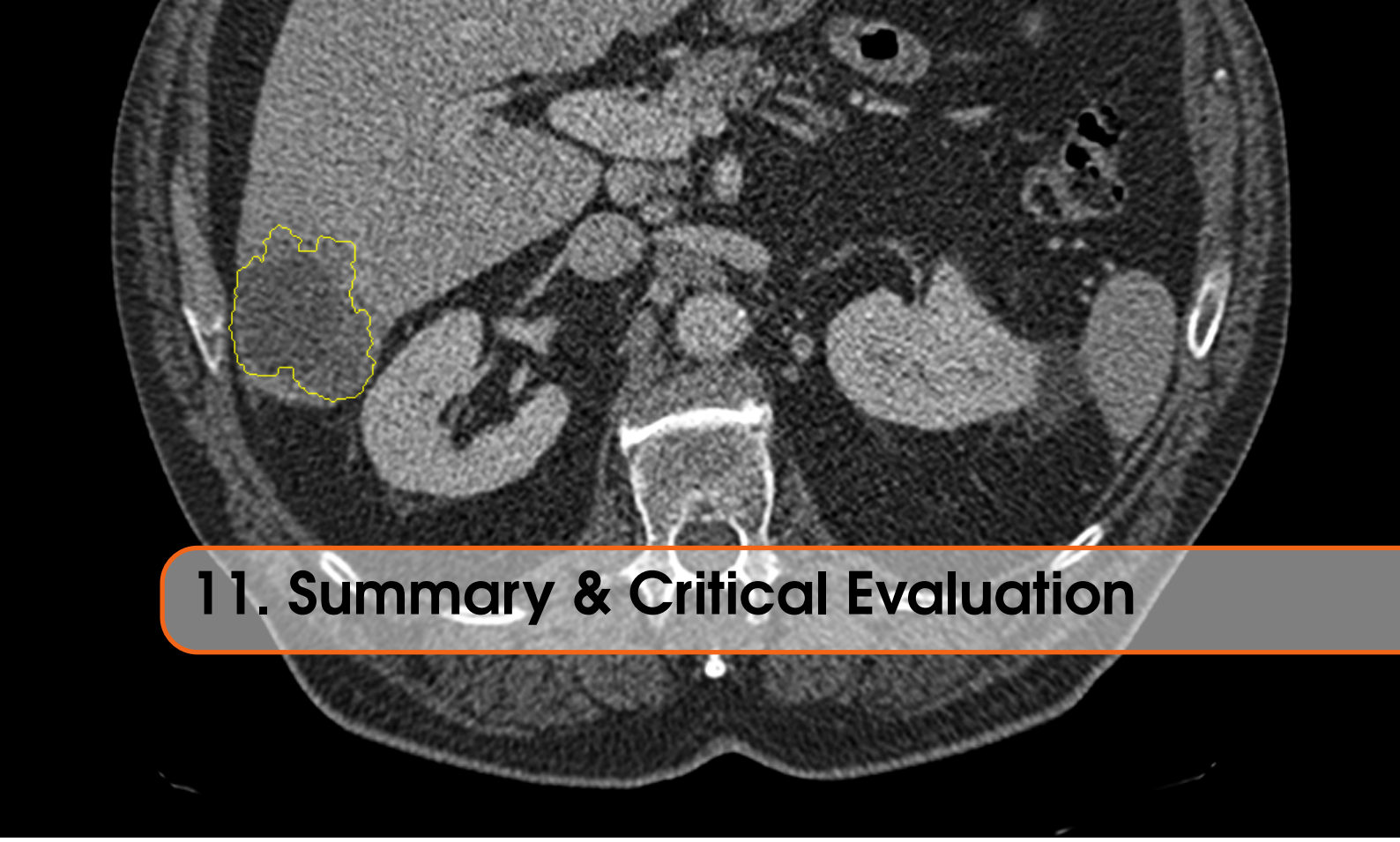
Figure 10.9: Histogram of preferability (bars) and kernel density estimation (line) by algorithm across images, as measured by extracting a general factor Φ using PCA on the LIDC-IDRI (top-left), BreastMNIST (top-right), Cats-vs.-Dogs (bottom-left) and Trucks-vs.-Cars (bottom-right) dataset, accounting for 84.6/86.5/76.6/88.0 % of the observed variance, respectively. The table, showing the factor load, demonstrates a similar factor structure for each dataset. Source: [Katzmann et al. 2021]

V

Conclusion

11 Summary & Critical Evaluation 115
11.1 Outcome Prediction in Oncology
11.2 Meta-Methods and Decision-Explanation

12 Outlook 119



11. Summary & Critical Evaluation

In the beginning, it was pointed out that deep learning is on its way to becoming an indispensable tool for clinical decision support and medical image analysis. It has demonstrated highly promising results in a variety of subdomains, sometimes even surpassing human gold-standard assessment (cf. Part I). Mainly arising from its data-driven nature, the use of deep learning has major obstacles which necessarily have to be overcome for achieving this goal. Next to its rather low comprehensibility, the data available for training deep approaches sustainably remains an issue, typically being sparse and often containing a relevant degree of data or label noise, which in turn results in practical implications for result quality and trainability. This work particularly addressed these issues, proposing methods for clinical decision support settled in the field of oncology. While being specifically tailored to this application, it was further demonstrated that these methods provide value beyond this scenario. Still, however, some open issues remain. Thus, the following chapter will briefly summarize the scientific contributions of this thesis, and relate them to the clinical requirements.

11.1 Outcome Prediction in Oncology

In Chapter 5, 6 and 7 methods for predicting lesion growth and patient survival in an oncological environment were discussed. The achieved results clearly indicate that the quantitative disease information gathered by using a combination of deep learning and radiological imaging data can be a highly promising source of additional information, and provides incremental value beyond size- or volume-based assessment, as it is currently used in clinical practice, as well as radiomics-based assessment. It has thus the potential to significantly enhance the clinical capabilities towards a quantitative, early assessment, reducing the risk of mistreatment, leading to higher efficiency, and improving overall healthcare quality.

Each of the presented methods, however, also has particular limitations. While these

have been discussed in more detail in the respective chapters already, this part will identify the key points to be addressed in future research.

First, while successfully addressing the issue of data sparsity, none of the approaches in Chapters 5–7 have addressed the issues of low transparency and comprehensibility for the treating practitioner, which arise from using deep-learning-based algorithms. While the results on lesion growth and one-year survival prediction clearly outperformed state-of-the-art (SoA) approaches, such as radiomics-based assessment, the latter has a clear benefit by being better comprehensible and easier to explain, being crucial factors in a clinical application. In fact, the algorithm's low comprehensibility can restrict the application of these approaches to scenarios with a fine-granular cross-check, or serving as a cross-check themselves, and to some degree limits possibilities for a transition into clinical practice. Their combination with methods for confidence estimation (cf. Chapter 8), or decision explanation (cf. Chapter 10) might therefore be highly beneficial in future applications.

Secondly, it has to be noted that patient survival prediction from radiological imaging data remains a rather difficult field, as patient survival is dependent on a variety of variables, such as age, gender, and comorbidities. Additionally, even lifestyle changes within therapy can lead to an increased or decreased survival expectancy (cf. [Parsons et al. 2010; Zutphen et al. 2017]). Notably, these factors can only partly be quantified using radiological imaging data, as was exemplarily shown by the insufficient results for continuous patient survival estimation in Chapter 7. Therefore, while such a prediction might be of highest clinical interest for the purpose of *risk stratification* and *therapy adaption*, a prediction of patient survival which is accurate *to the day* would require a wide range of additional modalities, will generally be unlikely within the near future, and even if possible would be highly debatable, as it goes hand-in-hand with a variety of ethical questions. This has impressively become clear by the widespread media attention in reaction to the work of Avati et al. [2017], which tried to predict patient mortality within the next 3-12 months for *proactively* reaching out to patients for palliative care [Mukherjee 2018]. In fact, it is reasonable to request that any approach for survival estimation has to be embedded into a workflow that ensures that estimates will always and only be used for the very best of the patient's health. This especially means that while such an algorithm might be used to compute the concrete risk of a given case, this information should never be used for deciding about access to medical care.

As this work has focused on the shape and textural lesion features, it did not take into account the multiplicity of lesions and their spatial distribution. In fact, the spatial distribution, i. e. *dispersion*, *does contain* valuable information for the disease outcome, as has exemplarily been shown in [Mühlberg et al. 2021a]. Future work should analyze a combination of both sources of information.

It has to be noted that the work proposed herein has primarily been evaluated in the field of oncological CT image classification, and so far is specific to it. However, the methods are expected to be easily transferable to different modalities, such as MRI or PET, as well as lesion types, e. g. renal or hepatocellular carcinomas. It would be highly desirable if this would be specifically addressed by future studies.

The algorithms presented within this work have served as a prototypical framework for outcome prediction as part of the BMBF[1] project "PANTHER". For a clinical application beyond this implementation, however, a significantly larger amount of data as well as a

[1] "Bildungsministerium für Bildung und Forschung", engl. Federal Ministry of Education and Research

thorough clinical evaluation is required and should be aimed for in the near future.

Finally, this work has focussed on radiological imaging data, thus aiming to assess the information which can be found within an image. In clinical practice, however, an interdisciplinary approach is crucial in order to achieve the best treatment response. Future work should thus additionally take into account the multimodality of clinical sources of information, as well as the interaction with other clinical practitioners.

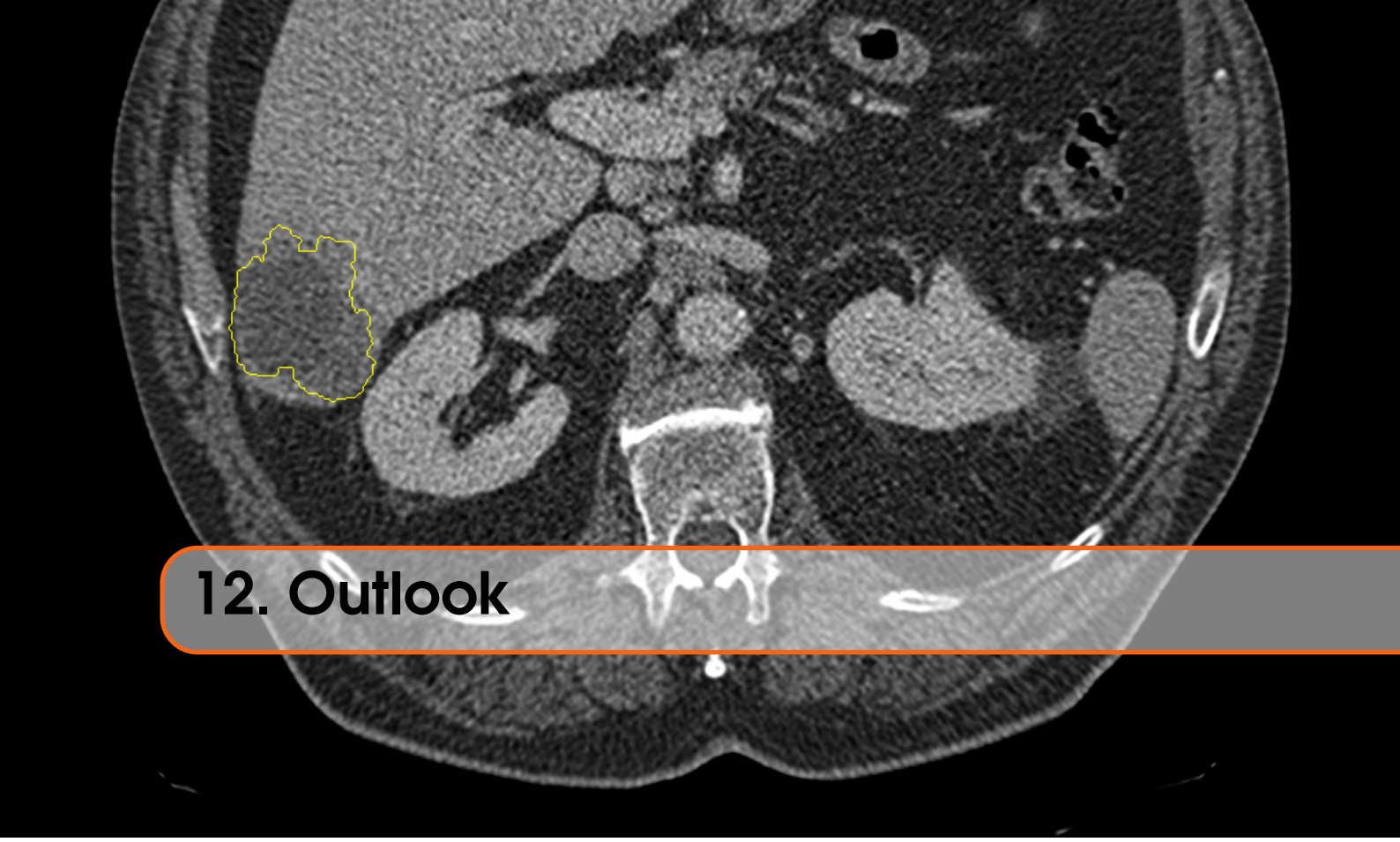
11.2 Meta-Methods and Decision-Explanation

While Part III has focussed on the successful application of deep learning for clinical decision support, in contrast, Part IV specifically addressed the issues which arise from this application, proposing methods particularly aiming for an improvement of classification performance as well as human comprehensibility, and being applicable when trained on medical imaging datasets, i. e. small datasets containing label noise and label imbalance (cf. Chapter 1.1).

First, providing a measure of classifier confidence is clinically highly important. It allows practitioners to interpret results according to their degree of confidence in a more targeted manner, and thus is a key criterion for clinical decision making and resulting treatment quality, as it allows an assessment of the algorithmic reliability. For this, a method has been proposed, based on a SoA algorithm, which not only provides an estimate of classifier confidence but can further be used for an improved training efficiency, yielding higher test-time performances. While both are beneficial for the final application, it should be noted that the approach alone cannot satisfactorily address all comprehensibility issues, as it only allows for a very limited view into the *reasons* for potential inconfidence.

To contend with this more specifically, further, a method for decision explanation has been proposed, aiming for a visualization of the network's understanding of an image, and outperforming recent SoA approaches when applied to two medical imaging datasets. In contrast to recent approaches, the network also provides reasonable explanations when only a few data are available, as this is often the case with medical imaging data. Although providing the best results in this scenario, it should be noted that even with this approach, however, the criteria in the user study had average raw values between 1.04 and 1.92 on a scale of -4 to 4, meaning that there is still considerable room for improvement. Future work should therefore even more focus on better comprehensibility for deep learning-based clinical decision support, as it is a key criterion for the transition into a clinical workflow. It is hoped that the approach presented herein can serve as a basis for this work.

Finally, data efficiency remains a topic of highest interest within the community. In Chapter 9 multiple methods for data efficiency enhancement have been proposed, each of them fundamentally based on the bootstrapping methodology from Efron [1979]. Although it was possible to clearly demonstrate the benefits of these architectures in few data scenarios, it should be noted that with increasing dataset sizes the incremental value of these methods diminishes. They are thus a temporary toolbox for the successful application of deep learning-based techniques for medical imaging data. Ultimately, however, the solution to superior clinical assessment lies in large and publicly available databases, providing the potential to reliably identify even rare medical conditions, and allowing the whole community to work together towards better healthcare for everyone.



12. Outlook

It was possible within this work to demonstrate a variety of applications for deep neural networks in clinical decision support. Including algorithms for lesion growth and patient survival prediction, the work presented within this thesis is among the first to use deep learning for computer-aided disease assessment in metastatic colorectal cancer patients. With the contributions to data efficiency and classifier comprehensibility enhancement, it addressed key factors for a translation of deep learning into clinical practice.

Still, clinical outcome prediction is at its very beginning, and likely future successes will clearly outperform current solutions. For now, clinical decision support will mainly serve as a cross-check, helping to prioritize cases, reduce the risk of mistreatment, and to accelerate everyday clinical tasks, leading to better availability of high-quality treatment for many patients. Fully-automated computer-aided diagnosis, however, will for now stay a topic of the future, and likely will be until its performance nearly exceptionless outperforms gold-standard human assessment. Next to the very practical reasons, such as legal liability, there are notable ethical and sociological reasons for this, as was discussed earlier in this work. Using a trade-off, such as the clinician-in-the-loop strategy [Tang et al. 2020], might help to overcome obstacles on this course. However, inevitably future research will, even more, have to focus on specific fields, some of which are briefly outlined in the following:

1. **Comprehensibility** - Future methods should even better explain classifier decisions. This especially may include a verbal description of decision processes in an interactive question-answer scheme, but also utilizing other means, such as the input-space visualizations presented in this work.

2. **Out-of-distribution (OOD) detection** - As the work from Goodfellow et al. [2014] impressively demonstrated, OOD samples, whether mistakenly or viciously introduced, can have a devastating influence on deep learning-based algorithms, with highly relevant implications for the medical imaging domain [Finlayson et al. 2019],

although recent research has identified ways to partly combat these issue (cf. [DeVries et al. 2018; Madry et al. 2018]). However, OOD errors do not align with human intuition. They are thus unexpected and may lead to serious consequences. Reliability in these scenarios, therefore, is of *highest importance* for clinical acceptance, creating a need for future research to address this issue more thoroughly.

