Inaugural – Dissertation

zur

Erlangung der Doktorwürde

der

Gesamtfakultät für Mathematik, Ingenieur- und Naturwissenschaften

der

Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Tim Julian Adler, M. Sc.

aus Sinsheim

Tag der mündlichen Prüfung: ______________________________

Uncertainty Quantification in Biophotonic Imaging using Invertible Neural Networks

Supervisor: Prof. Dr. Lena Maier-Hein

„[…] Ich find das immer so klugscheißermäßig mit vorangestellten Zitaten. Jeder Depp meint ja, wenn er vornweg Oscar Wilde, Brecht oder Kafka zitiert, dann wird's gleich Literatur, was er da verbricht."

Das Känguru

aus

Marc-Uwe Klings
„Die Känguru-Chroniken"

Abstract

Owing to high stakes in the field of healthcare, medical machine learning (ML) applications have to adhere to strict safety standards. In particular, their performance needs to be robust toward volatile clinical inputs. The aim of the work presented in this thesis was to develop a framework for uncertainty handling in medical ML applications as a way to increase their robustness and trustworthiness. In particular, it addresses three root causes for lack of robustness that can be deemed central to the successful clinical translation of ML methods:

First, many tasks in medical imaging can be phrased in the language of inverse problems. Most common ML methods aimed at solving such inverse problems implicitly assume that they are well-posed, especially that the problem has a unique solution. However, the solution might be ambiguous. In this thesis, we introduce a data-driven method for analyzing the well-posedness of inverse problems. In addition, we propose a framework to validate the suggested method in a problem-aware manner.

Second, simulation is an important tool for the development of medical ML systems due to small *in vivo* data sets and/or a lack of annotated references (e. g. spatially resolved blood oxygenation (sO_2)). However, simulation introduces a new uncertainty to the ML pipeline as ML performance guarantees generally rely on the testing data being sufficiently similar to the training data. This thesis addresses the uncertainty by quantifying the domain gap between training and testing data via an out-of-distribution (OoD) detection approach.

Third, we introduce a new paradigm for medical ML based on personalized models. In a data-scarce regime with high inter-patient variability, classical ML models cannot be assumed to generalize well to new patients. To overcome this problem, we propose to train ML models on a per-patient basis. This approach circumvents the inter-patient variability, but it requires training without a supervision signal. We address this issue via OoD detection, where the current status quo is encoded as in-distribution (ID) using a personalized ML model. Changes to the status quo are then detected as OoD.

While these three facets might seem distinct, the suggested framework provides a unified view of them. The enabling technology is the so-called invertible neural network (INN), which can be used as a flexible and expressive (conditional) density estimator. In this way, they can encode solutions to inverse problems as a probability distribution as well as tackle OoD detection tasks via density-based scores, like the widely applicable information criterion (WAIC).

The present work validates our framework on the example of biophotonic imaging. Biophotonic imaging promises the estimation of tissue parameters such as sO_2 in a non-invasive way by evaluating the “fingerprint” of the tissue in the light spectrum. We apply our framework to analyze the well-posedness of the tissue parameter estimation problem at varying spectral and spatial resolutions. We find that with sufficient spectral and/or spatial context, the sO_2 estimation problem is well-posed. Furthermore, we examine the realism of simulated biophotonic data using the proposed OoD approach to gauge the generalization capabilities of our ML models to *in vivo* data. Our analysis shows a considerable remaining domain gap between the *in silico* and *in vivo* spectra. Lastly, we validate the personalized ML approach on the example of non-invasive ischemia monitoring in minimally invasive kidney surgery, for which we developed the first-in-human laparoscopic multispectral imaging system. In our study, we find a strong OoD signal between perfused and ischemic kidney spectra. Furthermore, the proposed approach is video-rate capable.

In conclusion, we successfully developed a framework for uncertainty handling in medical ML and validated it using a diverse set of medical ML tasks, highlighting the flexibility and potential impact of our approach. The framework opens the door to robust solutions to applications like (recording) device design, quality control for simulation pipelines, and personalized video-rate tissue parameter monitoring. In this way, this thesis facilitates the development of the next generation of trustworthy ML systems in medicine.

Zusammenfassung

Aufgrund des hohen Risikos im Gesundheitswesen müssen medizinische Machine Learning (ML)-Anwendungen strenge Sicherheitsstandards einhalten. Insbesondere muss ihre Leistung gegenüber volatilen klinischen Eingaben robust sein. Das Ziel der in dieser Dissertation vorgestellten Arbeit war die Entwicklung eines Frameworks für den Umgang mit Unsicherheiten in medizinischen ML-Anwendungen. Solch ein Framework kann dazu beitragen die Robustheit und Vertrauenswürdigkeit medizinischer ML-Anwendungen zu erhöhen. Insbesondere werden drei Hauptursachen für mangelnde Robustheit adressiert, die als zentral für die erfolgreiche klinische Translation von ML-Methoden angesehen werden können:

Erstens lassen sich viele Aufgaben in der medizinischen Bildgebung als inverse Probleme formulieren. Die meisten gebräuchlichen ML-Methoden, die darauf abzielen, solche inversen Probleme zu lösen, gehen implizit davon aus, dass sie gut gestellt sind, insbesondere, dass das Problem eine eindeutige Lösung besitzt. Es kann jedoch mehrere Lösungen geben. In dieser Arbeit stellen wir eine datengetriebene Methode zur Analyse der Wohlgestelltheit inverser Probleme vor. Darüber hinaus schlagen wir ein Framework zur Auswahl geeigneter Metriken vor, um die vorgeschlagene Methode problemspezifisch zu validieren.

Zweitens ist die Simulation aufgrund kleiner In-vivo-Datensätze und/oder fehlender annotierter Referenzen (z. B. ortsaufgelöster Sauerstoffsättigungen (sO_2)) ein wichtiges Werkzeug für die Entwicklung medizinischer ML-Systeme. Die Simulation führt jedoch eine neue Unsicherheit in die ML-Pipeline ein, da ML-Leistungsgarantien im Allgemeinen davon abhängen, dass die Testdaten den Trainingsdaten ausreichend ähnlich sind. Diese Dissertation befasst sich mit dieser Unsicherheit, indem sie den Domain Gap zwischen Trainings- und Testdaten über einen Out-of-Distribution (OoD)-Detektionsansatz quantifiziert.

Drittens führen wir ein neues Paradigma für medizinisches ML ein, das auf personalisierten Modellen basiert. In einem datenarmen Regime mit hoher Inter-Patienten-Variabilität kann nicht davon ausgegangen werden, dass klassische ML-Modelle gut auf neue Patienten generalisieren. Um dieses Problem zu lösen, schlagen wir vor, ML-Modelle auf jedem

Patienten individuell zu trainieren. Dieser Ansatz umgeht die Inter-Patienten-Variabilität, erfordert jedoch ein Training ohne Überwachungs-Signal. Wir lösen dieses Problem über die OoD-Detektion, bei der der aktuelle Status quo als in-Distribution (ID) unter Verwendung eines personalisierten ML-Modells codiert wird. Änderungen am Status quo werden dann als OoD erkannt.

Während diese drei Facetten unterschiedlich erscheinen mögen, bietet das vorgeschlagene Framework eine einheitliche Sicht auf sie. Die Voraussetzungen des Frameworks bilden sogenannte Invertierbare Neuronale Netze (INNs), die als flexible und mächtige (bedingte) Dichteschätzer verwendet werden können. Auf diese Weise können sie Lösungen für inverse Probleme als Wahrscheinlichkeitsverteilung codieren sowie OoD-Detektionssaufgaben über dichtebasierte Scores wie das Widely Applicable Information Criterion (WAIC) angehen.

Die vorliegende Arbeit validiert unser Framework am Beispiel der biophotonischen Bildgebung. Die biophotonische Bildgebung verspricht die Schätzung von Gewebeparametern wie sO_2 auf nicht-invasive Weise, indem der „Fingerabdruck“ des Gewebes im Lichtspektrum ausgewertet wird. Wir wenden unser Framework an, um die Wohlgestelltheit des Gewebeparameterschätzungsproblems bei unterschiedlichen spektralen und räumlichen Auflösungen zu analysieren. Unsere Untersuchung ergibt, dass bei ausreichendem spektralem und/oder räumlichem Kontext das sO_2-Schätzproblem gut gestellt ist. Darüber hinaus untersuchen wir die Realitätstreue simulierter biophotonischer Daten mit dem vorgeschlagenen OoD-Ansatz, um die Generalisierungsmöglichkeiten unserer ML-Modelle auf In-vivo-Daten abzuschätzen. Unsere Analyse zeigt einen beträchtlichen Domain Gap zwischen den In-silico- und In-vivo-Spektren. Schließlich validieren wir den personalisierten ML-Ansatz am Beispiel des nicht-invasiven Ischämie-Monitorings bei minimal-invasiven Niereneingriffen. Im Zuge der Studie entwickelten wir das erste laparoskopische multispektrale Bildgebungssystem, das im Menschen eingesetzt wurde. In unserer Studie finden wir ein starkes OoD-Signal zwischen perfundierten und ischämischen Nierenspektren. Darüber hinaus ist der vorgeschlagene Ansatz schnell genug für Video-Anwendungen.

Zusammenfassend haben wir erfolgreich ein Framework für den Umgang mit Unsicherheiten im medizinischen ML entwickelt und es anhand einer Vielzahl medizinischer ML-Aufgaben validiert, was die Flexibilität und den potenziellen Impact unseres Ansatzes zeigt. Das Framework öffnet die Tür zu robusten Lösungen für Anwendungen wie dem Design von Aufnahme-Geräten, der Qualitätskontrolle von Simulationspipelines und der personalisierten Überwachung von Gewebeparametern mit Videorate. Auf diese Weise ermöglicht diese Arbeit die Entwicklung der nächsten Generation vertrauenswürdiger ML-Systeme in der Medizin.

Acknowledgments

More than 200 pages, 53 figures, 57 footnotes[1], over 400 commits to the thesis's LaTeX project alone, roughly 1000 commits spread over multiple git repositories, more than four years of work, and a global pandemic to boot — My time as a Ph. D. student has been quite the ride. I could not have succeeded by myself. Now, I would like to take the time to say thank you.

First, I would like to thank my supervisor Lena Maier-Hein. Sometimes, I had the feeling that you believed more in my work than I did. However, by doing so, you gave me the courage to publish and present my results. I very much appreciated your enthusiasm for discussing my research projects, challenges, and results. I felt that our discussions were always on eye level, and I could make myself heard. I do not think that any of this is self-evident, and I am very grateful that you built this environment.

My next shout-out goes to Leonardo Ayala. No matter my spectral imaging question, you always knew the answer, and you never became grumpy, even if I disturbed you ten times. I will always remember our drives to Karlsruhe and our discussions along the way[2]. Working with you on a joint paper was a blast, and having you as my group leader in spectral imaging was a great experience.

I am also very grateful for the opportunity to supervise students. Jan-Hinrich Nölke was one such student who wrote his master thesis with me and decided to stay for his Ph. D. studies. I always looked forward to our weekly discussions, and it was a joy to support you in your work and get feedback on my own problems and struggles. Thank you for your independent work and your clear presentations. I would also like to thank my interns Nina Sautter[3] and Elise Récéjac. It was a pleasure working with you, and your progress during your time in our department was astounding.

The next one up is my favorite external collaborator, Lynton Ardizzone. My INNs behaved strangely, and one e-mail later, I received the explanation, including the solution. Your knowledge about the architecture in particular and deep learning (DL) in general is inspiring. At the same time, you have always been down-to-earth and unassuming, which made me feel right at home. Being able to work on our "approval requirements" for the Ph. D. program together made the task very much enjoyable. My gratitude also goes out to the complete Explainable Machine Learning group at Heidelberg University, particularly to Ullrich Köthe, who always found time to discuss my current research projects and has an amazing overview of the DL literature.

[1] I love footnotes.

[2] Or while waiting because of a delayed surgery...

[3] Who stayed on as a research assistant.

This is the end of my Ph. D. journey, but there is no end without a beginning[4]. Thomas Kirchner and Janek Gröhl made sure that my beginning was smooth. I am very thankful that you took me under your wings and showed me the ropes. You made me feel very welcome, and I hope you are not too disappointed that my focus shifted from photoacoustics to spectral imaging.

Regarding the end of my journey, I wanted to express my gratitude to Annika Reinke and Leonardo Ayala, who, together with me, founded the "IMSY Thesis Writing Club". Our regular calls and exchanges about our respective thesis progress were very therapeutic. I feel very indebted to you for your willingness to endure my venting ☺. In addition, I wanted to thank all the hard-working proofreaders who heavily improved my thesis's readability and understandability. My thanks go out to[5] Leonardo Ayala, Christoph Bender, Patrick Godau, Marco Hübner, Keno März, Jan-Hinrich Nölke, Maike Rees, Alexander Seitel, Silvia Seidlitz, Jan Sellner, Minu Dietlinde Tizabi, and Fabian Wolf.

As a member and later the leader of the Intelligent Systems in Endoscopy subgroup, I am very happy to have met and worked with a group of such motivated and supportive people. You have made my job very easy, and I love our team spirit. Thank you Sebastian Gruber, Lucas-Raphael Müller, Annika Reinke, Henri Smidt, Akriti Srivastava, Thuy Nuong Tran, Dasha Trofimova, and Amine Yamlahi. I especially wanted to thank two people. First, I wanted to thank my predecessor, Tobias Roß, who established the reporting and other safeguards to ensure that no one was left behind. Second, I wanted to thank my successor, Patrick Godau, who made sure that I felt safe in handing over the responsibilities and could fully focus on my thesis.

As a member of a second subgroup, Intelligent Systems in Spectral Imaging, I could repeat the praise from the previous paragraph. Suffice it to say that you are great team players, and each and everyone is simultaneously very knowledgeable and very helpful. I am happy that I got the chance to work with you. Thank you Leonardo Ayala, Marco Hübner, Maike Rees, Silvia Seidlitz, Jan Sellner, and Ahmad Bin Qasim.