3. **Data availability** - As pointed out above, finally, data availability is key for high-quality clinical decision support algorithms. As for now the lack of large and publicly available datasets should be addressed, to this end future research will have to find ways to allow for continuous learning through automated data-mining in a fully-integrated IT landscape, while preserving data privacy and security when training on highly-sensitive patient data. A key technology towards this aim will be *Federated Learning* [Rieke et al. 2020].

Within the near future, highly dynamic progress in the field of medical image analysis is to be expected. With the work presented in this thesis, some of the major obstacles on the way there have been addressed, and may hopefully be used by future work, aiming toward even better treatment and globally improved patient healthcare.

VI

Appendix

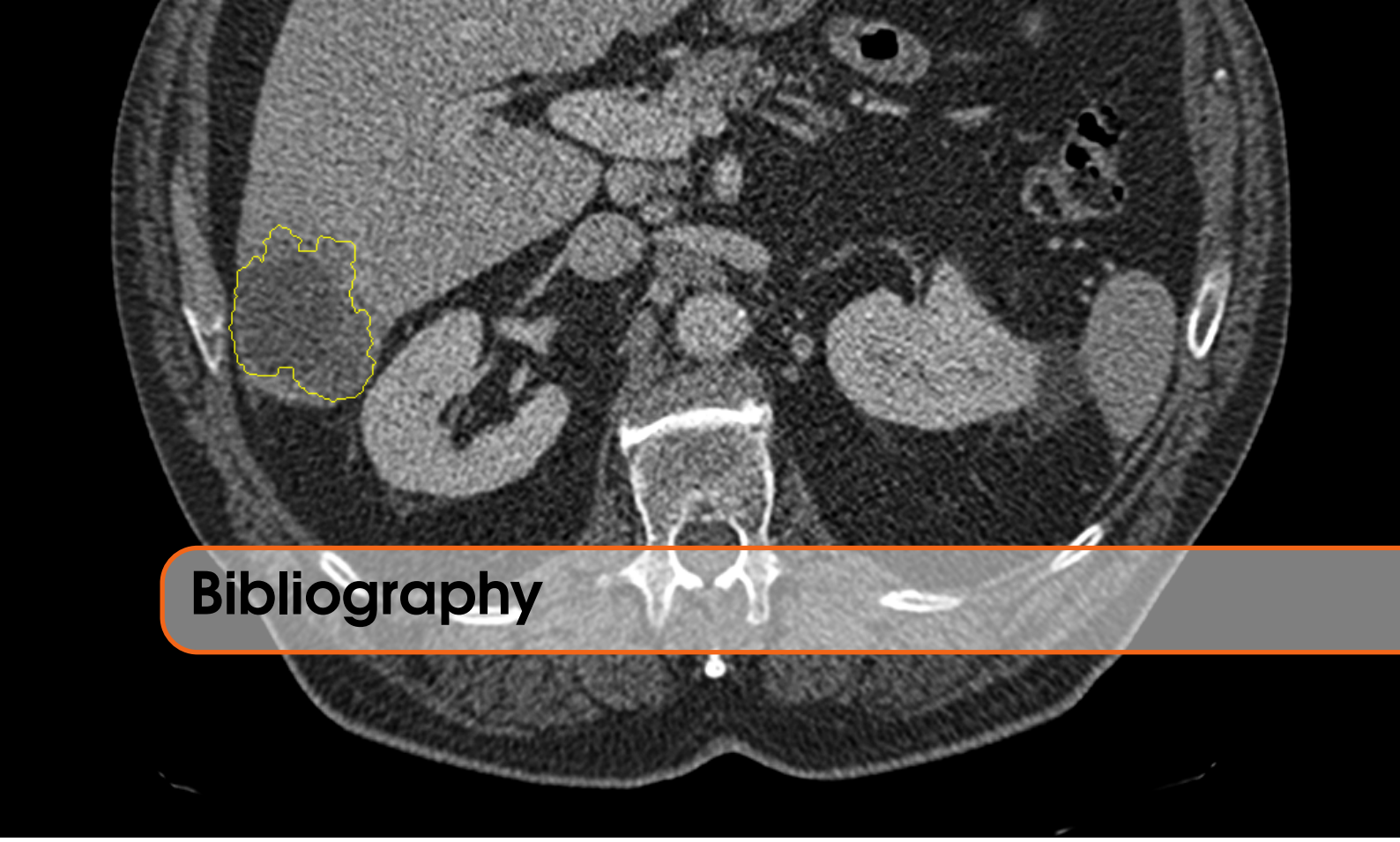
Bibliography 123

A Additional Background 141

B Tables & Figures 143

C Metrics & Measures 159

Index 165



Bibliography

Aalen, Odd O (1989). "A linear regression model for the analysis of life times". In: *Statistics in Medicine* 8.8, pp. 907–925 (cit. on p. 48).

AAMC (2021). *Association of American Medical Colleges - Physician Specialty Data Report*. URL: https://www.aamc.org/data-reports/workforce/report/physician-specialty-data-report (visited on 09/29/2021) (cit. on p. 17).

AARP (2021). *AARP Magazine - Cancer Treatment Costs*. URL: https://web.archive.org/https://www.aarp.org/money/credit-loans-debt/info-2018/the-high-cost-of-cancer-treatment.html (visited on 08/25/2021) (cit. on p. 13).

Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim (2018). "Sanity checks for saliency maps". In: *Advances in Neural Information Processing Systems*, pp. 9505–9515 (cit. on p. 97).

Aerts, Hugo J. W. L. et al. (2015). *Data From NSCLC-Radiomics*. DOI: 10.7937/k9/tcia.2015.pf0m9rei. URL: https://wiki.cancerimagingarchive.net/x/FgL1 (cit. on p. 88).

Aerts, Hugo JWL et al. (2014). "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach". In: *Nature Communications* 5 (cit. on pp. 19, 20, 29, 31, 34, 40, 41, 45, 53, 88).

Ahmed, Ejaz, Michael Jones, and Tim K Marks (2015). "An improved deep learning architecture for person re-identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3908–3916 (cit. on p. 3).

Alber, Maximilian et al. (2019). "iNNvestigate Neural Networks!" In: *Journal of Machine Learning Research* 20.93, pp. 1–8. URL: http://jmlr.org/papers/v20/18-540.html (cit. on p. 103).

Alshabibi, AS, ME Suleiman, KA Tapia, and PC Brennan (2020). "Effects of time of day on radiological interpretation". In: *Clinical Radiology* 75.2, pp. 148–155 (cit. on p. 27).

American Cancer Society (2021a). *Global Cancer Facts & Figures*. URL: https://web.archive.org/web/https://www.cancer.org/research/cancer-facts-statistics/global.html (visited on 08/25/2021) (cit. on p. 13).

— (2017a). *Key Statistics for Colorectal Cancer*. URL: https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html (visited on 10/03/2017) (cit. on p. 15).

— (2021b). *Lung Cancer Key Statistics*. URL: https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html (visited on 08/25/2021) (cit. on p. 16).

— (2021c). *Lung Cancer Survival Rates*. URL: https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/survival-rates.html (visited on 08/25/2021) (cit. on p. 16).

— (2021d). *NSCLC Statistics*. URL: https://www.cancer.net/cancer-types/lung-cancer-non-small-cell/statistics (visited on 08/25/2021) (cit. on p. 16).

— (2021e). *SCLC Statistics*. URL: https://www.cancer.net/cancer-types/lung-cancer-small-cell/statistics (visited on 08/25/2021) (cit. on p. 16).

— (2017b). *Survival Rates for Colorectal Cancer*. URL: https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html (visited on 11/16/2017) (cit. on p. 48).

Anthony, Thomas et al. (2020). "Learning to play no-press diplomacy with best response policy iteration". In: *arXiv preprint arXiv:2006.04635* (cit. on p. 3).

Apgar, V (1952). "A proposal for a new method of evaluation of the newborn". In: *Classic Papers in Critical Care* 32.449, p. 97 (cit. on p. 20).

Armato, Samuel G et al. (2011). "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans". In: *Medical Physics* 38.2, pp. 915–931 (cit. on pp. 82, 102).

Armato III, Samuel G, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Laurence P Clarke, et al. (2015). "Data from LIDC-IDRI. The Cancer Imaging Archive". In: *DOI http://doi. org/10.7937 K* 9, p. 7 (cit. on pp. 82, 102).

Assran, Mahmoud, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat (2021). "Semi-Supervised Learning of Visual Features by Non-Parametrically Predicting View Assignments with Support Samples". In: *arXiv preprint arXiv:2104.13963* (cit. on pp. 38, 97).

Atanov, Andrei, Arsenii Ashukha, Dmitry Molchanov, Kirill Neklyudov, and Dmitry Vetrov (2018). "Uncertainty Estimation via Stochastic Batch Normalization". In: *arXiv preprint arXiv:1802.04893* (cit. on p. 62).

Avati, Anand, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H Shah (2017). "Improving palliative care with deep learning". In: *arXiv preprint arXiv:1711.06402* (cit. on p. 116).

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek (2015). "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS One* 10.7, e0130140 (cit. on pp. 97, 103).

Begoli, Edmon, Tanmoy Bhattacharya, and Dimitri Kusnezov (2019). "The need for uncertainty quantification in machine-assisted medical decision making". In: *Nature Machine Intelligence* 1.1, pp. 20–23 (cit. on pp. 61, 62).

Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston (2009). "Curriculum learning". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, pp. 41–48 (cit. on pp. 62, 66, 71).

Benjamens, Stan, Pranavsingh Dhunnoo, and Bertalan Meskó (2020). "The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database". In: *NPJ Digital Medicine* 3.1, pp. 1–8 (cit. on p. 15).

Bogowicz, Marta, Oliver Riesterer, Luisa Sabrina Stark, Gabriela Studer, Jan Unkelbach, Matthias Guckenberger, and Stephanie Tanadini-Lang (2017). "Comparison of PET and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma". In: *Acta Oncologica* 56.11, pp. 1531–1536 (cit. on p. 31).

Bravais, Auguste (1844). *Analyse mathématique sur les probabilités des erreurs de situation d'un point*. Impr. Royale (cit. on p. 161).

Breiman, L, JH Friedman, R Olshen, and CJ Stone (1984). "Classification and Regression Trees". In: (cit. on p. 70).

Breiman, Leo (2001). "Random forests". In: *Machine Learning* 45.1, pp. 5–32 (cit. on pp. 74, 78).

Brierley, James D, Mary K Gospodarowicz, and Christian Wittekind (2016). *TNM classification of malignant tumours*. John Wiley & Sons (cit. on p. 52).

Brooks, Rodney A and Giovanni Di Chiro (1975). "Theory of image reconstruction in computed tomography". In: *Radiology* 117.3, pp. 561–572 (cit. on pp. 14, 142).

Brosch, Tom, Youngjin Yoo, Lisa YW Tang, David KB Li, Anthony Traboulsee, and Roger Tam (2015). "Deep convolutional encoder networks for multiple sclerosis lesion segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 3–11 (cit. on p. 23).

Brown, Tom B et al. (2020). "Language models are few-shot learners". In: *arXiv preprint arXiv:2005.14165* (cit. on p. 24).

Cambridge Online Dictionary (2021). *Medicine*. URL: `https://web.archive.org/web/https://dictionary.cambridge.org/de/worterbuch/englisch/medicine` (visited on 08/25/2021) (cit. on p. 13).

Caron, Mathilde, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin (2021). "Emerging properties in self-supervised vision transformers". In: *arXiv preprint arXiv:2104.14294* (cit. on p. 97).

Chattopadhay, Aditya, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian (2018). "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 839–847 (cit. on p. 97).

Chiou, Victoria L and Mauricio Burotto (2015). "Pseudoprogression and immune-related response in solid tumors". In: *Journal of Clinical Oncology* 33.31, p. 3541 (cit. on p. 40).

Chollet, François et al. (2015). *Keras*. `https://github.com/fchollet/keras` (cit. on pp. 35, 103).

Choo, Jaegul and Shixia Liu (2018). "Visual analytics for explainable deep learning". In: *IEEE computer graphics and applications* 38.4, pp. 84–92 (cit. on p. 97).

Clark, Kenneth et al. (2013). "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository". In: *Journal of Digital Imaging* 26.6, pp. 1045–1057 (cit. on pp. 82, 88, 102).

Clinical Oncology, American Society of (2021). *Colorectal Cancer: Statistics*. URL: https://www.cancer.net/cancer-types/colorectal-cancer/statistics (visited on 09/23/2021) (cit. on p. 16).

Cohen, Steven J et al. (2008). "Relationship of circulating tumor cells to tumor response, progression-free survival, and overall survival in patients with metastatic colorectal cancer". In: *Clin Oncol* 26, pp. 3213–3221 (cit. on p. 48).

Cowan, Ian A, Sharyn LS MacDonald, and Richard A Floyd (2013). "Measuring and managing radiologist workload: Measuring radiologist reporting times using data from a R adiology I nformation S ystem". In: *Journal of Medical Imaging and Radiation Oncology* 57.5, pp. 558–566 (cit. on p. 4).

Cowley, Helen C and Alastair G Gale (1997). "Time-of-day effects on mammographic film reading performance". In: *Medical Imaging 1997: Image Perception*. Vol. 3036. International Society for Optics and Photonics, pp. 212–221 (cit. on p. 27).

Cox, David R (1972). "Regression models and life-tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2, pp. 187–202 (cit. on pp. 47, 49, 54, 88).

Damiens, K, JPM Ayoub, B Lemieux, F Aubin, W Saliba, MP Campeau, and Mustapha Tehfe (2012). "Clinical features and course of brain metastases in colorectal cancer: an experience from a single institution". In: *Current Oncology* 19.5, pp. 254–258 (cit. on p. 17).

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso (2011). "Extraneous factors in judicial decisions". In: *Proceedings of the National Academy of Sciences* 108.17, pp. 6889–6892 (cit. on p. 27).

Davidson-Pilon, Cameron (2019). "lifelines: survival analysis in Python". In: *Journal of Open Source Software* 4.40, p. 1317 (cit. on pp. 52, 89).

De Vries, AMM, Y De Roten, C Meystre, J Passchier, J-N Despland, and F Stiefel (2014). "Clinician characteristics, communication, and patient outcome in oncology: a systematic review". In: *Psycho-Oncology* 23.4, pp. 375–381 (cit. on p. 18).

Desquilbet, L and L Meyer (2005). "Time-dependent covariates in the Cox proportional hazards model. Theory and practice". In: *Revue d'epidemiologie et de sante publique* 53.1, pp. 51–68 (cit. on p. 48).

DeVries, Terrance and Graham W Taylor (2017). "Dataset augmentation in feature space". In: *arXiv preprint arXiv:1702.05538* (cit. on p. 76).

— (2018). "Learning Confidence for Out-of-Distribution Detection in Neural Networks". In: *arXiv preprint arXiv:1802.04865* (cit. on pp. 62, 64–68, 71, 72, 120).

Devroye, Luc (1986). "Sample-based non-uniform random variate generation". In: *Proceedings of the 18th conference on Winter simulation*, pp. 260–265 (cit. on p. 67).

Dozat, Timothy (2016). "Incorporating nesterov momentum into adam". In: (cit. on pp. 35, 43).

Edoarado (Jan. 20, 2021). *Human Anatomy Planes and Latin derived orientations. License: Creative Commons Attribution-Share Alike 3.0 Unported: https://creativecommons.org/licenses/by-sa/3.0/deed.en*. URL: https://commons.wikimedia.org/wiki/File:Human_Anatomy_Planes_and_Latin_orientations.png (visited on 09/29/2021) (cit. on p. 141).

Efron, B (1979). "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics*, pp. 1–26 (cit. on pp. 73, 117).

Efron, Bradley (1982). *The jackknife, the bootstrap, and other resampling plans*. Vol. 38. Siam (cit. on pp. 68, 82).

— (1987). "Better bootstrap confidence intervals". In: *Journal of the American Statistical Association* 82.397, pp. 171–185 (cit. on pp. 43, 52, 74, 103).

Efron, Bradley and Robert Tibshirani (1997). "Improvements on cross-validation: the 632+ bootstrap method". In: *Journal of the American Statistical Association* 92.438, pp. 548–560 (cit. on p. 85).

Eisenbach, Markus, Daniel Seichter, Tim Wengefeld, and Horst-Michael Gross (2016). "Cooperative multi-scale convolutional neural networks for person detection". In: *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 267–276 (cit. on p. 3).

Eisenhauer, EA1 et al. (2009). "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)". In: *European Journal of Cancer* 45.2, pp. 228–247 (cit. on pp. 29, 34).

Elsken, Thomas, Jan Hendrik Metzen, and Frank Hutter (2019). "Neural architecture search: A survey". In: *The Journal of Machine Learning Research* 20.1, pp. 1997–2017 (cit. on p. 79).

Erhan, Dumitru, Yoshua Bengio, Aaron Courville, and Pascal Vincent (2009). "Visualizing higher-layer features of a deep network". In: *University of Montreal* 1341.3, p. 1 (cit. on p. 98).