Let me turn toward the unsung heroes. There is no bureaucratic problem they cannot solve. They are at the same time competent and kind, which I find an amazing combination. Thank you for all your help and your infinite nerve regarding administrative mistakes: Janina Dunning, Michaela Gelz, Theresa Klocke, and Stefanie Strzysch. I would imagine a Ph. D. in computer science without the technical infrastructure to be very hard. Hence, I would like to thank Marco Pascale, Stefan Dinkelacker, Melinda Mike, and the IT core facility for keeping everything running.

Naturally, IMSY consists of more people than I could name in this thesis, but they all contribute to our great work environment. Thank you so much for being the awesome people that you are! If you have the feeling that I left you out unfairly, I hope you can forgive me.

[4] Philosophers are advised not to think too much about this sentence.

[5] All lists are ordered alphabetically.

Last but by no means least, I would like to thank my family. First, I would like to thank my parents, Susanne and Jochen Adler, who have supported me in all my endeavors and sparked and nurtured my interest in the sciences. You gave me the safety that I needed and enabled me to find my own path. More tangibly, I am grateful that you gave me refuge while I was fully immersed in the writing process. Not needing to cook was incredible 😀. Finally, I would like to express my gratitude toward my wife, Sabrina Baier, who was there for me through the highs and lows of my journey. You endured my bad tempers when I felt uncertain and insecure. You brainstormed strategies and solutions with me and showed me a path forward. This thesis is to no small part to your credit. I am deeply thankful that you helped me throughout this adventure.

Contents

Lists

List of Acronyms

Machine Learning

ABC	approximate Bayesian computation
AI	artificial intelligence
AP	average precision
AUC	area under curve
AUROC	area under receiver operating characteristics curve
bw	bandwidth
cINN	conditional invertible neural network
CI	confidence interval
CNN	convolutional neural network
conv-cINN	convolutional cINN
DBSCAN	density-based spatial clustering of applications with noise
DCF-CSR	discriminative correlation filter with channel and spatial reliability
DL	deep learning
ECE	expected calibration error
EM	expectation-maximization
fc-cINN	fully-connected cINN
FN	false negative
FP	false positive
FrEIA	framework for easily invertible architectures
GAN	generative adversarial network
GMM	Gaussian mixture model
HSM	half-sample mode
ID	in-distribution
IQR	interquartile range
INN	invertible neural network
KDE	kernel density estimation
KL	Kullback-Leibler
k**-NN**	k-nearest neighbor
KS	Kolmogorov-Smirnov
MedAE	median absolute error
MAP	maximum a posteriori
ML	machine learning
MLP	multi-layer perceptron

MMD	maximum mean discrepancy
NN	neural network
OoD	out-of-distribution
PCA	principal component analysis
ReLU	rectified linear unit
RNN	recurrent neural network
SVM	support vector machine
TN	true negative
TP	true positive
VAE	variational auto-encoder
WAIC	widely applicable information criterion

Biophotonics

DAS	delay-and-sum
Hb	(deoxy-)hemoglobin
HbO_2	oxyhemoglobin
HSI	hyperspectral imaging
ICG	indocyanine green
MC	Monte Carlo
MCML	Monte Carlo Multi-Layered
MCX	Monte Carlo eXtreme
MSI	multispectral imaging
PAI	photoacoustic imaging
PDE	partial differential equation
ROI	region of interest
RTE	radiative transfer equation
SI	spectral imaging
SIMPA	Simulation and Image Processing for Photonics and Acoustics
sO_2	blood oxygenation
US	ultrasound
vHb	blood volume fraction

Miscellaneous

AI-HERO	Hackathon on Energy Efficient AI
BVM	Bildverarbeitung für die Medizin Workshop
CT	computed tomography
COVID-19	coronavirus disease 2019
DKFZ	German Cancer Research Center
EndoCV	Computer Vision in Endoscopy
IJCARS	International Journal of Computer Assisted Radiology and Surgery
IPCAI	International Conference on Information Processing in Computer-Assisted Interventions
MICCAI	International Conference on Medical Image Computing and Computer Assisted Intervention
MIDL	Medical Imaging with Deep Learning Conference

MOOD	Medical Out-of-Distribution Analysis Challenge
MRI	magnetic resonance imaging
NeurIPS	Conference on Neural Information Processing Systems
OR	operating room
UNSURE	Uncertainty for Safe Utilization of Machine Learning in Medical Imaging

List of Figures

List of Tables

Part I.

Introduction

1. Motivation

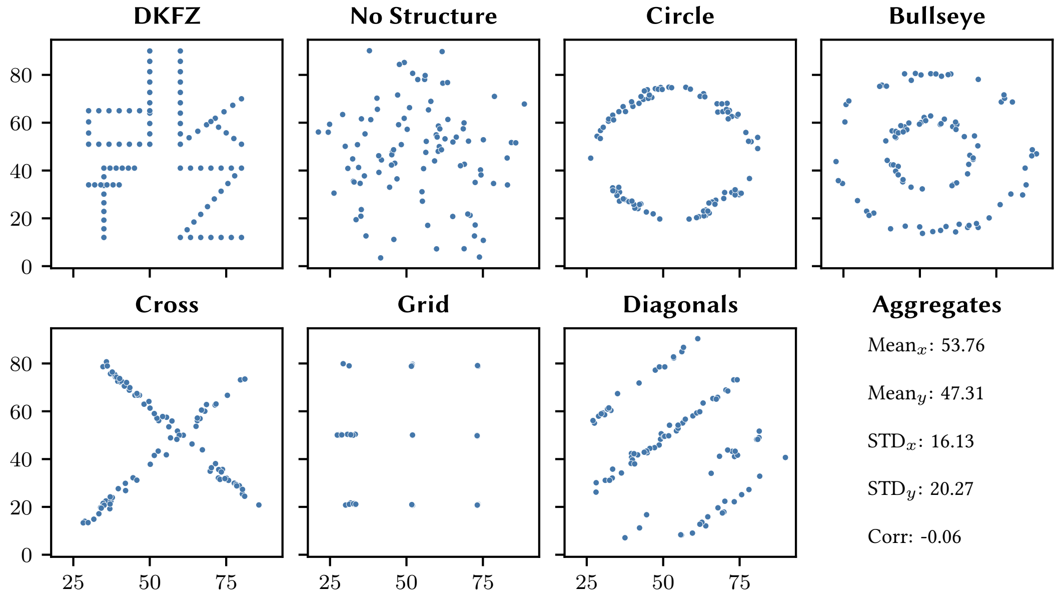


Figure 1.1.: **Summary statistics can hide structure in data.** The seven 2D data sets exhibit a wide range of patterns with identical mean, standard deviation, and correlation coefficient up to the second decimal. The data sets were inspired by and constructed using code from [MF17]. STD: standard deviation, DKFZ: German Cancer Research Center.