FDA (2018). *Computed Tomography (CT)*. Youtube. URL: `https://www.fda.gov/radiation-emitting-products/medical-x-ray-imaging/computed-tomography-ct` (visited on 09/29/2021) (cit. on pp. 14, 142).

Feldman, Myra K, Sanjeev Katyal, and Margaret S Blackwood (2009). "US artifacts". In: *Radiographics* 29.4, pp. 1179–1189 (cit. on p. 15).

Finlayson, Samuel G, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane (2019). "Adversarial attacks on medical machine learning". In: *Science* 363.6433, pp. 1287–1289 (cit. on p. 119).

Flanders, W Dana, Cathy A Lally, Bao-Ping Zhu, S Jane Henley, and Michael J Thun (2003). "Lung cancer mortality in relation to age, duration of smoking, and daily cigarette consumption: results from Cancer Prevention Study II". In: *Cancer research* 63.19, pp. 6556–6562 (cit. on p. 16).

Flavell, John H (1971). "First discussant's comments: What is memory development the development of?" In: *Human development* 14.4, pp. 272–278 (cit. on p. 64).

Food, FDA U.S., Drug Administration - Center for Devices, and Radiological Health (2021). *Artificial Intelligence and Machine Learning in Software*. URL: `https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device` (visited on 09/23/2021) (cit. on p. 15).

Fox, John and Sanford Weisberg (2002). "Cox proportional-hazards regression for survival data". In: *An R and S-PLUS companion to applied regression* 2002 (cit. on p. 53).

Gal, Yarin and Zoubin Ghahramani (2016). "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning". In: *International Conference on Machine Learning*, pp. 1050–1059 (cit. on pp. 62, 63).

Ganeshan, Balaji and Kenneth A Miles (2013). "Quantifying tumour heterogeneity with CT". In: *Cancer imaging* 13.1, p. 140 (cit. on p. 31).

Gastaldi, Xavier (2017). “Shake-shake regularization”. In: *arXiv preprint arXiv:1705.07485* (cit. on pp. 75, 77).

Ghesu, Florin C et al. (2019). “Quantifying and leveraging classification uncertainty for chest radiograph assessment”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 676–684 (cit. on p. 62).

Ghorbani, Amirata, Abubakar Abid, and James Zou (2019). “Interpretation of neural networks is fragile”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 3681–3688 (cit. on p. 97).

Gillies, Robert J, Paul E Kinahan, and Hedvig Hricak (2015). “Radiomics: images are more than pictures, they are data”. In: *Radiology* 278.2, pp. 563–577 (cit. on p. 31).

Gini, Corrado (1912). “Variabilità e mutabilità (Variability and Mutability)”. In: *Cuppini, Bologna* 156 (cit. on p. 70).

Glimelius, B, E Tiret, A Cervantes, D Arnold, et al. (2013). “Rectal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up”. In: *Ann Oncol* 24.Suppl 6) (cit. on p. 27).

González, Roberto E, Roberto P Munoz, and Cristian A Hernández (2018). “Galaxy detection and identification using deep learning and data augmentation”. In: *Astronomy and Computing* 25, pp. 103–109 (cit. on p. 3).

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2020). “Generative adversarial networks”. In: *Communications of the ACM* 63.11, pp. 139–144 (cit. on p. 98).

Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy (2014). “Explaining and harnessing adversarial examples”. In: *arXiv preprint arXiv:1412.6572* (cit. on pp. 35, 61, 119).

Graves, Alex, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu (2017). “Automated curriculum learning for neural networks”. In: *International Conference on Machine Learning*. PMLR, pp. 1311–1320 (cit. on p. 72).

Griethuysen, Joost JM van et al. (2017). “Computational Radiomics System to Decode the Radiographic Phenotype”. In: *Cancer Research* 77.21, e104–e107 (cit. on p. 41).

Grill, Jean-Bastien et al. (2020). “Bootstrap Your Own Latent: A new approach to self-supervised learning”. In: *Neural Information Processing Systems* (cit. on p. 38).

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti (2018). “Local rule-based explanations of black box decision systems”. In: *arXiv preprint arXiv:1805.10820* (cit. on p. 97).

Gurumurthy, Swaminathan, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu (2017). “Deligan: Generative adversarial networks for diverse and limited data”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 166–174 (cit. on p. 98).

Haarburger, Christoph, Philippe Weitz, Oliver Rippel, and Dorit Merhof (2018). “Image-based Survival Analysis for Lung Cancer Patients using CNNs”. In: *arXiv preprint arXiv:1808.09679* (cit. on pp. 49, 55, 62).

Hand, David and Peter Christen (2018). “A note on using the F-measure for evaluating record linkage algorithms”. In: *Statistics and Computing* 28.3, pp. 539–547 (cit. on p. 159).

Harrell Jr, Frank E, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati (1984). "Regression modelling strategies for improved prognostic prediction". In: *Statistics in Medicine* 3.2, pp. 143–152 (cit. on p. 23).

Hart, Julian T (1965). "Memory and the feeling-of-knowing experience." In: *Journal of Educational Psychology* 56.4, p. 208 (cit. on p. 64).

Hausen, Harald zur (2012). "Red meat consumption and cancer: reasons to suspect involvement of bovine infectious factors in colorectal cancer". In: *International Journal of Cancer* 130.11, pp. 2475–2483 (cit. on p. 16).

Havaei, Mohammad et al. (2017). "Brain tumor segmentation with deep neural networks". In: *Medical Image Analysis* 35, pp. 18–31 (cit. on p. 23).

Hayes, SA, MC Pietanza, D O'Driscoll, J Zheng, CS Moskowitz, MG Kris, and MS Ginsberg (2016). "Comparison of CT volumetric measurement with RECIST response in patients with lung cancer". In: *European Journal of Radiology* 85.3, pp. 524–533 (cit. on p. 34).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (cit. on pp. 50, 52, 53, 55, 72, 75, 81, 101).

Healthineers, Siemens (2021). *SOMATOM Force*. URL: `https://www.siemens-healthineers.com/de-ch/computed-tomography/dual-source-ct/somatom-force` (visited on 09/29/2021) (cit. on p. 142).

Hendrycks, Dan and Kevin Gimpel (2016). "A baseline for detecting misclassified and out-of-distribution examples in neural networks". In: *arXiv preprint arXiv:1610.02136* (cit. on p. 61).

Hinton, Geoffrey E and Ruslan R Salakhutdinov (2006). "Reducing the dimensionality of data with neural networks". In: *Science* 313.5786, pp. 504–507 (cit. on pp. 31, 41).

Hinton, Geoffrey E, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov (2012). "Improving neural networks by preventing co-adaptation of feature detectors". In: *arXiv preprint arXiv:1207.0580* (cit. on p. 32).

Holch, Julian Walter, Maximilian Demmer, Charlotte Lamersdorf, Marlies Michl, Christoph Schulz, Jobst Christian von Einem, Dominik Paul Modest, and Volker Heinemann (2017). "Pattern and dynamics of distant metastases in metastatic colorectal cancer". In: *Visceral Medicine* 33.1, pp. 70–75 (cit. on p. 15).

Houthooft, Rein, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel (2016). "Vime: Variational information maximizing exploration". In: *arXiv preprint arXiv:1605.09674* (cit. on p. 72).

Huang, Yan-qi, Chang-hong Liang, Lan He, Jie Tian, Cui-shan Liang, Xin Chen, Ze-lan Ma, and Zai-yi Liu (2016). "Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer". In: *Journal of Clinical Oncology* 34.18, pp. 2157–2164 (cit. on p. 45).

Hueper, Katja et al. (2015). "Pulmonary microvascular blood flow in mild chronic obstructive pulmonary disease and emphysema. The MESA COPD Study". In: *American journal of respiratory and critical care medicine* 192.5, pp. 570–580 (cit. on p. 20).

Hüllermeier, Eyke and Willem Waegeman (2021). "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods". In: *Machine Learning* 110.3, pp. 457–506 (cit. on p. 63).

Ioffe, Sergey and Christian Szegedy (2015). “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International Conference on Machine Learning*, pp. 448–456 (cit. on pp. 32, 42, 63, 81).

Ishwaran, Hemant, Udaya B Kogalur, Eugene H Blackstone, Michael S Lauer, et al. (2008). “Random survival forests”. In: *The Annals of Applied Statistics* 2.3, pp. 841–860 (cit. on pp. 48, 87–89).

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros (2017). “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134 (cit. on pp. 101, 157).

Joachim, Clarisse, Jonathan Macni, Moustapha Drame, Audrey Pomier, Patrick Escarmant, Jacqueline Veronique-Baudin, and Vincent Vinh-Hung (2019). “Overall survival of colorectal cancer by stage at diagnosis: Data from the Martinique Cancer Registry”. In: *Medicine* 98.35 (cit. on p. 42).

Johnson-Laird, PN, Paolo Legrenzi, Vittorio Girotto, Maria Sonino Legrenzi, and Jean-Paul Caverni (1999). “Naive probability: a mental model theory of extensional reasoning.” In: *Psychological Review* 106.1, p. 62 (cit. on p. 96).

Jumper, John et al. (2021). “Highly accurate protein structure prediction with AlphaFold”. In: *Nature*, p. 1 (cit. on p. 3).

Katharopoulos, Angelos and François Fleuret (2017). “Biased Importance Sampling for Deep Neural Network Training”. In: *arXiv preprint arXiv:1706.00043* (cit. on p. 35).

Katzman, Jared L, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger (2016). “Deep survival: A deep cox proportional hazards network”. In: *stat* 1050, p. 2 (cit. on pp. 49, 55).

Katzmann, Alexander, Alexander Muehlberg, Michael Suehling, Dominik Noerenberg, Julian Walter Holch, Volker Heinemann, and Horst-Michael Gross (2018a). “Predicting lesion growth and patient survival in colorectal cancer patients using deep neural networks”. In: *1st International Conference on Medical Imaging with Deep Learning (MIDL 2018)* (cit. on pp. 33, 39–46, 53, 69, 70, 93, 94, 146, 147).

Katzmann, Alexander, Alexander Muehlberg, Michael Suehling, Dominik Nörenberg, Julian Walter Holch, and Horst-Michael Gross (2020). “Deep Random Forests for Small Sample Size Prediction with Medical Imaging Data”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 1543–1547 (cit. on pp. 78–80, 82).

Katzmann, Alexander, Alexander Mühlberg, Michael Sühling, Dominik Nörenberg, and Horst-Michael Groß (2019a). “Deep Metamemory-A Generic Framework for Stabilized One-Shot Confidence Estimation in Deep Neural Networks and its Application on Colorectal Cancer Liver Metastases Growth Prediction”. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, pp. 1298–1302 (cit. on pp. 62, 64, 67).

Katzmann, Alexander, Alexander Mühlberg, Michael Sühling, Dominik Nörenberg, Julian Walter Holch, and Horst-Michael Groß (2018b). “TumorEncode-Deep Convolutional Autoencoder for Computed Tomography Tumor Treatment Assessment”. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8 (cit. on pp. 23, 27, 29, 31, 32, 34–38, 41–43, 53, 144–146).

Katzmann, Alexander, Alexander Mühlberg, Michael Sühling, Dominik Nörenberg, Stefan Maurus, Julian Walter Holch, Volker Heinemann, and Horst-Michael Groß (2019b). "Computed Tomography Image-Based Deep Survival Regression for Metastatic Colorectal Cancer Using a Non-proportional Hazards Model". In: *International Workshop on PRedictive Intelligence In MEdicine (MICCAI-PRIME), International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI 2019)*. Springer, pp. 73–80 (cit. on pp. 48–52).

Katzmann, Alexander, Oliver Taubmann, Stephen Ahmad, Alexander Mühlberg, Michael Sühling, and Horst-Michael Groß (2021). "Explaining clinical decision support systems in medical imaging using cycle-consistent activation maximization". In: *Neurocomputing* 458, pp. 141–156 (cit. on pp. 93, 98, 100–106, 109–112, 148–152, 156, 157).

Kelemen, William L (2000). "Metamemory cues and monitoring accuracy: Judging what you know and what you will know." In: *Journal of Educational Psychology* 92.4, p. 800 (cit. on p. 64).

Kelley, Harold H (1973). "The processes of causal attribution." In: *American Psychologist* 28.2, p. 107 (cit. on p. 96).

Kendall, Alex and Yarin Gal (2017). "What uncertainties do we need in bayesian deep learning for computer vision?" In: *Advances in Neural Information Processing Systems*, pp. 5574–5584 (cit. on p. 63).

Kim, Beomsu, Junghoon Seo, Seunghyeon Jeon, Jamyoung Koo, Jeongyeol Choe, and Taegyun Jeon (2019). "Why are saliency maps noisy? cause of and solution to noisy saliency maps". In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, pp. 4149–4157 (cit. on p. 97).

Kindermans, Pieter-Jan, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim (2019). "The (un) reliability of saliency methods". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, pp. 267–280 (cit. on p. 97).

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (cit. on pp. 35, 68).

Kirkpatrick, James et al. (2017). "Overcoming catastrophic forgetting in neural networks". In: *Proceedings of the National Academy of Sciences* 114.13, pp. 3521–3526 (cit. on p. 76).

Krizhevsky, Alex, Geoffrey Hinton, et al. (2009). "Learning multiple layers of features from tiny images". In: (cit. on pp. 68, 82, 103).

Krizhevsky, Alex and Geoffrey E Hinton (2011). "Using very deep autoencoders for content-based image retrieval." In: *ESANN* (cit. on pp. 31, 41).

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems* 25, pp. 1097–1105 (cit. on pp. 3, 31, 41).

Kruger, Justin and David Dunning (1999). "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments." In: *Journal of Personality and Social Psychology* 77.6, p. 1121 (cit. on p. 64).

Kumar, Virendra et al. (2012). "Radiomics: the process and the challenges". In: *Magnetic Resonance Imaging* 30.9, pp. 1234–1248 (cit. on pp. 19, 31).

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell (2017). "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in Neural Information Processing Systems*, pp. 6402–6413 (cit. on pp. 62, 63).

Lambin, Philippe et al. (2012). "Radiomics: extracting more information from medical images using advanced feature analysis". In: *European Journal of Cancer* 48.4, pp. 441–446 (cit. on p. 31).

Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data". In: *Biometrics*, pp. 159–174 (cit. on p. 107).

LeCun, Yann, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel (1989). "Backpropagation applied to handwritten zip code recognition". In: *Neural Computation* 1.4, pp. 541–551 (cit. on pp. 50, 101).

Lee, Changhee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar (2018). "Deephit: A deep learning approach to survival analysis with competing risks". In: *Thirty-Second AAAI Conference on Artificial Intelligence* (cit. on pp. 49, 50).

Leijenaar, Ralph TH et al. (2013). "Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability". In: *Acta Oncologica* 52.7, pp. 1391–1397 (cit. on pp. 31, 45).

Li, Hui et al. (2016). "MR imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of MammaPrint, Oncotype DX, and PAM50 gene assays". In: *Radiology* 281.2, pp. 382–391 (cit. on p. 45).

Li, Xiao-Hui et al. (2020). "A survey of data-driven and knowledge-aware explainable ai". In: *IEEE Transactions on Knowledge and Data Engineering* (cit. on p. 97).

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár (2017). "Focal loss for dense object detection". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988 (cit. on pp. 62, 81).

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis (2021). "Explainable AI: A review of machine learning interpretability methods". In: *Entropy* 23.1, p. 18 (cit. on pp. 24, 96).

Liu, Shusen, Bhavya Kailkhura, Donald Loveland, and Yong Han (2019). "Generative counterfactual introspection for explainable deep learning". In: *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, pp. 1–5 (cit. on p. 98).

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems*, pp. 4765–4774 (cit. on pp. 61, 97, 103).

Maas, Andrew L, Awni Y Hannun, and Andrew Y Ng (2013). "Rectifier nonlinearities improve neural network acoustic models". In: *Proc. ICML*. Vol. 30. 1, p. 3 (cit. on pp. 32, 81).

Mackin, Dennis et al. (2015). "Measuring CT scanner variability of radiomics features". In: *Investigative radiology* 50.11, p. 757 (cit. on p. 45).

Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu (2018). "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *International Conference on Learning Representations* (cit. on p. 120).

Mamassian, Pascal, David C Knill, and Daniel Kersten (1998). "The perception of cast shadows". In: *Trends in Cognitive Sciences* 2.8, pp. 288–295 (cit. on p. 96).

Martín Abadi et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. URL: https://www.tensorflow.org/ (cit. on p. 35).

Ming, Yao, Huamin Qu, and Enrico Bertini (2018). "Rulematrix: Visualizing and understanding classifiers with rules". In: *IEEE transactions on visualization and computer graphics* 25.1, pp. 342–352 (cit. on p. 97).

Misiakos, Evangelos P, Nikolaos P Karidis, and Gregory Kouraklis (2011). "Current treatment for colorectal liver metastases". In: *World Journal of Gastroenterology: WJG* 17.36, p. 4067 (cit. on p. 15).

Moltz, Jan Hendrik (2019). "Stability of radiomic features of liver lesions from manual delineation in CT scans". In: *Medical Imaging 2019: Computer-Aided Diagnosis*. Vol. 10950. International Society for Optics and Photonics, 109501W (cit. on p. 21).