Artificial intelligence (AI) and machine learning (ML) in particular have led to leaps in many computer science and real-world problems (e. g. [Goo+14; RFB15; Sil+16; Sil+17; Dev+18; Sil+18; Bro+20; Sen+20; Jum+21; Ram+22]). Therefore, it comes as no surprise that ML has found its way into the medical domain. However, off-the-shelf methods oftentimes do not meet the requirements due to the high-risk nature of medical care. Errors of a recommender system might be acceptable, but an undetected polyp could have far-reaching consequences. Hence, uncertainty quantification plays a central role in most medical ML applications. Colloquially, uncertainty quantification is concerned with teaching ML models to know what they do not know. The first approaches boiled down to interpreting the actual prediction as a mean of some distribution and augmenting it with a standard

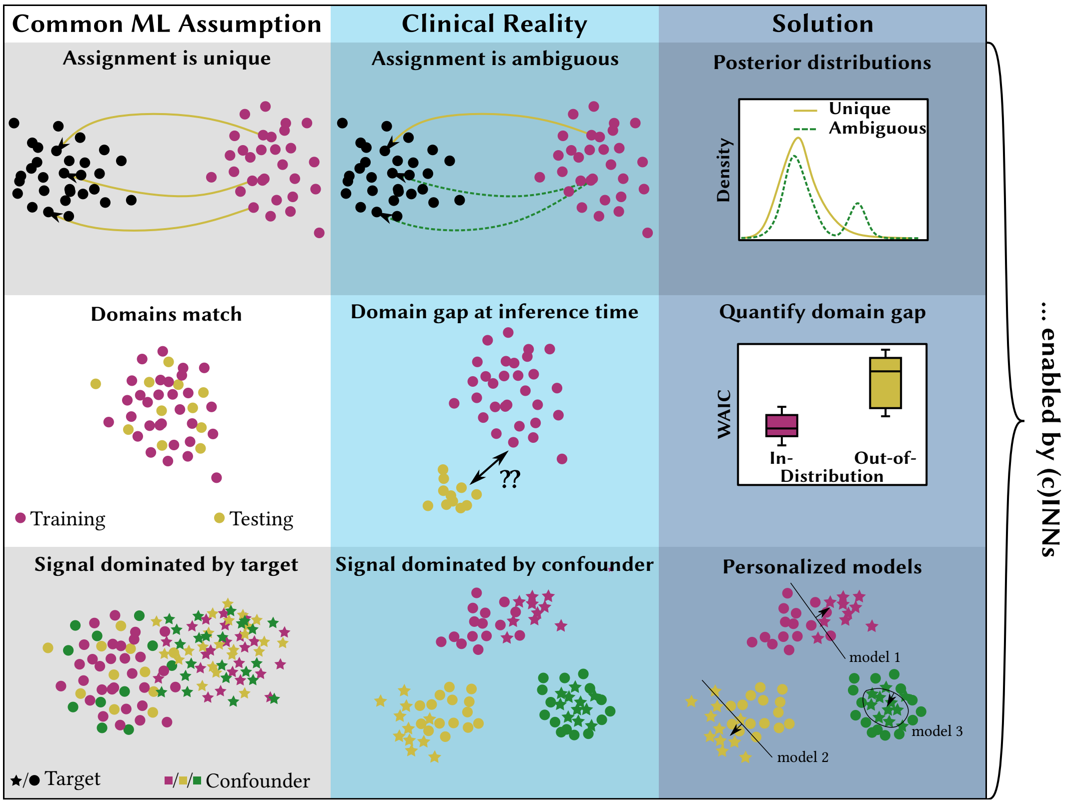


Figure 1.2.: **ML assumption violations in the clinic and how to solve them with INNs. First row:** Inverse problems encountered in a clinical setting are often ill-posed, i. e. the solution to the problem might be ambiguous. We propose cINNs to encode this ambiguity in multimodal posterior distributions. **Second row:** Due to data sparsity, training data is often generated via simulation. However, there might be a domain gap between the training and the testing domain. We propose INNs to detect and quantify this domain gap as an out-of-distribution (OoD) detection task. **Third row:** Clinical data sets are often small, and the variability between patients (confounders) can cover the physiological signal of interest (target). We propose INNs as unsupervised OoD detectors to train personalized models circumventing the confounders. INN: invertible neural network, cINN: conditional INN, WAIC: widely applicable information criterion, ML: machine learning.

deviation (e. g. [KG17]). The standard deviation becomes the uncertainty score. Figure 1.1 gives an idea of why this might be insufficient. As introduced in the work of Matejka and Fitzmaurice [MF17], we see that summary statistics can hide a plethora of structure in the data. This observation motivates our quest to access the underlying distribution.

An overview of our contributions can be found in figure 1.2. Each row contrasts a common ML assumption with the clinical reality. It is of note that all addressed examples are concerned with representing or coping with uncertainty. The last column hints at the proposed solutions, which are methodologically powered by so-called invertible neural networks (INNs) [Ard+18b]. In fact, this thesis's core is built around a framework with INNs at the center to wholistically handle the uncertainty in medical imaging applications. INNs are a special neural network (NN) architecture that builds a family of very flexible density estimators. This property is useful as many questions in uncertainty quantification can be formulated in a probabilistic manner.

We will explore the proposed framework with the example of biophotonic imaging. Biophotonic imaging is a new and fast-paced research area with great potential for patient care. When light interacts with tissue, the resulting spectrum contains information about the tissue composition on a molecular level. The value proposition of biophotonic imaging is the recovery of the physiological tissue composition from the spectra (cf. figure 1.3, left column). This approach is non-invasive, and if successful, it can pave the way to new clinical applications like identifying the target for tissue ablation in cancer or cardiac arrhythmia, temperature monitoring during the ablation, and tissue classification for polyp detection (cf. figure 1.3, right column). The first success stories in biophotonic imaging include monitoring hemodynamic changes in the porcine brain [Aya+19; Kir+19] and the detection of sepsis [Die+21]. However, while these results are promising, we need to ensure that the ML models are robust before proceeding with the clinical translation. We will discuss this robustness along the common ML assumptions and how they are violated in a clinical setting introduced in figure 1.2.

The first assumption (cf. figure 1.2, first row) concerns the well-posedness of inverse problems. This might sound rather abstract, but the regression of tissue parameters from spectra in biophotonic imaging is an example of solving an inverse problem. Most classic ML algorithms implicitly or explicitly assume that the inverse problem is well-posed. One aspect of this is that a solution exists and is unique. If this assumption is violated, the consequences in an interventional setting might be drastic. Consider the example of perfusion monitoring. If the biophotonic inverse problem is ill-posed, then there might be two tissue parameter configurations, one with a high blood oxygenation (sO_2) and one with a low sO_2, with nigh identical associated spectra. A classic ML method might collapse to one of the two tissue parameter configurations. In the worst case, the model could predict that an organ is still perfused, while it might actually have been severed from the blood supply for some time. We will address this problem using cINNs [Ard+19]. Instead of predicting a single tissue configuration per spectrum (point estimate), they provide a full probability distribution conditioned on the spectral input, a so-called *posterior*. Such

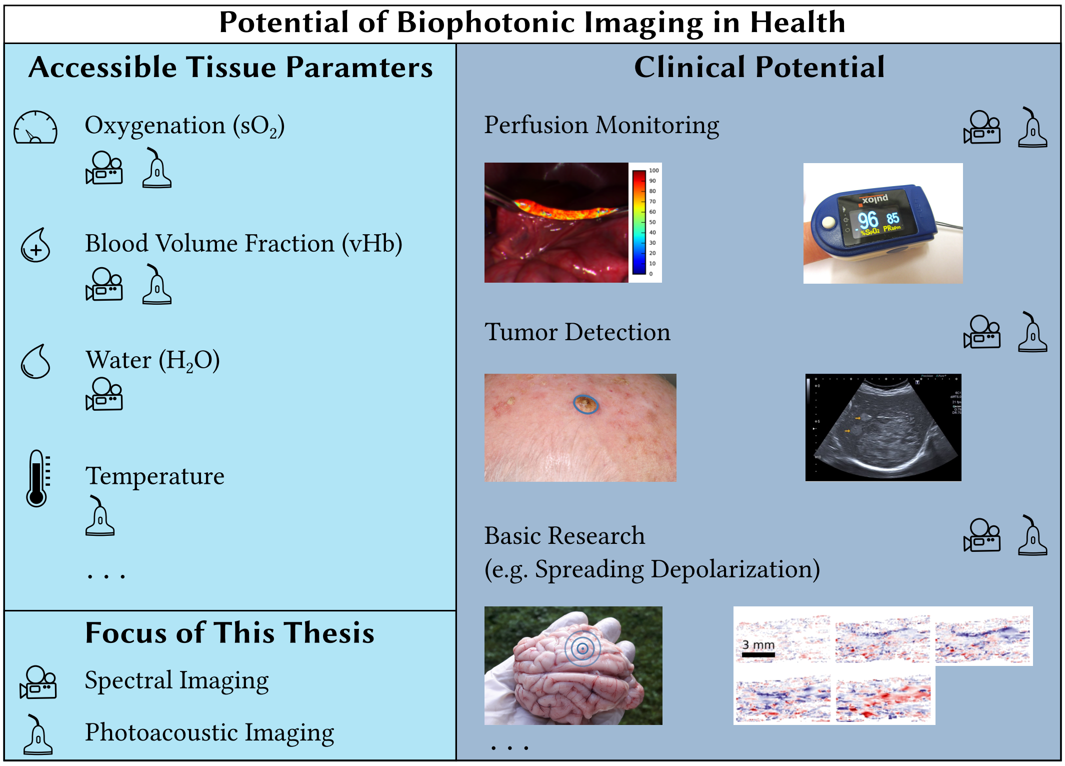


Figure 1.3.: **Potential of biophotonic imaging in health. Left Column:** Example physiological tissue parameters that can potentially be reconstructed using biophotonic imaging. **Right Column:** Example applications of biophotonic imaging in health. The example images were taken and adapted from [Bel14; Wir+17; Kir+19; Wai19; Wer20; Tie21]. The icons indicate if the tissue parameter or clinical application is accessible via spectral imaging or photoacoustic imaging, which are examples of biophotonic imaging modalities.

a posterior can encode ambiguous solutions as multiple modes, making the ill-posedness visible and actionable by the practitioners.

The second assumption (cf. figure 1.2, second row) concerns a domain gap between the training and testing domain. The goal of biophotonic imaging is to predict spatially resolved tissue parameters, but to date, there is no gold standard reference method that could be used to collect training data for the ML models. However, in order to train a supervised regression model, we need training samples consisting of spectra with associated tissue parameters. The community-accepted solution to this conundrum is simulation. Using a "digital tissue twin", we can simulate the light-tissue interaction (mostly using Monte Carlo (MC) methods) to generate the associated spectrum. However, human tissue is complex, and certain modeling assumptions are necessary for a feasible simulation implementation. These assumptions might come at the cost of the realism of the simulation. The problem is that most ML methods are only guaranteed to perform on testing data that is sufficiently similar to the training data. Even worse, oftentimes ML models fail very confidently on so-called OoD data [Big+13; GSS14; Bro+17; CW17; Mad+17]. The implications for medical applications are that if the domain gap between the (simulated) training data and the (*in vivo*) testing data is too large, the ML model might predict high sO_2 levels, and even with high confidence[1], even though the sO_2 is once again very low. There are multiple facets to solving this problem. One is hardening models against domain gaps. Another is working toward diminishing the domain gap. In this thesis, we will take one step back and aim to quantify the domain gap. As it turns out, it is easy to say that the testing data needs to be "sufficiently similar", but to put this into equations and numbers is not as simple. While not solving the drop in model performance, a robust domain gap quantification would at least allow to filter OoD data and prevent spurious predictions of the main model. We will use ensembles consisting of INNs to represent the density of the training or in-distribution (ID) data. We can use the ensembles at testing time to estimate the likelihood of new data. We will enrich this information using the epistemic uncertainty of the ensemble members, which leads to the widely applicable information criterion (WAIC). WAIC can then be used as an OoD score, which we propose to use to quantify domain gaps. This allows us to examine our simulation pipeline with regard to its realism, and WAIC can be used as an indicator of the generalization capabilities for models trained on *in silico* data when transferred to *in vivo* data.

The third assumption (cf. figure 1.2, third row) is directly related to the previous OoD discussion. While the OoD approach allows us to detect if our simulation pipeline is insufficient to cover *in vivo* data, the approach does not provide a constructive way to adapt the simulation pipeline. Hence, we might be in a setting where we need to train on *in vivo* data. As mentioned previously, with a lack of a gold standard reference method to collect spatially resolved tissue parameters, we are already restricted in the types of models we can train, i. e. regressing tissue parameters will be hard, but a coarse classification, e. g. into

[1] If the model is capable of producing such a score, like the cINN.

perfused and ischemic tissue, might be possible. However, medical data sets are notoriously small. This is understandable because of the sensitivity of the data. The downside is that often the patient cohorts are very small, and if the data exhibits high inter-patient variability, we cannot hope to train a classifier that would generalize to new, unseen patients because the data would be OoD with regard to the training patients. The direct solution to this problem is to collect larger data sets such that the variability is better covered. This thesis will pursue a complementary approach. Instead of representing the inter-patient variability in the data, we will circumvent it with a personalized ML model, i. e. if we train a model on each patient individually, we can ignore the inter-patient variability. If successful, such an approach would be very interesting because it would allow us to ignore one of the main confounders (the patient) in medical applications. However, such an approach requires some care. Let us return to the perfused/ischemic classification example. A regular classifier would require a training data set containing perfused and ischemic spectra. The spectra would need to be labeled with a reference model. However, if we can classify spectra by an alternative means (the reference), what is the point of the personalized model? Hence, we rephrase the problem as another OoD detection problem. For example, at the beginning of a surgical intervention, we can generally be sure that the organ in question is perfused. We can collect spectra and define them as ID data. In other words, we train an ML model to recognize the current tissue status quo. Later in the surgery, we can then detect changes to the status quo as OoD data. In this way, we can detect ischemic spectra as a deviation from the ID perfused spectra. As for the domain gap quantification, we propose INN ensembles and WAIC as our OoD detectors of choice. The INNs, as flexible density estimators, are trained to represent the ID data, and the likelihood, together with the epistemic uncertainty of the ensemble members, will be aggregated in WAIC and used as an OoD score on unseen data. INNs can be adapted to train fast enough (less than 1 min) such that training at the beginning of the actual procedure becomes realistic. In addition, the inference time is fast enough to enable video rate perfusion monitoring after training. This personalized approach is a completely new paradigm for ML in interventional health care.