Moltz, Jan Hendrik, Stefan Braunewell, Jan Rühaak, Frank Heckel, Sebastiano Barbieri, Lennart Tautz, Horst K Hahn, and H-O Peitgen (2011). "Analysis of variability in manual liver tumor delineation in CT scans". In: *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, pp. 1974–1977 (cit. on p. 27).

Mongan, John P, Camilo E Fadul, Bernard F Cole, Bassem I Zaki, Arief A Suriawinata, Gregory H Ripple, Tor D Tosteson, and J Marc Pipas (2009). "Brain metastases from colorectal cancer: risk factors, incidence, and the possible role of chemokines". In: *Clinical Colorectal Cancer* 8.2, pp. 100–105 (cit. on p. 17).

Montavon, Grégoire, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller (2017). "Explaining nonlinear classification decisions with deep taylor decomposition". In: *Pattern Recognition* 65, pp. 211–222 (cit. on pp. 97, 103).

Mühlberg, Alexander et al. (2020). "The technome-a predictive internal calibration approach for quantitative imaging biomarker research". In: *Scientific reports* 10.1, pp. 1–15 (cit. on pp. 14, 21, 23, 45, 47).

Mühlberg, Alexander et al. (2021a). "The relevance of CT-based geometric and radiomics analysis of whole liver tumor burden to predict survival of patients with metastatic colorectal cancer". In: *European Radiology* 31.2, pp. 834–846 (cit. on pp. 47, 116).

Mühlberg, Alexander et al. (2021b). "Unraveling the Interplay of Image Formation, Data Representation and Learning in CT-based COPD Phenotyping Automation: The Need for a Meta-Strategy". In: *Medical Physics* (cit. on p. 21).

Mukherjee, Siddhartha (Jan. 3, 2018). "This Cat Sensed Death. What if Computers Could, Too?" In: *The New York Times Magazine*. URL: https://www.nytimes.com/2018/01/03/magazine/the-dying-algorithm.html (visited on 09/18/2021) (cit. on p. 116).

Nelson, Thomas O (1990). "Metamemory: A theoretical framework and new findings". In: *Psychology of Learning and Motivation*. Vol. 26. Elsevier, pp. 125–173 (cit. on p. 64).

Nelson, Wayne (1972). "Theory and applications of hazard plotting for censored failure data". In: *Technometrics* 14.4, pp. 945–966 (cit. on p. 87).

Ng, Andrew et al. (2011). "Sparse autoencoder". In: *CS294A Lecture notes* 72.2011, pp. 1–19 (cit. on pp. 31, 32, 41).

Nguyen, Anh, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune (2016). "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks". In: *Advances in Neural Information Processing Systems* 29, pp. 3387–3395 (cit. on p. 98).

Nibali, Aiden, Zhen He, and Dennis Wollersheim (2017). “Pulmonary nodule classification with deep residual networks”. In: *International Journal of Computer Assisted Radiology and Surgery* 12.10, pp. 1799–1808 (cit. on pp. 33, 82, 103).

Nie, Dong, Han Zhang, Ehsan Adeli, Luyan Liu, and Dinggang Shen (2016). “3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 212–220 (cit. on p. 40).

NIH-NCI (2017). “Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2015), National Cancer Institute, DCCPS, Surveillance Research Program, released April 2018, based on the November 2017 submission.” In: (cit. on p. 52).

— (2019). *National Institue of Health-National Cancer Institute - Overall Survival*. URL: `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/os` (visited on 04/01/2019) (cit. on p. 48).

— (2021). *National Institute of Health-National Cancer Institute - Mortality and Person-Years of Life Lost*. URL: `https://web.archive.org/web/https://progressreport.cancer.gov/tables/end` (visited on 08/25/2021) (cit. on pp. 13, 15).

Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han (2015). “Learning deconvolution network for semantic segmentation”. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1520–1528 (cit. on p. 97).

NVIDIA Corporation (2016). *NASA Program Deploys Deep Learning to Ward Off Asteroid Attack*. URL: `http://web.archive.org/web/20201125111935/https://blogs.nvidia.com/blog/2016/11/21/nasa-deep-learning-asteroids/` (visited on 07/06/2021) (cit. on p. 3).

Orhan, Emin and Xaq Pitkow (2018). “Skip Connections Eliminate Singularities”. In: *International Conference on Learning Representations* (cit. on p. 52).

Oxnard, Geoffrey R, Michael J Morris, F Stephen Hodi, Laurence H Baker, Mark G Kris, Alan P Venook, and Lawrence H Schwartz (2012). “When progressive disease does not mean treatment failure: reconsidering the criteria for progression”. In: *Journal of the National Cancer Institute* 104.20, pp. 1534–1541 (cit. on p. 40).

Parmar, Chintan, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo JWL Aerts (2015). “Machine learning methods for quantitative radiomic biomarkers”. In: *Scientific reports* 5, p. 13087 (cit. on p. 78).

Parsons, A, Amanda Daley, Rachna Begh, and Paul Aveyard (2010). “Influence of smoking cessation after diagnosis of early stage lung cancer on prognosis: systematic review of observational studies with meta-analysis”. In: *Bmj* 340 (cit. on p. 116).

Perslev, Mathias, Erik Bjørnager Dam, Akshay Pai, and Christian Igel (2019). “One network to segment them all: A general, lightweight system for accurate 3d medical image segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 30–38 (cit. on p. 33).

Pezzotti, Nicola, Thomas Höllt, Jan Van Gemert, Boudewijn PF Lelieveldt, Elmar Eisemann, and Anna Vilanova (2017). “Deepeyes: Progressive visual analytics for designing deep neural networks”. In: *IEEE transactions on visualization and computer graphics* 24.1, pp. 98–108 (cit. on p. 97).

Popat, Sanjay and Richard S Houlston (2005). “A systematic review and meta-analysis of the relationship between chromosome 18q genotype, DCC status and colorectal cancer prognosis”. In: *European Journal of Cancer* 41.14, pp. 2060–2070 (cit. on p. 40).

Powers, David Martin (2011). “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation”. In: *Journal of Machine Learning Technologies* 2.1, pp. 37–63 (cit. on p. 43).

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). “Language Models are Unsupervised Multitask Learners”. In: URL: https://openai.com/blog/better-language-models/ (cit. on pp. xi, 23).

Rao, Qing and Jelena Frtunikj (2018). “Deep learning for self-driving cars: Chances and challenges”. In: *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems*, pp. 35–38 (cit. on p. 3).

RBC Capital Markets (2021). *The healthcare data explosion*. URL: https://web.archive.org/web/https://www.rbccm.com/en/gib/healthcare/episode/the_healthcare_data_explosion (visited on 08/25/2021) (cit. on pp. 4, 17).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). “" Why should i trust you?" Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144 (cit. on pp. 61, 97).

— (2018). “Anchors: High-precision model-agnostic explanations”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1 (cit. on p. 97).

Rieke, Nicola et al. (2020). “The future of digital health with federated learning”. In: *NPJ Digital Medicine* 3.1, pp. 1–7 (cit. on pp. 86, 120).

Ries, Lynn A Gloeckler, Malcolm A Smith, JG Gurney, M Linet, T Tamra, JL Young, GRe Bunin, et al. (1999). “Cancer incidence and survival among children and adolescents: United States SEER Program 1975-1995.” In: *Cancer incidence and survival among children and adolescents: United States SEER Program 1975-1995.* (cit. on p. 52).

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 234–241 (cit. on pp. 23, 101, 156, 157).

Roser, Max and Hannah Ritchie (2015). “Cancer”. In: *Our World in Data*. URL: https://ourworldindata.org/cancer (cit. on p. 16).

Roth, Holger R, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers (2015). “Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 556–564 (cit. on p. 23).

Rothe, Jan Holger et al. (2013). “Size determination and response assessment of liver metastases with computed tomography—comparison of RECIST and volumetric algorithms”. In: *European Journal of Radiology* 82.11, pp. 1831–1839 (cit. on pp. 29, 34).

Santos, Iria, Luz Castro, Nereida Rodriguez-Fernandez, Alvaro Torrente-Patino, and Adrian Carballal (2021). “Artificial Neural Networks and Deep Learning in the Visual Arts: A review”. In: *Neural Computing and Applications*, pp. 1–37 (cit. on p. 3).

Sauerbrei, Willi and Patrick Royston (1999). “Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials”.

In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 162.1, pp. 71–94 (cit. on p. 88).

Scarselli, Franco, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini (2008). "The graph neural network model". In: *IEEE Transactions on Neural Networks* 20.1, pp. 61–80 (cit. on p. 55).

Schemper, Michael (1992). "Cox analysis of survival data with non-proportional hazard functions". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 41.4, pp. 455–465 (cit. on p. 50).

Schlemper, Jo, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert (2019). "Attention gated networks: Learning to leverage salient regions in medical images". In: *Medical Image Analysis* 53, pp. 197–207 (cit. on pp. 50, 55, 97).

Schumacher, M et al. (1994). "Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group." In: *Journal of Clinical Oncology* 12.10, pp. 2086–2093 (cit. on p. 88).

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626 (cit. on p. 97).

Sennaar, Kumba (2021). *Emerj - How America's 5 Top Hospitals are Using Machine Learning Today*. URL: `https://emerj.com/ai-sector-overviews/top-5-hospitals-using-machine-learning/` (visited on 08/25/2021) (cit. on p. 24).

Seymour, Lesley et al. (2017). "iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics". In: *The Lancet Oncology* 18.3, e143–e152 (cit. on p. 31).

Shallue, Christopher J and Andrew Vanderburg (2018). "Identifying exoplanets with deep learning: A five-planet resonant chain around kepler-80 and an eighth planet around kepler-90". In: *The Astronomical Journal* 155.2, p. 94 (cit. on p. 3).

Shankar, Devashish, Sujay Narumanchi, HA Ananya, Pramod Kompalli, and Krishnendu Chaudhury (2017). "Deep learning based large scale visual recommendation and search for e-commerce". In: *arXiv preprint arXiv:1703.02344* (cit. on p. 3).

Shen, Dinggang, Guorong Wu, and Heung-Il Suk (2017). "Deep learning in medical image analysis". In: *Annual Review of Biomedical Engineering* 19, pp. 221–248 (cit. on p. 4).

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017). "Learning important features through propagating activation differences". In: *International Conference on Machine Learning*. PMLR, pp. 3145–3153 (cit. on p. 97).

Silver, David et al. (2016). "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587, pp. 484–489 (cit. on p. 3).

Silver, David et al. (2017). "Mastering the game of go without human knowledge". In: *nature* 550.7676, pp. 354–359 (cit. on p. 3).

Silver, David et al. (2018). "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362.6419, pp. 1140–1144 (cit. on p. 3).

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2013). "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (cit. on pp. 61, 93, 94, 98).

Singla, Sumedha, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich (2019). "Explanation by progressive exaggeration". In: *arXiv preprint arXiv:1911.00483* (cit. on p. 98).

Song, Jiangdian, Yanjie Yin, Hairui Wang, Zhihui Chang, Zhaoyu Liu, and Lei Cui (2020). "A review of original articles published in the emerging field of radiomics". In: *European Journal of Radiology*, p. 108991 (cit. on p. 21).

Statista (2021). *Total amount of global healthcare data generated in 2013 and a projection for 2020*. URL: https://www.statista.com/statistics/1037970/global-healthcare-data-volume/ (visited on 08/25/2021) (cit. on p. 4).

Stoehlmacher, Jan, David J Park, Wu Zhang, Susan Groshen, Denice D Tsao-Wei, Mimi C Yu, and Heinz-Josef Lenz (2002). "Association between glutathione S-transferase P1, T1, and M1 genetic polymorphism and survival of patients with metastatic colorectal cancer". In: *Journal of the National Cancer Institute* 94.12, pp. 936–942 (cit. on p. 40).

Sun, Yi, Ding Liang, Xiaogang Wang, and Xiaoou Tang (2015). "Deepid3: Face recognition with very deep neural networks". In: *arXiv preprint arXiv:1502.00873* (cit. on p. 3).

Szegedy, Christian et al. (2015). "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9 (cit. on p. 24).

Tan, Mingxing and Quoc Le (2019). "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 6105–6114 (cit. on p. 103).

Tang, Shengpu, Aditya Modi, Michael Sjoding, and Jenna Wiens (2020). "Clinician-in-the-loop decision making: Reinforcement learning with near-optimal set-valued policies". In: *International Conference on Machine Learning*. PMLR, pp. 9387–9396 (cit. on p. 119).

Taylor, Graham (2018). *Efficient techniques for learning confidence - MIDL (Accessed: 2021-05-07)*. Youtube. URL: https://www.youtube.com/watch?v=YedM4Cslj0g (visited on 07/16/2021) (cit. on p. 65).

Teng, Hao-Wei et al. (2012). "BRAF mutation is a prognostic biomarker for colorectal liver metastasectomy". In: *Journal of Surgical Oncology* 106.2, pp. 123–129 (cit. on p. 40).

Thanh-Tung, Hoang and Truyen Tran (2020). "Catastrophic forgetting and mode collapse in GANs". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–10 (cit. on p. 99).

Thanikachalam, Kannan and Gazala Khan (2019). "Colorectal cancer and nutrition". In: *Nutrients* 11.1, p. 164 (cit. on p. 16).

TOFT, P (1996). "The Radon Transform: Theory and Implementation". In: *PhD thesis, Technical University of Denmark* (cit. on pp. 14, 142).

Torrado, Ruben Rodriguez, Philip Bontrager, Julian Togelius, Jialin Liu, and Diego Perez-Liebana (2018). "Deep reinforcement learning for general video game ai". In: *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, pp. 1–8 (cit. on p. 3).

Tulving, Endel and Stephen A Madigan (1970). "Memory and verbal learning". In: *Annual Review of Psychology* 21.1, pp. 437–484 (cit. on p. 64).

Van Cutsem, Eric et al. (2016). "ESMO consensus guidelines for the management of patients with metastatic colorectal cancer". In: *Annals of Oncology* 27.8, pp. 1386–1422 (cit. on p. 27).

Verma, Vikas, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio (2019). "Manifold mixup: Better representations by interpolating hidden states". In: *International Conference on Machine Learning*. PMLR, pp. 6438–6447 (cit. on p. 76).

Waite, Stephen, Arkadij Grigorian, Robert G Alexander, Stephen L Macknik, Marisa Carrasco, David J Heeger, and Susana Martinez-Conde (2019). "Analysis of perceptual expertise in radiology–Current knowledge and a new perspective". In: *Frontiers in Human Neuroscience* 13, p. 213 (cit. on p. 18).

Wang, Rui, Karthik Kashinath, Mustafa Mustafa, Adrian Albert, and Rose Yu (2020). "Towards physics-informed deep learning for turbulent flow prediction". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1457–1466 (cit. on p. 3).

Wang, Sida and Christopher Manning (2013). "Fast dropout training". In: *International Conference on Machine Learning*. PMLR, pp. 118–126 (cit. on p. 63).

Wang, Zhou, Eero P Simoncelli, and Alan C Bovik (2003). "Multiscale structural similarity for image quality assessment". In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Ieee, pp. 1398–1402 (cit. on p. 101).

Waters, Harriet Salatas (1982). "Memory development in adolescence: Relationships between metamemory, strategy use, and performance". In: *Journal of Experimental Child Psychology* 33.2, pp. 183–195 (cit. on p. 66).

Willemink, Martin J, Pim A de Jong, Tim Leiner, Linda M de Heer, Rutger AJ Nievelstein, Ricardo PJ Budde, and Arnold MR Schilham (2013a). "Iterative reconstruction techniques for computed tomography Part 1: technical principles". In: *European Radiology* 23.6, pp. 1623–1631 (cit. on pp. 14, 142).

Willemink, Martin J, Tim Leiner, Pim A de Jong, Linda M de Heer, Rutger AJ Nievelstein, Arnold MR Schilham, and Ricardo PJ Budde (2013b). "Iterative reconstruction techniques for computed tomography part 2: initial results in dose reduction and image quality". In: *European Radiology* 23.6, pp. 1632–1642 (cit. on pp. 14, 142).

World Medical Association (Nov. 2013). "World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects". In: *JAMA* 310.20, pp. 2191–2194. ISSN: 0098-7484. DOI: 10.1001/jama.2013.281053. eprint: https://jamanetwork.com/journals/jama/articlepdf/1760318/jsc130006.pdf. URL: https://doi.org/10.1001/jama.2013.281053 (cit. on p. 3).

Wu, Aaron, Ziyue Xu, Mingchen Gao, Mario Buty, and Daniel J Mollura (2016). "Deep vessel tracking: A generalized probabilistic approach via deep learning". In: *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, pp. 1363–1367 (cit. on p. 23).

Xiao, Jian et al. (2015). "Tumor volume reduction rate is superior to RECIST for predicting the pathological response of rectal cancer treated with neoadjuvant chemoradiation: Results from a prospective study". In: *Oncology Letters* 9.6, pp. 2680–2686 (cit. on p. 34).

Xie, Saining, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He (2017). "Aggregated residual transformations for deep neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500 (cit. on p. 75).

Yakubovskiy, Pavel (2020). *EfficientNet Keras*. URL: `https://github.com/qubvel/efficientnet` (visited on 09/15/2020) (cit. on pp. 86, 103).