ML breakthroughs in medicine are still missing, but we are confident that the proposed framework for uncertainty quantification and INNs, with their versatility, can bring those breakthroughs one step closer to reality.

2. Research Questions

This chapter introduces the research questions that this thesis aims to answer. As ML for medical applications is a highly interdisciplinary area, we have divided the research questions into a technical part (T.1 – T.3) and a domain part (D.1 – D.3). The technical research questions are related to methodological challenges that are broadly applicable beyond the scope of medical imaging. The domain research questions concern concrete domain-related problems that we need to address to bring ML methods to the patient.

T.1. How can we encode inherent ambiguities in inverse problems?

As introduced in the previous chapter, most classic ML models assume that inverse problems have unique solutions (cf. figure 1.2, first row). However, we would like to avoid this a priori assumption and empirically analyze the well-posedness of biophotonic inverse problems. To this end, we require a technique to encode ambiguous solutions. This is the content of research question T.1.

When choosing our method to analyze the biophotonic imaging setting, the main consideration is its diversity. There is not a single, static inverse problem, but the exact problem depends on the chosen recording device, which influences the amount of spectral information, and on the data in question. If we restrict our attention to porcine spectra, the inverse problem might be unique, but including human spectra might lead to ambiguities. In addition, due to the speed requirements for e. g. video rate applications, we do not necessarily want to operate on whole images all of the time. This introduces another source of possible ambiguities because of the reduced amount of spatial context. Hence, we require a flexible method that can be easily adapted and applied to this changing setting and provide a unified view of the well-posedness of the different instances of the inverse problem. To the best of our knowledge, there is no prior work systematically providing such a framework.

In this thesis, we propose cINNs to tackle this question. They encode the solution of an inverse problem in a conditional probability distribution (the posterior). If the posterior consists of more than one mode, this corresponds to an ambiguous solution. As an ML model, they are perfectly suited for our volatile setting, as we can train them on the appropriate data set for the chosen recording device, spectral resolution, spatial resolution, and other factors like the species. The posteriors open the door for a fine-grained analysis of uncertainties for the inverse problem. Next to the number of modes, we have access to measures of variability like the variance or the interquartile range (IQR). Hence, even for unimodal posteriors, we can use the width of the posterior to gauge the certainty of the cINN in its predictions and determine to what extent the inverse problem is solvable at all.

The posteriors of the cINN are accessible as a sample. Hence, this thesis will also propose suitable post-processing steps to extract information from this representation. In particular, we will discuss mode detection algorithms as the modes are not automatically labeled in the sample. We will resort to unsupervised clustering algorithms taking into account that we generally do not wish to specify the number of clusters in advance as the detection and analysis of ambiguities is our main goal. Furthermore, the cINN can produce artifacts in the posteriors, influencing our choice of mode detection algorithms.

We will approach research question T.1 using the example of biophotonic imaging, but the methods should generalize to other inverse problems.

T.2. How can we validate posteriors?

While T.1 is concerned with the construction of posterior distributions, T.2 addresses the question of how we can trust the posteriors. Before we base patient decisions on detected multimodalities in a posterior, we better be certain that the posterior adequately describes a solution to the inverse problem and that the modes are no artifacts. This validation of the posteriors is by no means trivial.

Let us assume that we have a testing data set. In biophotonic imaging, such a data set will generally consist of pairs of a spectrum and the corresponding tissue parameter configuration. The proposed cINN approach will generate a complete posterior for a given spectrum. Let us consider the case that the posterior is bimodal. Then we are faced with the problem that we have exactly one reference tissue parameter configuration. In the best-case scenario, this single reference coincides with one of the modes, but even in this scenario, we have a second mode without an associated reference. Is the second mode a false positive (FP), i. e. falsely generated by the cINN? Or is the model correct, and the fault lies with the data set containing only a single reference per spectrum?

These and similar questions need to be answered to validate the posteriors. The extent to which this is possible depends on many factors, like the exact representation of the reference data or the dimensionality of the problem. To the best of our knowledge, this problem has only been addressed in a very ad hoc manner in the sense that a sensible set of metrics was chosen for the inverse problem setting at hand, but leaving out the larger structure and patterns of the setting that might inform a systematic, consistent, and transferable choice of metrics. Such a framework could potentially lead to a community-accepted standard for the validation of posteriors.

This thesis will approach research question T.2 by proposing a posterior validation framework that attempts to cover a wide range of inverse problem settings and guide the user to a set of metrics that can be used to validate posteriors for their respective application. To this end, we will explore an analogy between the computer vision task of object detection and posterior validation for inverse problems. This analogy will allow us to transfer metrics from object detection, which is a more mature field, to our setting. At the same time, we will extend the object detection metrics by taking cINN peculiarities, like the necessity to calibrate the posterior width, into account.

T.3. How can we detect out-of-distribution data in biophotonic imaging?

There is a hoard of empirical evidence that ML models are only guaranteed to perform on data that is sufficiently close to the training data [Big+13; GSS14; CW17; Mad+17]. This is an important restriction as it is often impossible to create labeled *in vivo* data in biophotonic imaging and many other fields. Hence, the community uses simulated data to address data scarcity. However, simulation comes at the cost of added uncertainty. We have to ask ourselves whether the simulation framework captures enough of the reality to allow for robust ML models. What exactly "enough" means is hard to quantify, and this problem is the content of research question T.3. We propose to analyze the problem through the lens of OoD detection. In the OoD setting, there is ID data at training time, and we aim to find a model that can detect any data that is not drawn from the same distribution[1].

The detection of OoD data is common, and so many fields have developed their own approaches. For low-dimensional data, we could hope to get away with quantile-based methods. The computer vision community has developed approaches based on auto-encoders and representation learning. The field of medical spectral imaging comes with its own challenges and peculiarities. For example, we commonly operate on single pixels or small regions of interest (ROIs), such that image-based approaches are hard to transfer. At the same time, due to the higher spectral resolution, we are not in a low-dimensional setting, even for single-pixel data. Thus, we have to check the fit of OoD methodologies and adapt them to our special case. To the best of our knowledge, there is no prior work on OoD detection methods tailored towards spectral imaging (SI). To approach research question T.3, we will explore the potential of INNs as flexible density estimators and propose INN ensembles with WAIC as robust OoD detectors on SI data.

D.1. Are inverse problems in biophotonic imaging well-posed?

Biophotonic imaging has the potential to recover physiological tissue parameters during a medical intervention using only non-ionizing radiation. Still, the clinical impact of biophotonic imaging hinges on research question D.1. An answer in the affirmative would

[1] Hence, OoD data.

highly increase the trust that we can put into (classical) ML methods. However, even if we have to answer in the negative, a thorough analysis and the capability to reliably detect ambiguous situations can lead to an added value for the practitioner. If there are cases where the same spectrum corresponds to multiple markedly different tissue parameter configurations, we could warn the user about the situation. In a next step, this could even lead to automated suggestions on how the ambiguity might be resolved, e. g. by proposing additional recording poses.