Yamada, Yoshihiro, Masakazu Iwamura, Takuya Akiba, and Koichi Kise (2018). "Shakedrop regularization for deep residual learning". In: *arXiv preprint arXiv:1802.02375* (cit. on pp. 75–77).

Yang, Jiancheng, Rui Shi, and Bingbing Ni (2021). "MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 191–195 (cit. on pp. 82, 103).

Yao, Jiawen, Sheng Wang, Xinliang Zhu, and Junzhou Huang (2016). "Imaging biomarker discovery for lung cancer survival prediction". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 649–657 (cit. on p. 40).

Yao, Jiawen, Xinliang Zhu, Feiyun Zhu, and Junzhou Huang (2017). "Deep correlational learning for survival prediction from multi-modality data". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 406–414 (cit. on p. 40).

Yip, Stephen SF and Hugo JWL Aerts (2016). "Applications and limitations of radiomics". In: *Physics in Medicine & Biology* 61.13, R150 (cit. on p. 31).

Zagoruyko, Sergey and Nikos Komodakis (2016). "Wide residual networks". In: *Proceedings of the British Machine Vision Conference (BMVC)* (cit. on p. 75).

Zauber, Ann G et al. (2012). "Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths". In: *New England Journal of Medicine* 366.8, pp. 687–696 (cit. on p. 52).

Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba (2016). "Learning deep features for discriminative localization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929 (cit. on p. 97).

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A Efros (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232 (cit. on pp. 98–100).

Zutphen, Moniek van, Ellen Kampman, Edward L Giovannucci, and Fränzel JB van Duijnhoven (2017). "Lifestyle after colorectal cancer diagnosis in relation to survival and recurrence: a review of the literature". In: *Current colorectal cancer reports* 13.5, pp. 370–401 (cit. on pp. 16, 116).

A. Additional Background

Anatomical Orientations

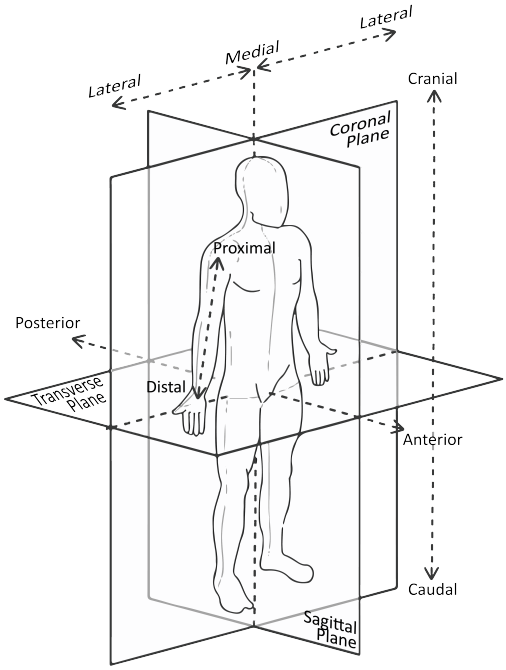


Figure A.1: The figure shows the anatomical planes and direction terms, most notably depicting the *transversal* (or *axial*), *saggital* and *coronal* planes, which are typically used to described image orientations in CT images. Source: [Edoarado 2021]

Computed Tomography Imaging

A short overview of the process of computed tomography imaging was already given in Chapter 2. Revisiting the figure from the beginning (see Fig. A.2) the process in the following will be described in some more detail.

In order to create a computed tomography image, the patient is slowly moved through a tube, containing a ring construction which is commonly denoted as *gantry*. The gantry holds both the X-ray source and detectors. Source and detectors are exactly opposed to each other, in order for the radiation to pass the human body before hitting the detectors. When passing the human body, the radiation is attenuated, depending on the characteristics of the crossed tissue. This attenuation can be measured on the opposite side by the detectors, if slice-wise concatenated yielding a so-called *sinogram*. The sinograms are collected over the scanning process, and can finally be transformed, i. e. *reconstructed*, into a volume by using an image transformation called *Radon transform* [TOFT 1996], which is related to the Fourier transformation. More specifically, in practice commonly an inverse transformation is used called *filtered backprojection* [Brooks et al. 1975]. Lately, other techniques, such as the iterative reconstruction [Willemink et al. 2013a,b] have been employed. Modern CT-scanners can acquire multiple slices at once (e. g. up to 384 with a Siemens SOMATOM Force) and achieve up to 4 full rotations a second [Healthineers 2021].

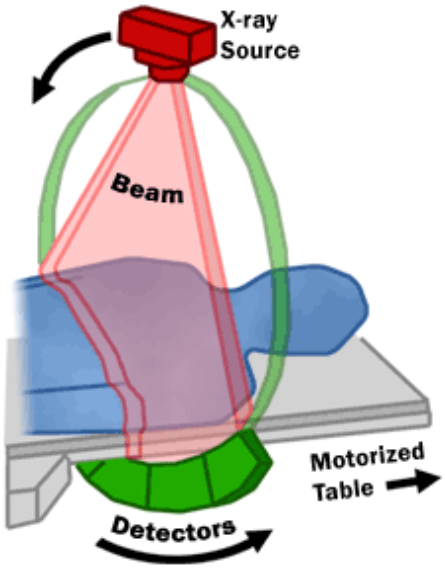
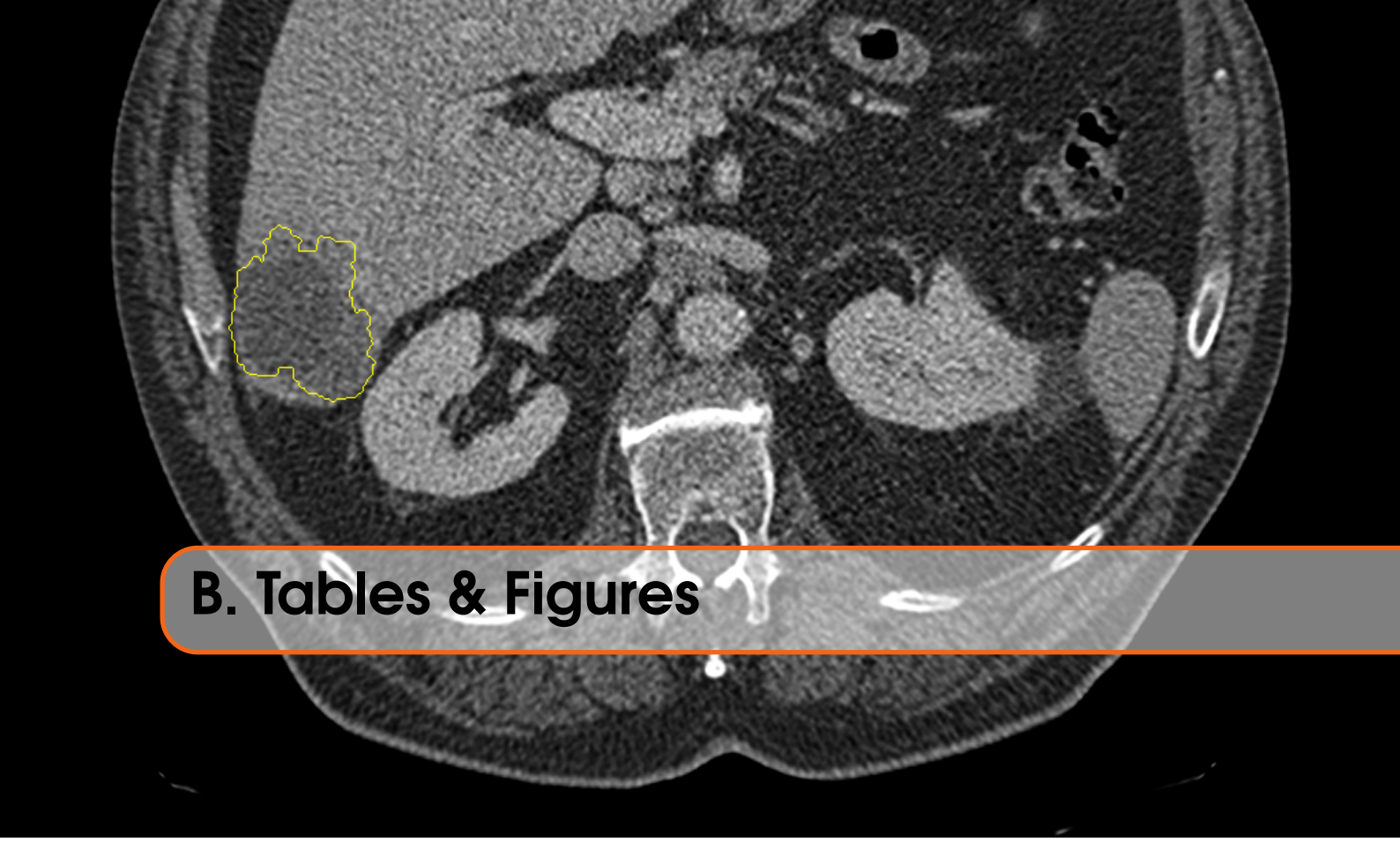


Figure A.2: This figure shows the basic concept of CT imaging. The patient is slowly moved through a *gantry*, holding X-ray source and detectors, being opposed to each other. The gantry is revolving around the patient, yielding linewise attenuation curves which are concatenated to a *sinogram*, which can finally be transformed into a volumetric array, i. e. the final CT image, by using an image reconstruction technique. Source: [FDA 2018]



B. Tables & Figures

In the following, additional tables and figures are depicted which are not necessarily needed for the understanding of the work, but which may give interesting additional information to the work presented. At multiple points within this work, it is referred to these figures if adequate.

type	filter size	strides	regularization	output	# params
in			BN	2×256×256×1	
conv3d	2×1×1	1×1×1	BN	256×256×32	224
conv2d	5×5	1×1	BN	256×256×32	25760
pool	4×4	4×4	—	64×64×32	
conv2d	5×5	1×1	BN	64×64×48	38640
pool	4×4	4×4	—	16×16×48	
conv2d	3×3	1×1	BN	16×16×64	27968
pool	2×2	2×2	—	8×8×64	
conv2d	3×3	1×1	BN	8×8×96	55776
pool	2×2	2×2	—	4×4×96	
conv2d	3×3	1×1	BN	4×4×128	111232
flatten				(2048)	
dense	(20)		L1+BN	(20)	41060
dense	(2048)		BN	(2048)	51200
reshape	(2048)		—	4×4×128	
dconv2d	3×3	1×1	BN	4×4×96	111072
up	2×2	1×1	—	8×8×96	
dconv2d	3×3	1×1	BN	8×8×64	55616
up	2×2	1×1	—	16×16×64	
dconv2d	3×3	1×1	BN	16×16×48	27888
up	4×4	1×1	—	64×64×48	
dconv2d	5×5	1×1	BN	64×64×32	38560
up	4×4	1×1	—	256×256×32	
dconv2d	5×5	1×1	BN	256×256×32	25760
dconv3d	2×1×1	1×1×1	BN	2×256×256×1	69

Table B.1: Network architecture of the sparse convolutional autoencoder used for pretraining the liver lesion growth predictor from [Katzmann et al. 2018b] which was used in Chapter 5. The table is based on the original work from [Katzmann et al. 2018b].

type	filter size	strides	regularization	output	# params
in				2×256×256×1	
conv3d	2×1×1	1×1×1	BN	256×256×32	224
conv2d	5×5	1×1	BN	256×256×32	25760
pool	4×4	4×4	—	64×64×32	
conv2d	5×5	1×1	BN	64×64×48	38640
pool	4×4	4×4	—	16×16×48	
conv2d	3×3	1×1	BN	16×16×64	27968
pool	2×2	2×2	—	8×8×64	
conv2d	3×3	1×1	BN	8×8×96	55776
pool	2×2	2×2	—	4×4×96	
conv2d	3×3	1×1	BN	4×4×128	111232
flat				(2048)	
fc	(20)		L1+BN	(20)	41060
dense	(8)		BN	(8)	200
dense	(2)		—	(2)	18

Table B.2: Predictor network architecture from [Katzmann et al. 2018b] used for liver lesion growth prediction in Chapter 5. Notably, all layers before the final dense layers are shared with the sparse autoencoder (see Tab. B.1). As dense layers tend to easily overfit, both are dimensioned with very few neurons, enforcing a latent space seperation in the convolutional layers. The table is based on the original work from [Katzmann et al. 2018b].

	type	filters	stride	regularization	output	# parameters
	input	–	–	BN	2x64x64x1 0:	0
	lambda	–	–	BN	64x64x1 1: 64x64x1	0
2x	conv	5x5	1x1	BN	64x64x32	960
	pool	2x2	2x2	–	32x32x32	–
	conv	5x5	1x1	BN	32x32x48	38640
	pool	2x2	2x2	–	16x16x48	–
	conv	3x3	1x1	BN	16x16x64	27968
	pool	2x2	2x2	–	8x8x64	–
	conv	3x3	1x1	BN	8x8x96	55776
	pool	2x2	2x2	–	4x4x96	–
	conv	3x3	1x1	BN	4x4x128	111232
	reshape	–	–	–	4096	–
	dense	–	–	BN	20	82020
2x	dense	–	–	BN	2048	51200
	reshape	–	–	–	4x4x128	–
	deconv	3x3	1x1	BN	4x4x96	111072
	up	2x2	2x2	–	8x8x96	–
	deconv	3x3	1x1	BN	8x8x64	55616
	up	2x2	2x2	–	16x16x64	–
	deconv	3x3	1x1	BN	16x16x48	27888
	up	2x2	2x2	–	32x32x48	–
	deconv	5x5	1x1	BN	32x32x32	38560
	up	2x2	2x2	–	64x64x32	–
	deconv	5x5	1x1	BN	64x64x1	932
	out	–	–	–	2x64x64x1	–

Table B.3: Network architecture of the used sparse convolutional autoencoder in Chapter 6 which was used for lesion growth and patient survival prediction. In contrast to the architecture from [Katzmann et al. 2018b], the network uses two separate processing lanes for baseline and followup images (cf. Tab. B.1). The table is based on the original work from [Katzmann et al. 2018a].

	type	filters	stride	regularization	output	# parameters
	input	–	–	BN	2x64x64x1	0
	lambda	–	–	BN	0: 64x64x1 1: 64x64x1	0
2x	conv	5x5	1x1	BN	64x64x32	960
	pool	2x2	2x2	–	32x32x32	–
	conv	5x5	1x1	BN	32x32x48	38640
	pool	2x2	2x2	–	16x16x48	–
	conv	3x3	1x1	BN	16x16x64	27968
	pool	2x2	2x2	–	8x8x64	–
	conv	3x3	1x1	BN	8x8x96	55776
	pool	2x2	2x2	–	4x4x96	–
	conv	3x3	1x1	BN	4x4x128	111232
	reshape	–	–	–	4096	–
	dense	–	–	BN	20	82020
	dense	–	–	BN	8	200
	softmax	–	–	–	2	18

Table B.4: Predictor network architecture for prediction lesion growth and patient survival as presented in Chapter 6 based on the sparse convolutional autoencoder which was described in Tab. B.3. The table based on the original work from [Katzmann et al. 2018a].