As highlighted in research question T.1, we need a flexible, data-driven method as we want to answer the well-posedness question empirically for the different recording devices and data sets. Furthermore, we want to be able to predict ambiguous solutions for each data point (e. g. spectrum) individually instead of being restricted to results about the general well-posedness of the problem. This property would allow us to incorporate our method into medical devices and use it at the bedside. The last requirement due to our aim for use in interventional care is the video rate capability of the method.

To the best of our knowledge, the well-posedness in biophotonic imaging has only been analyzed using analytical methods based on partial differential equations (PDEs) [Arr99; RNR13; RT19]. There is previous work regarding uncertainty quantification in biophotonic imaging [Grö+18], but this work used ML methods that produced a point estimate and an (un-)certainty score, where the score was used to gauge the reliability of the point estimate. While this is a valuable contribution to uncertainty quantification in general, this approach cannot be used to detect ambiguities as it still produces a single prediction per input.

We propose the use of cINNs to approach research question D.1. We will use them to explore how the amount of spectral context (different recording devices) and spatial context (single pixels up to whole images) influences the well-posedness of the associated inverse problem.

D.2. Are synthetic spectral images sufficiently realistic?

As the previous research question mentioned, biophotonic imaging opens the door for non-invasive, interventional, functional imaging. The most prominent example is the estimation of sO_2 from spectral data. However, a big challenge in training and evaluating ML models is a lack of a gold standard method to determine spatially resolved tissue parameter maps (e. g. the sO_2 distribution over an organ surface). The solution is to resort to synthetic, i. e. simulated, biophotonic imaging data where we start with a “digital twin” of the tissue in question, including all tissue parameters, and, using a physical model, generate corresponding spectral data. This solves the missing reference problem, but at the cost of introducing new errors and uncertainties due to inadequacies in the simulation framework. The domain gap between the *in silico* and *in vivo* domain might be subtle and not at all obvious to the human eye. For example, the simulated spectra not only depend on the tissue in question but also on the light source. The human eye will quickly adapt to and ignore a change in light source, but an ML model which was trained on a specific light source will deteriorate in performance after a light source switch. In an open surgery setting, this might happen as easily as opening or closing the blinds or turning on the screen of a medical device. Hence, it is important to quantify the difference between the *in silico* and the *in vivo* domain.

As highlighted in the previous example, the realism of the simulated spectra can change dynamically based on environmental factors. Therefore, we treat research question D.2 as a continuous OoD detection task, where we filter an incoming stream of new spectral data for OoD data points. Depending on the exact application, the OoD data points could simply be dropped, or the practitioners could be made aware of a detected change in the spectra. This setting informs the requirements that the proposed method needs to satisfy. Due to its continuous nature, we need video rate capabilities. As we want to cover the broadest possible range of changes, we need a so-called unsupervised learning method that does not require OoD examples at training time but is trained on ID data only. To the best of our knowledge, there is no prior work toward continuous, task-agnostic OoD filtering in biophotonic imaging.

We propose INN ensembles with WAIC as OoD detectors for biophotonic imaging. They can be trained in an unsupervised manner, and they can run inference fast enough to enable video rate filtering. Therefore, we think they are natural candidates to make progress toward research question D.2.

D.3. Can physiological tissue changes be detected via an out-of-distribution approach?

With the previous research question, we can answer if the simulated data fit the *in vivo* data, but what do we do if they do not fit? In such a case, an ML model trained on simulated data will not generalize to the *in vivo* domain. The remaining option is to directly leverage the *in vivo* data. However, this option comes with its own challenges. First, we have already noted that we are lacking a gold standard reference method to annotate physiological tissue parameters. That is why we resorted to simulation in the first place. So, we cannot hope to regress tissue parameters. Instead, we can aim at a more coarse prediction of the tissue state, like if the tissue is currently perfused or not. Second, there is the problem that *in vivo* medical data sets are notoriously small. This is understandable because of the sensitive nature of the data. Nevertheless, it might prevent us from training standard classification models. Each human is different, and if this difference strongly impacts the data, in more technical terms, if the data exhibits a high inter-patient variability, then an ML model trained on a small patient cohort cannot generalize well to new, unseen patients. The new patient would always be OoD with regard to the (small) training cohort. The natural solution to this dilemma is to increase the cohort size such that the inter-patient variability is better represented in the data. However, this is a protracted process involving many regulatory hurdles. Instead, we propose a complementary approach. The approach is based on the observation that inter-patient variability is only of concern if we want to apply a fixed model to multiple patients. If we could train a *personalized* model, i. e. a separate model for each patient, we could ignore the inter-patient variability. To the best of our knowledge, such a personalized approach has not yet been used to detect physiological tissue parameter changes.

One requirement of a personalized model is that it cannot be trained in a supervised manner. For example, if we need perfused and ischemic spectra samples for the same patient, then these samples were already correctly classified by an alternative means, which would render our model obsolete. Instead, we phrase the problem as an OoD detection task, where we train an ML model to recognize the current tissue status quo as ID. Later in the procedure, changes to the tissue status quo can then be detected as OoD. This process directly introduces a second requirement. If we want to use the model on the same patient it was trained on, then the training time must be fast enough to be performed at the beginning of a surgical intervention. As a last requirement, the inference needs to run fast enough to be video rate capable.

With all these requirements in mind, we propose INN ensembles with WAIC as our unsupervised OoD detectors of choice. With suitable offline pre-training, they can be adapted fast enough (less than 1 min) and run inference fast enough for a surgical setting. We validate the approach with the example of a minimally invasive kidney surgery, more precisely, a partial nephrectomy. During such a procedure, the kidney is temporarily removed from the blood supply, and we will use our proposed method to monitor this change in perfusion state as a means to approach research question D.3.

3. Outline

This thesis consists of four parts: The introduction I[1], the background II, the contributions III, and the closing IV.

The introduction is made up of the motivation in section 1, the research questions of this thesis in section 2, and the outline in this section (section 3).

The background consists of two chapters. First, chapter 4 contains the fundamentals of inverse problems, biophotonics, and ML. The chapter closes with a section introducing the central architecture of this thesis: the INN. Second, chapter 5 lists the related work in the field of uncertainty quantification, posterior validation, and OoD detection.

Technical		**Medical**	
Research Question	Section	Research Question	Section
T.1	6.1	D.1	7.1 7.2
T.2	6.2	D.2	7.3
T.3	6.3.1	D.3	7.4

Table 3.1.: **Outline overview table.** Relation between research questions and where they are addressed in this thesis.

The contributions part is made up of a description of our framework for uncertainty handling in biophotonic imaging using INNs (section 6) and the experiments and results (section 7). Table 3.1 shows how the contribution sections relate to the research questions. The framework chapter addresses the technical research questions (T.1 – T.3). The chapter explains how cINNs can be used to generate posteriors for biophotonic inverse problems (section 6.1), how suitable metrics for the validation of multimodal posteriors can be chosen (section 6.2), and it introduces the widely applicable information criterion (WAIC) in section 6.3 and how the criterion can be used for OoD detection and tissue parameter monitoring. The experiments and results chapter describes our experiments to answer the domain research questions (D.1 – D.3). Sections 7.1 and 7.2 analyze the well-posedness of the inverse problems in multispectral imaging (MSI) and photoacoustic imaging (PAI) respectively. Section 7.3 contains our experiments regarding the realism of synthetic MSI

[1] Which we have almost finished.

data using OoD methodology based on INNs and WAIC. Finally, section 7.4 describes our study to detect ischemia in minimally invasive surgery via an OoD approach based on INNs and WAIC as in the previous section.