	Intuitivity	**Semantics**	**Quality**
Questionnaire results (raw)			
Ours	**1.92** [0.50, 3.13]	**1.76** [0.25, 3.13]	**1.04** [-0.63, 2.50]
DeepSHAP	-1.56* [-3.13, 0.13]	-1.54* [-3.13, 0.25]	-2.02* [-3.25,-0.63]
DeepTaylor	-1.52* [-3.25, 0.38]	-1.72* [-3.38, 0.13]	-0.85 [-2.50, 0.75]
LRP	-2.53** [-3.63,-1.25]	-2.57** [-3.75,-1.13]	-2.16* [-3.38,-0.75]
Questionnaire results (z-adjusted)			
Ours	**2.84** [1.50, 3.98]	**2.78** [1.39, 4.10]	**2.04** [0.50, 3.41]
DeepSHAP	-0.64* [-2.00, 0.86]	-0.52* [-1.97, 1.05]	-1.02* [-2.12, 0.16]
DeepTaylor	-0.60* [-2.16, 1.10]	-0.71* [-2.23, 1.00]	0.14 [-1.32, 1.57]
LRP	-1.61** [-2.72,-0.33]	-1.55** [-2.73,-0.11]	-1.16* [-2.24,-0.02]
Average rank			
Ours	**1.28** [1.00,1.75]	**1.36** [1.00,1.88]	**1.52** [1.06,2.19]
DeepSHAP	2.78** [2.19,3.38]	2.75* [2.13,3.31]	3.04* [2.50,3.50]
DeepTaylor	2.71* [2.06,3.31]	2.75* [2.06,3.38]	2.36 [1.75,3.00]
LRP	3.23*** [2.75,3.63]	3.14** [2.63,3.56]	3.07* [2.50,3.56]

Table B.5: Average questionnaire results (higher is better) and rank (lower is better) per algorithm on the LIDC-IDRI dataset from Chapter 10 with 95 % CI. p-values for two-tailed t-test with $t(6)$ are indicated as $p < .05$*, $p < .01$**, $p < .001$***. The best result is marked bold. Source: [Katzmann et al. 2021]

	Intuitivity	**Semantics**	**Quality**
Questionnaire results (raw)			
Ours	**1.45** [0.36, 2.45]	**1.33** [0.27, 2.36]	**1.72** [0.64, 2.64]
DeepSHAP	-1.77** [-2.82,-0.64]	-1.95** [-3.00,-0.82]	-1.42** [-2.55,-0.27]
DeepTaylor	0.91 [-0.55, 2.18]	0.90 [-0.55, 2.18]	0.75 [-0.64, 2.00]
LRP	-1.41** [-2.55,-0.18]	-1.63** [-2.73,-0.45]	-1.19** [-2.36, 0.00]
Questionnaire results (z-adjusted)			
Ours	**1.66** [0.65, 2.57]	**1.67** [0.63, 2.62]	**1.76** [0.70, 2.71]
DeepSHAP	-1.56** [-2.63, 0.39]	-1.62** [-2.70,-0.43]	-1.38** [-2.47,-0.20]
DeepTaylor	1.11 [-0.26, 2.37]	1.24 [-0.12, 2.49]	0.78 [-0.53, 1.99]
LRP	-1.21** [-2.33, 0.02]	-1.29** [-2.41,-0.07]	-1.15** [-2.29, 0.04]
Average rank			
Ours	**1.75** [1.36,2.23]	**1.77** [1.32,2.32]	**1.64** [1.23,2.14]
DeepSHAP	3.26** [2.82,3.64]	3.23** [2.77,3.64]	3.12** [2.59,3.59]
DeepTaylor	1.99 [1.41,2.64]	1.99 [1.45,2.64]	2.27 [1.73,2.86]
LRP	3.00** [2.45,3.45]	3.02* [2.50,3.45]	2.97* [2.36,3.50]

Table B.6: Average questionnaire results (higher is better) and rank (lower is better) per algorithm on the BreastMNIST dataset from Chapter 10 with 95 % CI. p-values for two-tailed t-test with $t(6)$ are indicated as $p < .05$*, $p < .01$**. The best result is marked bold. Source: [Katzmann et al. 2021]

	Intuitivity	**Semantics**	**Quality**
Questionnaire results (raw)			
Ours	0.67 [-0.67, 1.89]	0.72 [-0.78, 2.11]	-0.12 [-1.44, 1.22]
DeepSHAP	-1.41 [-2.78, 0.00]	-1.48 [-2.89, 0.00]	-0.23 [-2.00, 1.44]
DeepTaylor	**1.91** [0.56, 3.00]	**2.16** [0.78, 3.22]	**0.70** [-0.78, 2.11]
LRP	-1.81* [-3.00,-0.44]	-1.91* [-3.22,-0.44]	-1.45 [-2.56, 0.33]
Questionnaire results (z-adjusted)			
Ours	0.83 [-0.47, 2.03]	0.85 [-0.58, 2.13]	0.08 [-1.18, 1.36]
DeepSHAP	-1.25 [-2.60, 0.16]	-1.35 [-2.71, 0.07]	-0.03 [-1.83, 1.69]
DeepTaylor	**2.07** [0.75, 3.12]	**2.29** [0.96, 3.32]	**0.90** [-0.58, 2.28]
LRP	-1.65* [-2.88,-0.32]	-1.79* [-3.07,-0.36]	-0.95 [-2.39, 0.48]
Average rank			
Ours	2.13 [1.61,2.72]	2.15 [1.61,2.72]	2.61 [1.94,3.27]
DeepSHAP	3.04 [2.50,3.50]	3.06 [2.50,3.56]	2.30 [1.61,3.00]
DeepTaylor	**1.52** [1.06,2.17]	**1.50** [1.06,2.11]	**2.16** [1.50,2.89]
LRP	3.30* [2.83,3.72]	3.29* [2.78,3.67]	2.92 [2.39,3.44]

Table B.7: Average questionnaire results (higher is better) and rank (lower is better) per algorithm on the CIFAR-10 cats-vs.-dogs dataset from Chapter 10 with 95 % CI. p-values for two-tailed t-test with $t(7)$ are indicated as $p < .05$*. The best result is marked bold. Source: [Katzmann et al. 2021]

	Intuitivity	**Semantics**	**Quality**
	Questionnaire results (raw)		
Ours	0.95 [-0.67, 2.33]	1.49 [0.00, 2.67]	0.49 [-1.00, 1.83]
DeepSHAP	-1.78* [-3.33, 0.00]	-1.97* [-3.50,-0.17]	-1.51 [-3.33, 0.33]
DeepTaylor	**2.43** [0.66, 3.50]	**2.70** [1.17, 3.67]	**1.75** [-0.17, 3.33]
LRP	-2.20* [-3.66,-0.33]	-2.46** [-3.83,-0.83]	-1.87 [-3.50,-0.00]
	Questionnaire results (z-adjusted)		
Ours	1.10 [-0.28, 2.30]	1.55 [0.33, 2.54]	0.77 [-0.67, 2.10]
DeepSHAP	-1.64* [-3.30, 0.18]	-1.91* [-3.54,-0.04]	-1.22* [-3.01, 0.67]
DeepTaylor	**2.58** [1.25, 3.58]	**2.76** [1.53, 3.60]	**2.03** [0.31, 3.51]
LRP	-2.05* [-3.65,-0.18]	-2.40** [-3.78,-0.72]	-1.58* [-3.27, 0.30]
	Average rank		
Ours	2.12 [1.67,2.67]	1.94 [1.58,2.33]	2.39 [1.75,3.08]
DeepSHAP	3.16 [2.67,3.58]	3.24* [2.75,3.58]	2.83 [2.00,3.50]
DeepTaylor	**1.32** [1.00,2.00]	**1.31** [1.00,1.92]	**1.72** [1.08,2.50]
LRP	3.40* [2.83,3.83]	3.51** [3.17,3.83]	3.05 [2.33,3.67]

Table B.8: Average questionnaire results per algorithm (higher is better) and rank (lower is better) on the CIFAR-10 trucks-vs.-cars dataset from Chapter 10 with 95 % CI. p-values for two-tailed t-test with $t(4)$ are indicated as $p < .05$*, $p < .01$**. The best result is marked bold. Source: [Katzmann et al. 2021]

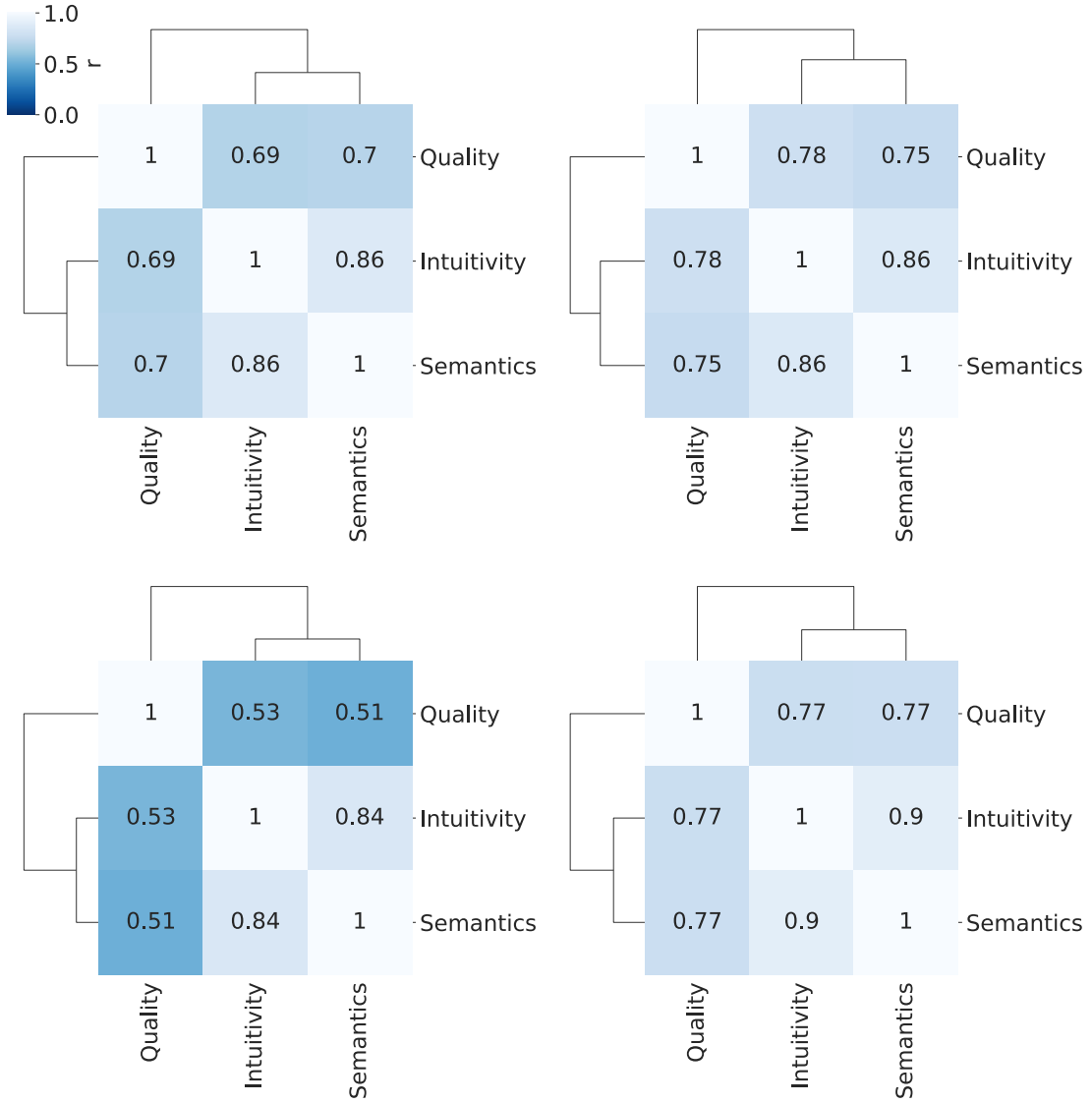


Figure B.1: Dendrogram of the pairwise Pearson correlation matrices for the analyzed criteria on the LIDC-IDRI (top-left), BreastMNIST (top-right), cats-vs.-dogs (bottom-left) and trucks-vs.-cars (bottom-right) datasets in Chapter 10. Source: [Katzmann et al. 2021]

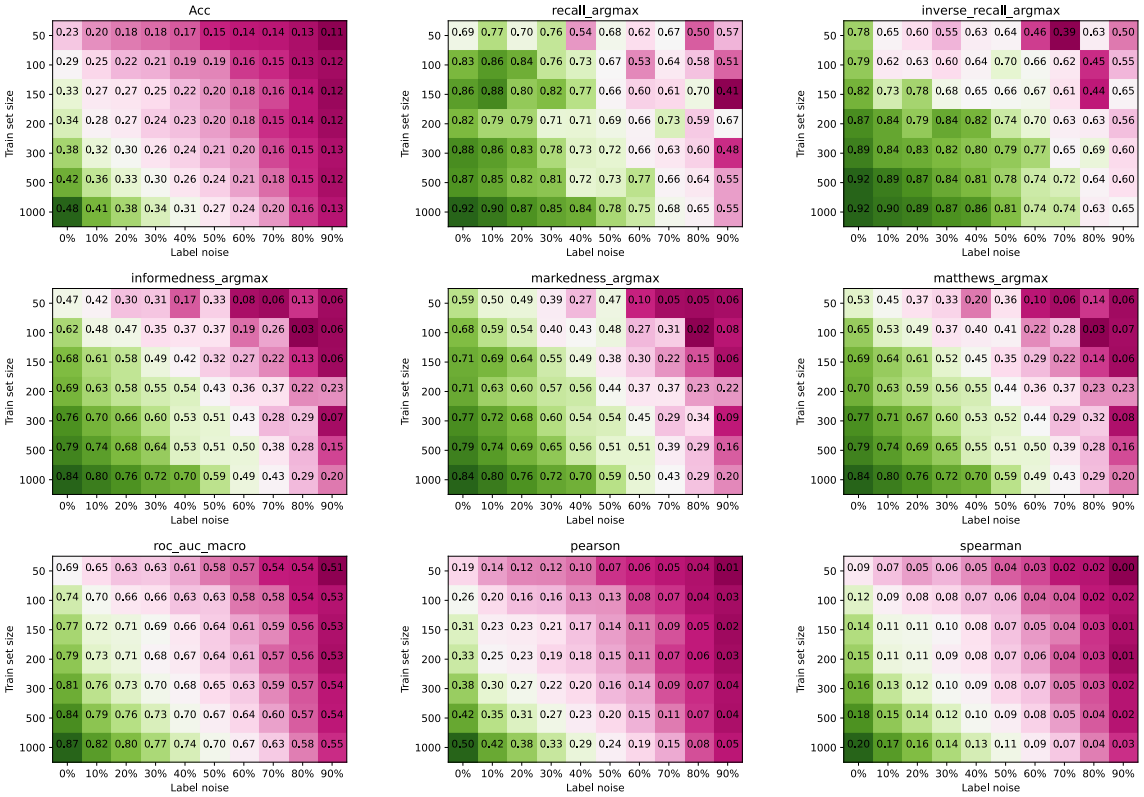


Figure B.2: Raw results on the CIFAR-10 dataset for the classifier ensemble evaluated in Chapter 9.2.5 with respect to (f.l.t.r.) accuracy, recall, inverse recall (top), informedness, markedness, matthews correlation coefficient (middle), AUC, Pearson product-moment and Spearman correlation (bottom).

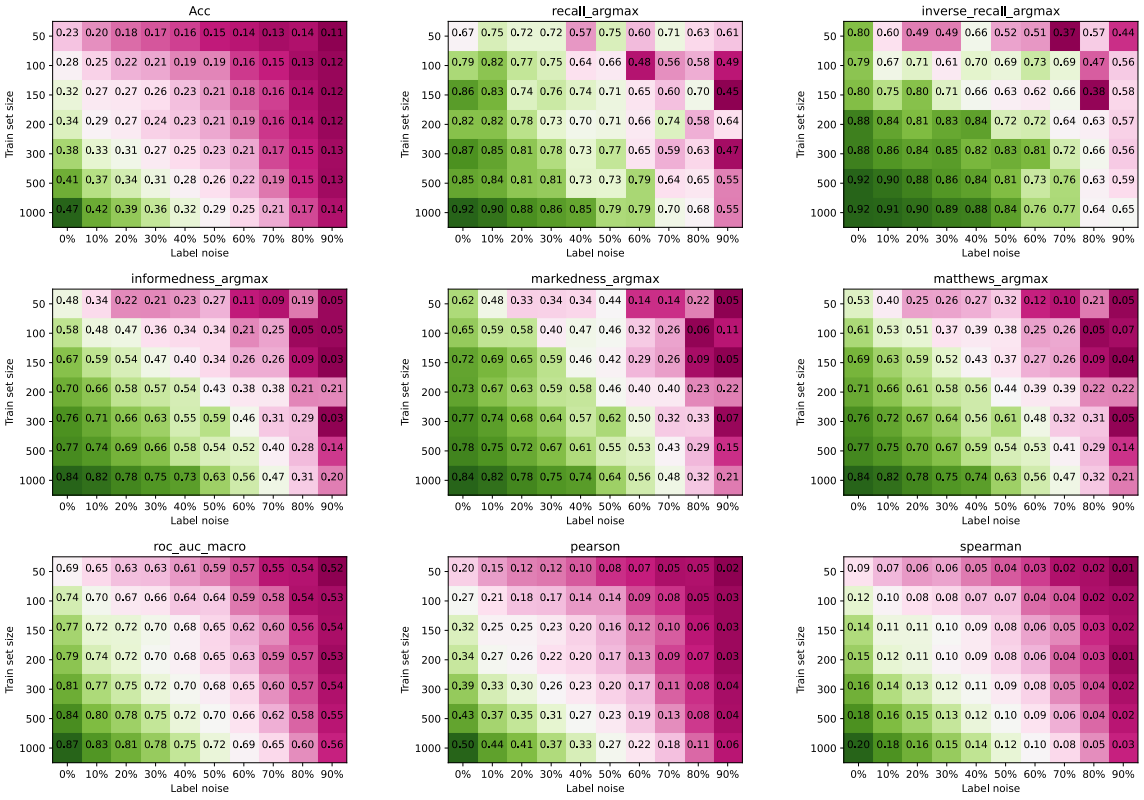


Figure B.3: Raw results on the CIFAR-10 dataset for the BTNet evaluated in Chapter 9.2.5 with respect to (f.l.t.r.) accuracy, recall, inverse recall (top), informedness, markedness, matthews correlation coefficient (middle), AUC, Pearson product-moment and Spearman correlation (bottom).

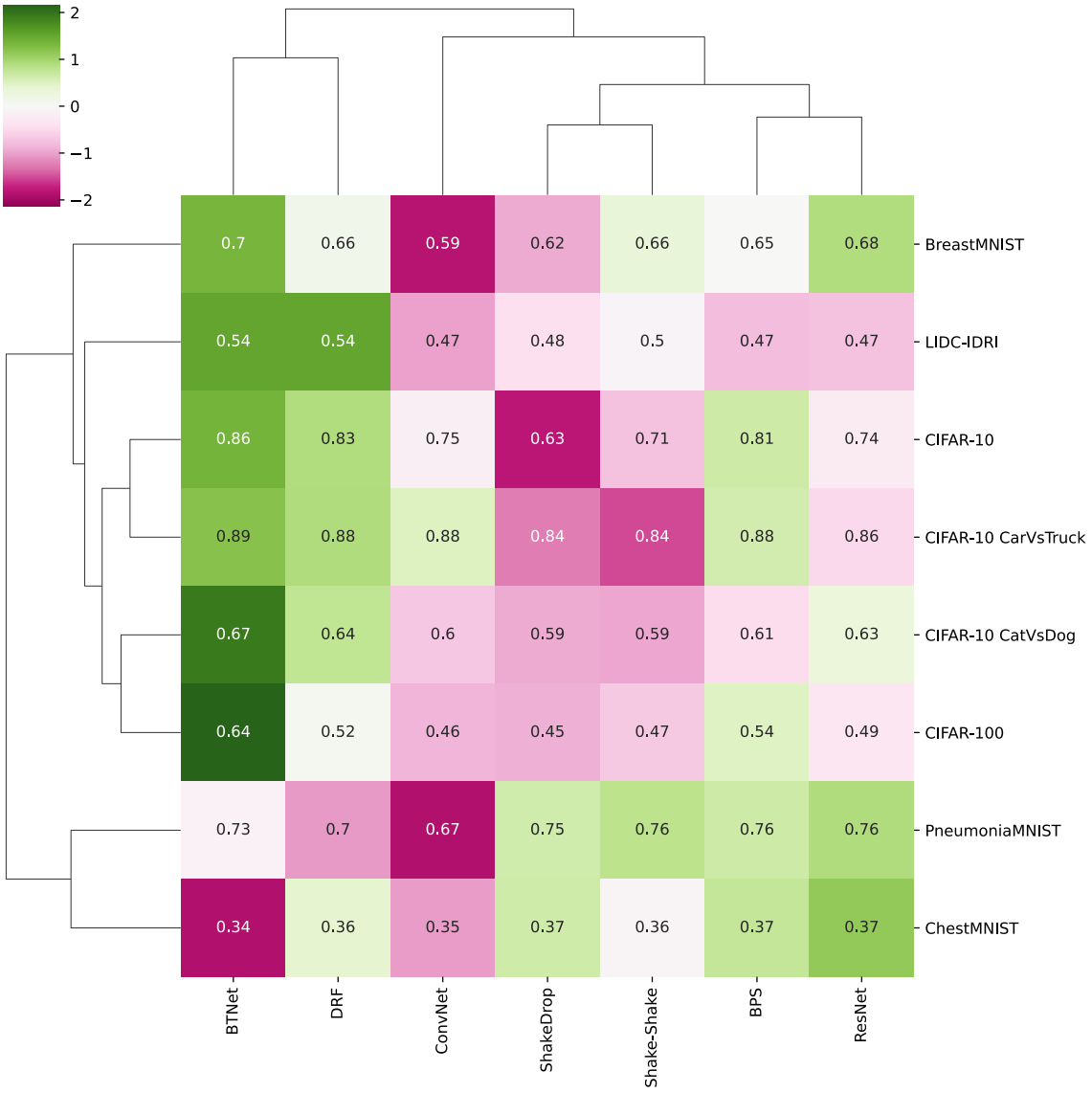


Figure B.4: Dendrogram showing the results of the bootstrapped and non-bootstrapped classifiers from Chapter 9.2.4 on medical (LIDC-IDRI, BreastMNIST, PneumoniaMNIST, ChestMNIST) and non-medical (CIFAR-10, its subsets, and CIFAR-100) imaging datasets with respect to the Matthews correlation coefficient (MCC). Results are z-normalized per dataset to account for varying dataset difficulty. Original values are annotated per cell. As shown by the linkage, BTNet and DRF, both being bootstrapped ensembling methods, follow a similar pattern across all datasets. Similarly, all residual approaches are linked together, with the strongest similarities between the Shake-Shake and ShakeDrop as well as the BPS and plain ResNet approach. PneumoniaMNIST and ChestMNIST, both being low-quality, high sample amount

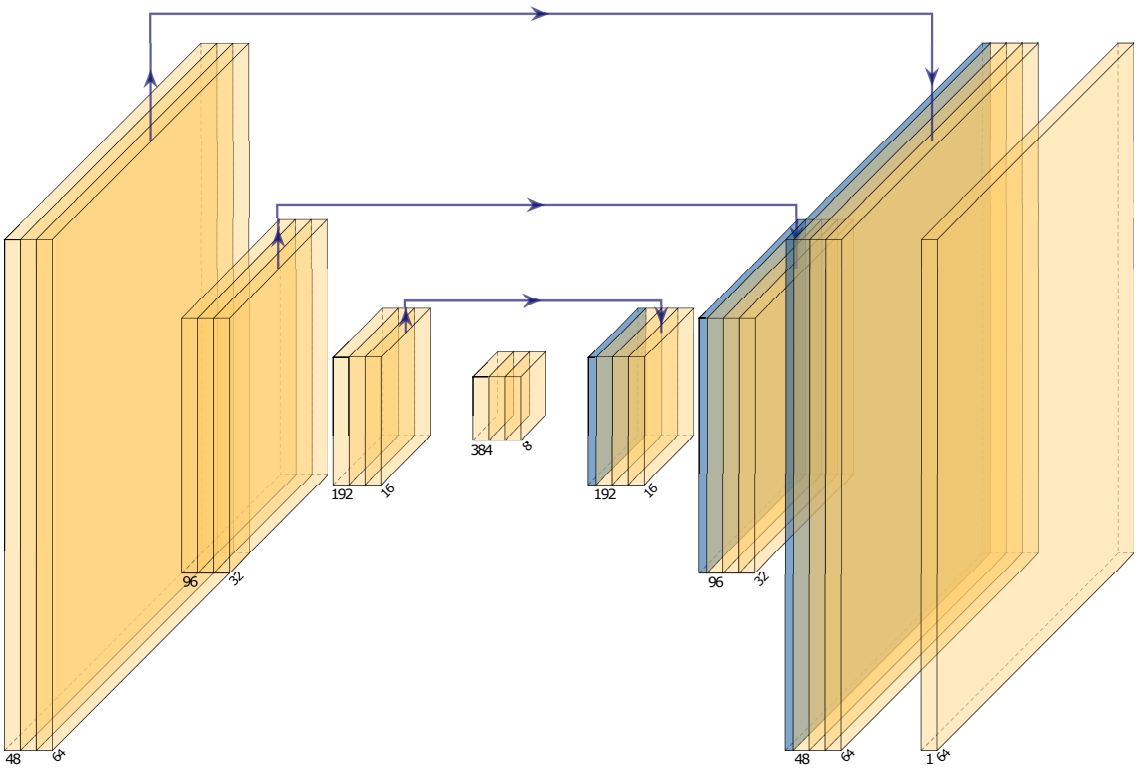


Figure B.5: Generator architecture used for the cycle-consistent deep decision explanation from Chapter 10 based on the UNet architecture from Ronneberger et al. [2015]. Source: [Katzmann et al. 2021]

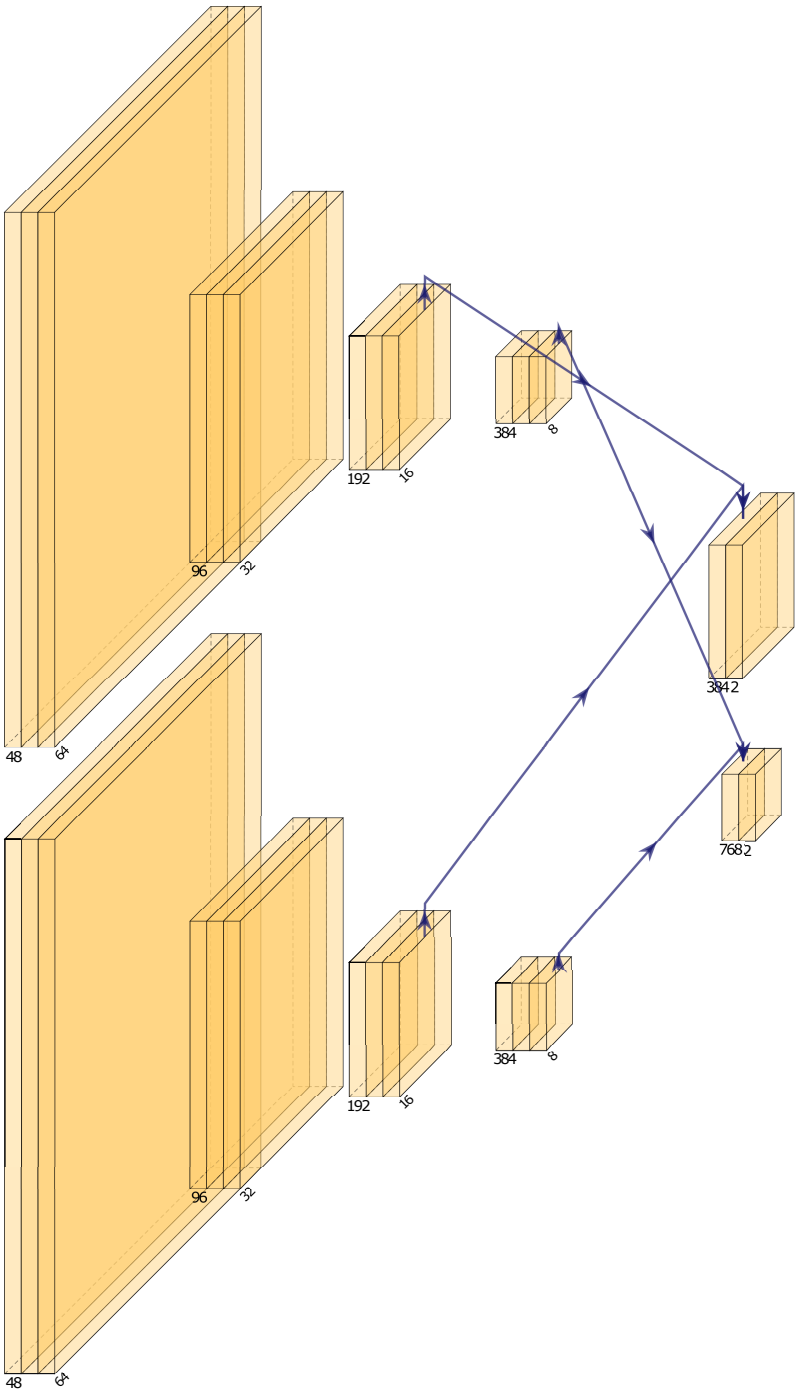
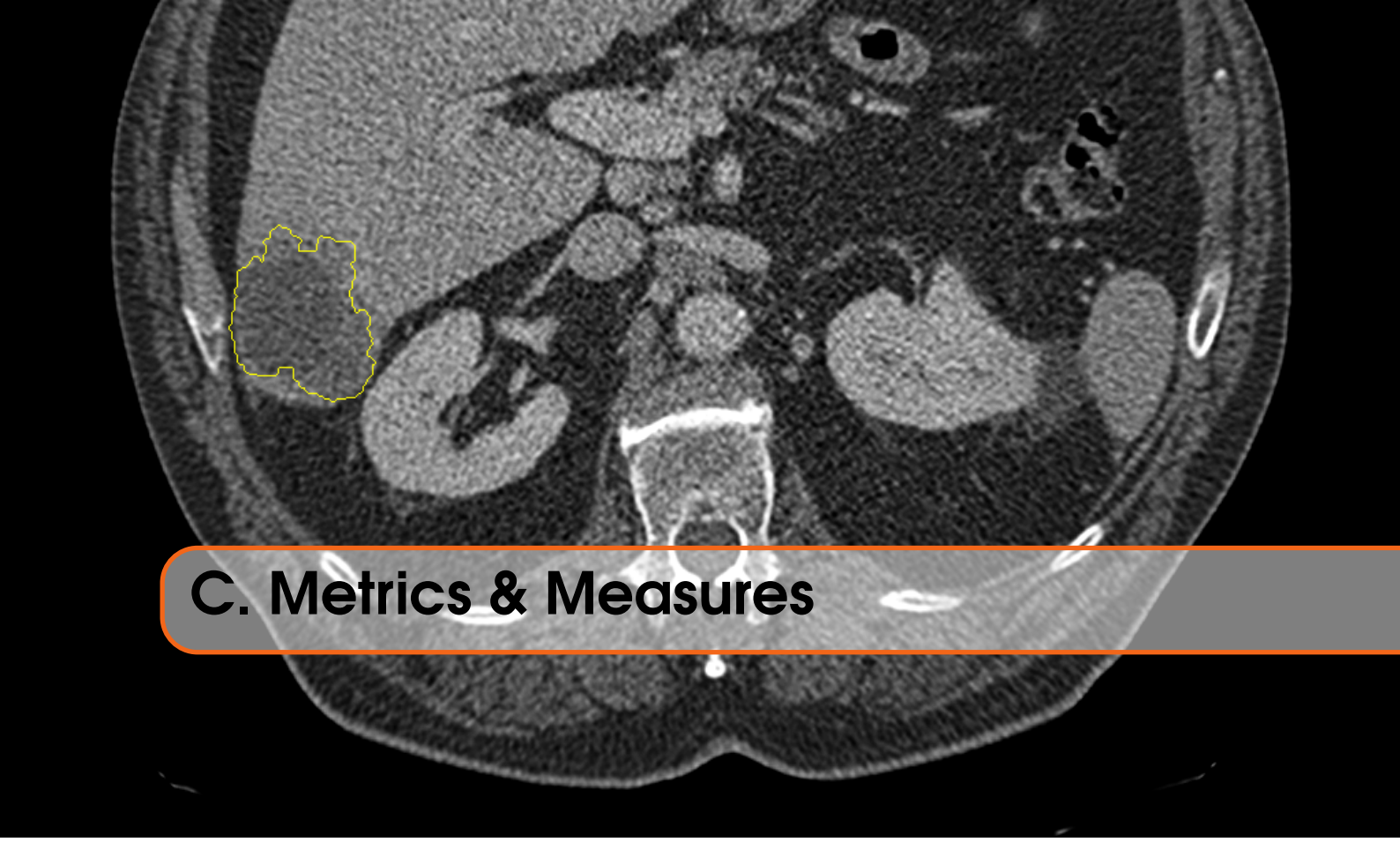


Figure B.6: Discriminator architecture used for the cycle-consistent deep decision explanation from Chapter 10 based on the UNet architecture from Ronneberger et al. [2015] (cf. Fig. B.5) and the PatchGAN approach from [Isola et al. 2017]. Source: [Katzmann et al. 2021]



C. Metrics & Measures

Throughout this work, a variety of algorithms has been introduced and quantitatively evaluated. As partly different goals have to be achieved by the proposed architectures, a variety of evaluation measures has been employed. These will briefly be discussed within this part.

Binary Classification Metrics

An overview of all used classification metrics within this work is found in Tab. C.1. First, it should be noted that the meaning of a value for one specific metric can largely depend on the context. Accuracy, for example, is a widely used metric that provides a highly intuitive understanding. It is, however, highly prone to misinterpretations due to label imbalance, and thus only applicable to balanced datasets. This similarly applies to other widely used metrics, such as sensitivity (true positive rate), specificity (true negative rate), positive, and negative predictive value (precision/inverse precision). As pointed out multiple times within this thesis, medical imaging data often have highly unequal label distributions (cf. Chapters 5 and 6), creating a need for using balanced measures if possible.

One of these measures is the F1 coefficient (see Tab. C.1), also known as Sørensen-Dice or Dice-similarity coefficient. A major downside, however, is that it does not take into account the number of true negatives, giving equal weight to precision and recall. First, depending on the concrete use case, this weighting is highly debatable (cf. [Hand et al. 2018]). Secondly, in nearly all of the application scenarios assessed within this work, e. g. lesion growth prediction, the true negative class is in fact highly important. Finally, it should be noted, that the formula of the F1 makes it prone to changes in the problem definition (e. g. prediction of *growth* vs. prediction of *non-growth* formulated as being the positive class).

The concrete class weight in a clinical setting is mostly problem-dependent and is related to the costs and/or acceptance problems resulting from false positives and negatives,

respectively. As a trade-off, one could suggest using balanced metrics where possible, such as informedness, markedness, or the Matthews correlation coefficient (MCC). Especially the latter has a variety of benefits, as it combines the former two, is zero-centered, and has an intuitive interpretation, which is that a value of zero corresponds to guessing, while a value of one corresponds to a perfect correlation[1].

Measure	Equation
true positive rate, sensitivity, recall	$TPR = \frac{tp}{tp+fn}$
true negative rate, specifity, inverse recall	$TNR = \frac{tn}{tn+fp}$
positive predictive value, precision	$PPV = \frac{tp}{tp+fp}$
negative predictive value, inverse precision	$NPV = \frac{tn}{tn+fn}$
F_1 score	$F_1 = \frac{2 \cdot tp}{2 \cdot tp+fp+fn}$
Informedness Youden's J statistic	$IFD = TPR + TNR - 1$
Markedness	$MKD = PPV + NPV - 1$
Matthews correlation coefficient/ϕ-coefficient	$MCC = \sqrt{IFD \cdot MKD}$

Table C.1: Overview on the classification metrics which are used within this work. The variables tp, tn, fp, fn represent the numbers of true positives, true negatives, false positives and false negatives given predictions and ground truth labels in a binary classification problem.

Non-binary Classification, Micro and Macro Averaging

In case of non-binary classification problems, binary metrics can be applied if the multiclass properties of the problem are explicitely taken into account. To achieve this, two possible ways are the use of micro- and macro-averaging. The basic idea of both is to binarize the problem and to apply the metrics accordingly. The main difference is the point of aggregation. More specifically, assuming a metric m which takes true positives tp, true negatives tn, false positives fp, and false negatives fn, applying macro averaging corresponds to first class-wise counting tp_c, tn_c, fp_c, fn_c and then to subsequently calculate m_c

[1] In fact, as the so-called ϕ-coefficient the Matthews correlation coefficient is widely used in statistics and for binary classification reduces to the Pearson's product-moment correlation coefficient.

for each class $c \in C$ with

$$m_c = m(tp_c, tn_c, fp_c, fn_c) \tag{C.1}$$

and deriving the final metric value m_{macro} as:

$$m_{\text{macro}} = \frac{1}{|C|} \sum_{c}^{C} m_c \tag{C.2}$$

Micro-averaging follows a similar approach, but aggregates at the counting level (i. e. micro), rather than averaging after calculating the metrics (*macro*), i. e.:

$$m_{\text{micro}} = m(\sum_{c}^{C} tp_c, \sum_{c}^{C} tn_c, \sum_{c}^{C} fp_c, \sum_{c}^{C} fn_c) \tag{C.3}$$

Generally, both options are valid but are somewhat different in their interpretation. While macro averaging provides a better overview of the accuracy for each class, it is more prone to outliers, e. g. classes that are very well or only poorly recognized. Micro averaging, on the other hand, better takes care of the number of samples in each class, as aggregation is applied only once. It is thus less prone to outliers and can be applied to imbalanced data without creating misleading results. It is, however, less applicable when the distributions between the evaluation and the final application differ, as it can happen when stratifying training or validation data. It is thus important that the used averaging is explicitly denoted for multiclass problems.

Regressive Evaluation Metrics

Pearson correlation coefficient

The Pearson correlation coefficient[2] r is a measure of the linear dependency of two variables. It can take values in the interval $[-1, 1]$, with -1 representing a perfect negative, and $+1$ a perfect positive correlation, respectively. No (linear) correlation is given if the coefficient is 0. It is calculated as:

$$r(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]\mathbb{E}[(Y - \mathbb{E}[Y])^2]}} \tag{C.4}$$

i. e. for two vectors x, y as:

$$r(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{C.5}$$

Spearman's rank correlation coefficient

Spearman's rank correlation coefficient r_s (or ρ) is defined analogously to the Pearson correlation coefficient, but uses the rank transformations $R(x), R(y)$ of the variables x, y:

$$r_s = r(R(x), R(y)) = \frac{cov(R(c), R(p_y))}{\sigma_{R(c)} \sigma_{R(p_y)}} \tag{C.6}$$

[2] The Pearson, or rarely Bravais-Pearson correlation coefficient, is named after Karl Pearson and Auguste Bravais. Its first use dates back to at least 1844 [Bravais 1844].

Mean squared / Mean absolute / Median absolute error

Assuming ground truth samples y and predictions $\hat{y}$, the mean squared error is given by:

$$\text{MSE} = \frac{1}{N}\sum_{1}^{N} ||y_{-}\hat{y}_i||^2 \tag{C.7}$$

with $N = |Y|$ being the number of samples. The mean absolute error is constructed analogously as

$$\text{MAE} = \frac{1}{N}\sum_{1}^{N} |y_i - \hat{y}_i| = \frac{1}{N} e_i \tag{C.8}$$

Finally, the median absolute error MedAE is the median $\tilde{e}$.

While there are a variety of metrics that are much more common for regression, such as the R^2 measure, these were not relevant with respect to the work at hand, as regression was mainly conducted as a side-topic of survival regression.

Survival Metrics

For survival estimation, such as in Chapter 7, the most commonly used evaluation metric is Harrel's concordance index (CI, or C-index) *CI*, which is calculated by counting concordant and discordant pairs of survival estimates. Concordance denotes the correctness of the order of the estimated survival times with respect to the ground truth. Assuming estimated survival times $\hat{Y} = \{\hat{y}_1, ..., \hat{y}_N\}$ and ground truth data $Y = \{y_1, ..., y_N\}$ the *CI* can be calculated as:

$$CI = \frac{\sum_{i \neq j} |\{(i,j)|(y_i < y_j) \wedge (\hat{y}_i < \hat{y}_j)\}|}{\sum_{i \neq j} |\{(i,j)|(y_i < y_j)\}|} \tag{C.9}$$

More intuitively, it represents the fraction of concordant pairs, i. e. pairs that have the same time order in both the predicted as well as the ground truth data, and takes into account each of these possible pairings (T_i, T_j). A major downside of the C-index is that it only takes into account time order, but not absolute times. It is thus an ordinal measure and does not give any information beyond this scale level. If exemplarily an algorithm would predict the double survival times for each patient, the resulting CI would be unchanged as long as the order is kept. This would even then be the case if the time differences between the patients would be completely arbitrary, being the main reason for the downsides of the Cox proportional hazards model (cf. Chapter 7).

Of course, regression metrics, such as the mean (MAE) or median absolute error (MedAE), can be applied to the problem of survival estimation as with standard regression.

Area under the AUC

A major downside of Harrel's C-index is that it only takes into account relative order. While it is therefore practical for *risk stratification*, i. e. by answering which patient has the higher survival in direct comparison, it does not allow the assessment of the absolute prediction quality over time. The area under the AUC (AUAUC), analogously to the AUC, binarizes the problem at each timepoint $\{y_i, \hat{y}_i\}$ for ground truth and estimated survival

times $Y, \hat{Y}$, respectively. All patients k are assigned a temporary label y_k and a temporary, binary prediction $\hat{y_k}$ for each time step as:

$$y_{k,i} = \begin{cases} 1 & T_k \geq y_i \\ 0 & \text{else} \end{cases} \tag{C.10}$$

$$\hat{y_{k,i}} = \begin{cases} 1 & \hat{y}_k \geq y_i \\ 0 & \text{else} \end{cases} \tag{C.11}$$

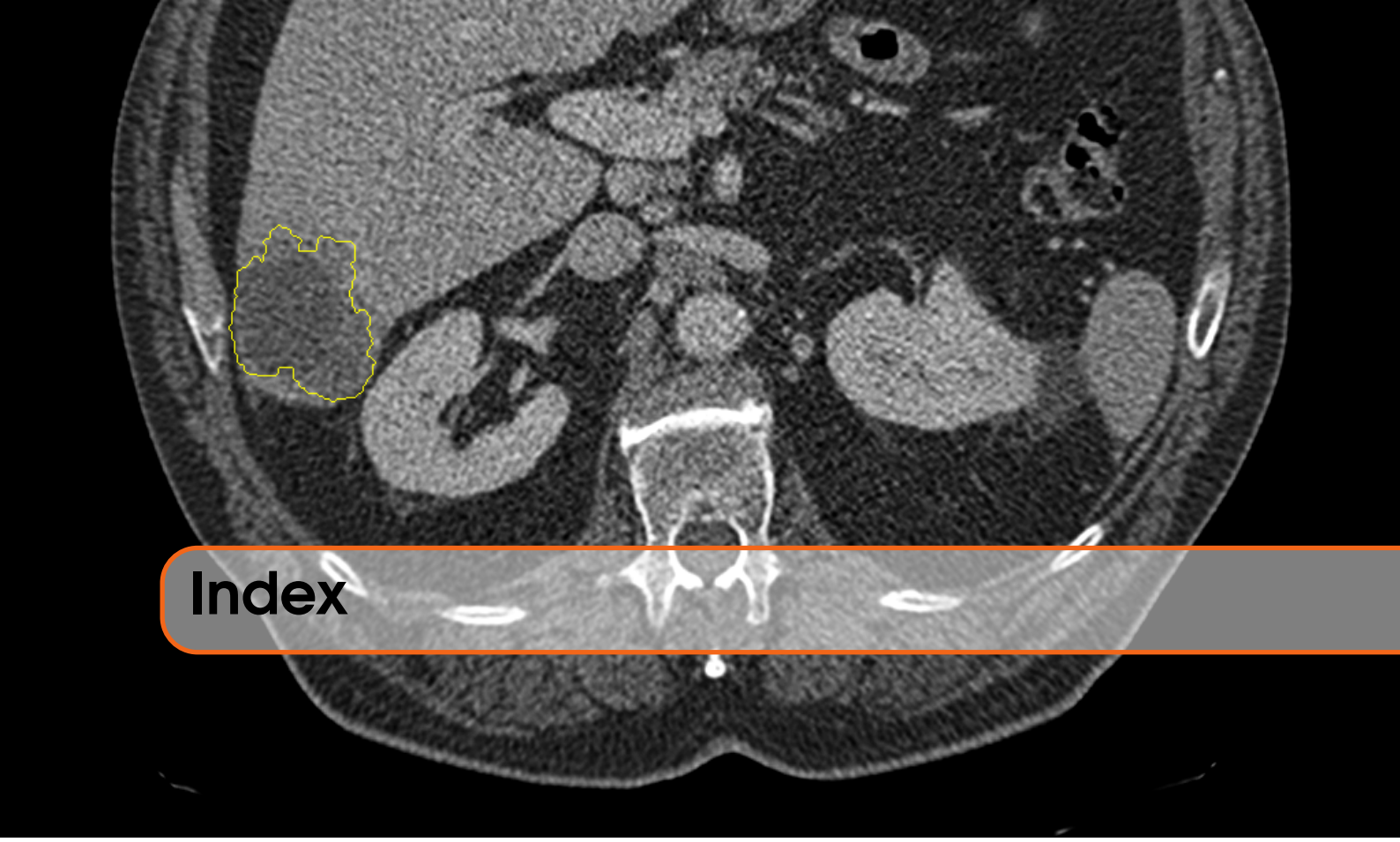
This temporary prediction can now be used to calculate an AUC for each timestep. Finally, an integral over the resulting curve can numerically be calculated, e. g. using trapezoid or Simpon's rule. This integral is then normalized by the range $(y_{\max} - y_{\min})$, yielding a value which, analogously to the original AUC, is in the range $[0,1] \in \mathbb{R}$.

Distributional inequality

Gini coefficient

Several times within this work, a measure for distributional inequality was needed, e. g. for the quantification of the confidence estimation quality in Chapter 8 and the split calculation in Chapter 9. For this purpose, the use of the Gini coefficient is widely common. The Gini coefficient takes values in the range $[0,1] \in \mathbb{R}$, with a value of 0 indicating an equal distribution and a value of 1 indicating a distribution in which all but one bin have a probability of zero with this one bin having a probability of one. As already described in Eq. 8.9, the Gini coefficient for a discrete probability distribution p can be calculated as:

$$\frac{\sum_i^m \sum_k^m |p_i - p_k|}{2n \sum_i^m p_i} \tag{C.12}$$



Index

A

Acquisition protocol 21
Activation maximization 98
Activations
  Leaky ReLU 32, 68, 81
  ReLU 75
  Tanh 35
Adversarial samples 118

B

Batch normalization 41, 75
Bias-variance dilemma 79
Bootstrapping 73, 102
Bootstrapping architectures 73, 87
  Bootstrapped Networks (BTNet) . . 80
  Bootstrapped Path Shaking (BPS) . 76, 85, 86
  Deep Random Forests (DRF) . . . 8, 78
  Out-of-bag (OOB) 80, 89
  Random survival forest (RSF) 87
BYOL 37

C

Cancer 13
Clinical workflow 17, 27
Colorectal cancer 7, 15, 16, 42, 47, 51, 52, 88, 90, 117
  Metastatic (mCRC), 16, 29
Lesion progression 29
Lung cancer
  Non-small-cell (NSCLC), 16
  Small-cell (SCLC), 16
Mean sojourn time 52
Primarius 16, 17, 29
Pseudoprogression 39
RECIST 29
Tumor genotype 39
Tumor phenotype . . 39, 46, 53, 59, 94
Case preparation 18
Catastrophic forgetting 99
Clinical decision support 113
Clinician-in-the-loop 117
Comprehensibility 113, 117
Confidence estimation . . . 5, 7, 61, 73, 114
  Monte-Carlo dropout 62
  One-shot 63
  Stochastic batch normalization 62
  Variational inference 62
Continious learning 118
Cross validation 34, 37, 42, 46, 69
  Randomized search 34, 41
Curriculum learning 7, 62, 71

D

Data augmentation 31, 86
Data efficiency 115
Data sparsity 114
Datasets
BreastMNIST . . 82, 83, 102, 103, 105, 108
ChestMNIST 82
CIFAR-10 . . . . . 67, 82, 102
CIFAR-100 . . . . . 67, 82
GBSG2 88
LIDC-IDRI 82, 88, 102, 104
MedMNIST 82, 86, 102
NSCLC 88
PneumoniaMNIST 82
Rossi Criminal Recidivism 52, 53, 88
SEER 51, 55, 88, 90
Date of first diagnosis (DOFD) 17, 48
Decision explanation 5, 93, 115
Additive feature attribution 97
CAM 97
Decision relevance 99
DeconvNet 97
DeepLIFT 97
DeepSHAP 97
DeepTaylor 97
GradCAM 97
LRP 97
Model-agnostic 97
Model-specific 97
Quality criteria 104
Semi-synthetic 97
SHAP 97
Decision trees 78
Randomized (RDT) 78
Deep Neural Networks . 23, 27, 40, 61, 73, 117
DINO 97
Dispersion 114

E

Early assessment 30, 37, 113
Explainable AI (XAI) 24, 97

F

Feature selection 20, 45, 78, 87, 88
ANOVA 41
Federated learning 86, 118

H

Hand-crafted feature design . 4, 20, 21, 31, 48, 54, 78, 89
Healthcare data 4

I

Image-to-image translation
Unpaired 98
Imaging modalities
Computed tomography (CT) . . 13, 17, 27, 114
Magnetic resonance (MRI) . . . 13, 114
Ultrasound (US) 13, 82, 103
Innovative contributions 5, 7, 113
Inter-observer reliability 104
Inter-observer variability 27, 30
Interpolation
Bicubic 33, 40

L

Label imbalance 4, 35, 36, 42, 81, 82, 115
Label noise 4, 82, 85, 91, 113
Lesion growth prediction 7, 29, 31, 59, 66, 69, 70, 73
Losses
Binary crossentropy 65, 88, 101
Categorical crossentropy 34, 62
Focal loss 62, 81
Hinge loss 62
Structural dissimilarity 100, 102

M

Machine learning . 3, 9, 20, 34, 41, 47, 74
Classical 21
Random Forest Classifier 41
Measures
Mean absolute error 52, 65, 89

Mean squared error 47, 65
RECIST diameter 29, 44, 53
Volume 31, 40
Medical image analysis 113
Meta-architectures 59
Deep Metamemory 62
Metacognition 63
Metamemory 63
Metrics
AUAUC 89
AUC 36, 68, 103
Concordance index . . . 52, 55, 89, 160
F1 score 68
F1-score 36
Gini coefficient 70, 78
Informedness 36
Markedness 36
Matthews correlation . 36, 68, 83, 103
Mean absolute error 34
Mean squared error 67
Spearman's rank correlation 68
Mode collapse 99
Model architectures
Attention-gated networks 50, 97
Autoencoder 7, 31, 34, 41, 73
Sparse, 31, 34, 69, 88
ConvNet . . . 32, 50, 54, 68, 75, 81, 85, 101, 103
CycleGAN 98
EfficientNet 86, 103
GAN 98
GPT-2 xi
Graph-based 54
PatchGAN 101
ResNet 50, 53, 54, 72, 75, 81, 85, 101, 103, 153
U-Net 23, 100
Multi-class averaging
Micro 68
Multicollinearity 20, 45, 48

N

Neural architecture search 79

O

One in ten rule 23
One-year survival 5, 39, 40, 42, 43, 47, 59, 114
Optimizers
Adam 42, 67
NAdam 34, 42
Stochastic gradient descent 95
Out-of-distribution (OOD) 62, 117

P

PANTHER 33, 69, 114
PAWS 37, 97
Pooling 75
Global maximum pooling 81
Maximum pooling 68, 75
Precision medicine 38, 46
Principle component analysis 104

R

Radiomics . . 19, 30, 34, 40, 43, 44, 78, 88, 113
Regression 13, 47, 48
Regularization
Shake-Shake 75
ShakeDrop 75
Resampling
Isotropic 33, 40
Risk stratification 48, 114

S

Small data 4, 18, 20, 23, 47, 93, 99
Spatial distribution 114
Survival estimation 47, 48, 87, 114
Cox proportional hazards 47
Cumulative hazard function (CHF) 87, 91
Right-censoring 50
Survival regression 5, 47, 48, 59

T

Therapy adaption 37, 114

Transfer learning 86

U

Uncertainty quantification 61

Z

Z-normalization 41, 53, 83, 104