The closing part contains the discussion of our findings (chapter 8) and a conclusion with an outlook on open questions (chapter 9). Furthermore, the conclusion references back to the research questions in the introduction.

Part II.

Background

4. Fundamentals

This chapter is concerned with the necessary background material, i. e. the fundamentals, to benefit from this thesis. We will start with some terminology regarding inverse problems and how they can be phrased in a probabilistic framework (section 4.1). Afterward, we will introduce the inverse problems addressed in this thesis which stem from the biophotonic imaging domain (section 4.2). Then, there is a section about ML (section 4.3) in general, where we start with a placement of machine learning (ML) within the wider range of artificial intelligence (AI) and its relation to deep learning (DL). We continue with definitions of supervised and unsupervised learning and introduce out-of-distribution (OoD) detection, deep learning, and uncertainty in ML. The last section of this chapter (section 4.4) will introduce invertible neural networks (INNs) in detail as they are the main focus of this thesis. Next to the architecture, we will present the concept of calibration and common data preprocessing strategies.

4.1. Inverse Problems

Many tasks and problems we face in science and engineering can be phrased as so-called inverse problems. They arise whenever there are parameters of interest that are not directly observable but which are related to parameters we can observe, called *observables*. A general pattern is that based on e. g. physical models, we have an understanding of how the parameters of interest cause the observations. This is the *forward problem*. However, the way from the observables back to the hidden parameters of interest is often more involved[1].

A well-known example in the medical imaging domain is computed tomography (CT) imaging. A CT scanner collects narrow-angle x-ray measurements (the observations), but what we are actually interested in is a tomographic image. This tomographic image is created using a reconstruction algorithm, and this process is an instance of solving an inverse problem.

In section 4.1.1, we will introduce the mathematical definition of an inverse problem and what it means for an inverse problem to be *well-* or *ill-posed*. This is central terminology for the remainder of the thesis. In section 4.1.2, we will introduce a probabilistic view on inverse problems. This view is useful if the inverse problem is suspected to be ambiguous (i. e. to have multiple solutions) because it allows the representation of the solution to the problem as a probability distribution over the solution space.

[1] This is not to say that the forward problem might not be involved, too.

4.1.1. Mathematical Formulation

In the most general definition, an inverse problem is equivalent to solving an equation given by some function. Let X, Y be sets, where we call X the space of parameters and Y the space of observables. Let

$$F\colon X \to Y, x \mapsto y$$

be a mapping. We call F a *forward problem.* An *inverse problem* is given by the equation

$$F(x) = y, \tag{4.1}$$

for a given observable $y \in Y$, which we want to solve for $x \in X$.

In our settings, the sets X and Y, as well as the map F, have additional structure. We will restrict our attention to cases where X and Y are Banach spaces (i. e. normed, complete vector spaces) and F is continuous. In fact, X and Y will be finite-dimensional most of the time so that they can be thought of as some $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, but the following observations hold in the more general case[2].

As a first instructive example, let us have a look at a finite-dimensional case, where F is a linear map represented by a matrix $A \in \mathbb{R}^{m\times n}$, i. e.

$$F\colon X = \mathbb{R}^n \to Y = \mathbb{R}^m,\ x \mapsto y = Ax.$$

Then the corresponding inverse problem is solving the system of linear equations given by A for a fixed $y \in \mathbb{R}^m$. It is a well-established fact in linear algebra that there are three cases regarding the solvability of this equation [Fis14]: The linear system of equations has

1. no solution,
2. infinitely many solutions, or
3. exactly one solution.

This observation for even the most simple case motivates the following definition: An inverse problem given by equation (4.1) is called *well-posed*, if

Existence: for every observable $y \in Y$ there exists a solution $x \in X$ such that $F(x) = y$,

Uniqueness: the solution $x \in X$ is unique, and

Stability: the solution is stable, i. e. for a sequence $(y_n)_{n\in\mathbb{N}} \subset Y$ with $y_n \xrightarrow{n\to\infty} y \in Y$ and corresponding parameters $(x_n)_{n\in\mathbb{N}} \subset X$ with $F(x_n) = y_n$ and $x \in X$ with $F(x) = y$, we have that $x_n \xrightarrow{n\to\infty} x$.

If any of the above properties are violated, the inverse problem is called *ill-posed.* This definition goes back to Hadamard [Had02; Had23] who originally only used it for linear functions F. We note that the three properties build upon one another: Uniqueness

[2] An example when we need this more general case is if the inverse problem is given by a differential equation. In these cases, X is a space of functions that is often infinite dimensional.

only makes sense if existence is established, and for stability, we need the one-to-one correspondence established by existence and uniqueness. The three properties can be rephrased in more modern terms. Existence corresponds to F being onto (or surjective), and uniqueness corresponds to F being one-to-one (or injective). Both together imply that F is a bijection, i. e. that an inverse F^{-1} of F exists. The stability property is a bit technical, but it is equivalent to saying that the inverse F^{-1} is continuous. Taking everything together, we see that a well-posed inverse problem corresponds to a map F that is continuous and bijective with a continuous inverse. Such a map has a special name. It is called a *homeomorphism*.

It turns out that homeomorphisms introduce many restrictions on the form an inverse problem might take. For example, if we return to the finite-dimensional case[3] there is a topological result with the name *invariance of dimension* by Brouwer [Bro12]. This theorem states that if there is a homeomorphism between an $\mathbb{R}^n$ and an $\mathbb{R}^m$, then $n = m$. By this theorem alone, we can already see that any inverse problem where the X and Y dimensions differ is ill-posed!

As a last example in this section, let us quickly revisit the CT inverse problem. If we ignore artifacts introduced because of undersampled angles, one can show that for each $y \in Y$, a solution to the problem exists and is unique, but the stability criterion is violated. The reason for this is technical and lies in the fact that the Eigenvalues of the inverse problem[4] diverge [Her81]. This implies that the inverse is not continuous. Hence, small errors in y could lead to large deviations in x. So we see that even one of the most classic inverse problems in medical imaging is, in fact, ill-posed.

To conclude, the notion of a well-posed inverse problem is important, but we need a framework to address problems where one or more of the well-posedness properties are violated. The tools to represent them will be introduced in the next section.

4.1.2. Probabilistic Formulation

This section has been adapted from Stuart [Stu10]. In this section, we would like to perform a conceptual shift from a deterministic point of view, implied by well-posed inverse problems, to a probabilistic point of view, which is more suitable for ill-posed inverse problems. This shift is hinted at in figure 4.1.

As a starting point, let us consider how to find a solution $x \in X$ to an inverse problem given by equation (4.1). Even in the well-posed case, the properties are not constructive, i. e. they do not come with instructions to find x given y. In a very limited and special number of cases, an explicit formula for x given y might be derived. For all other cases, we need an alternative. This alternative often comes in the form of phrasing the problem as an

[3] Please note that F need not be linear.

[4] Which is a linear map.

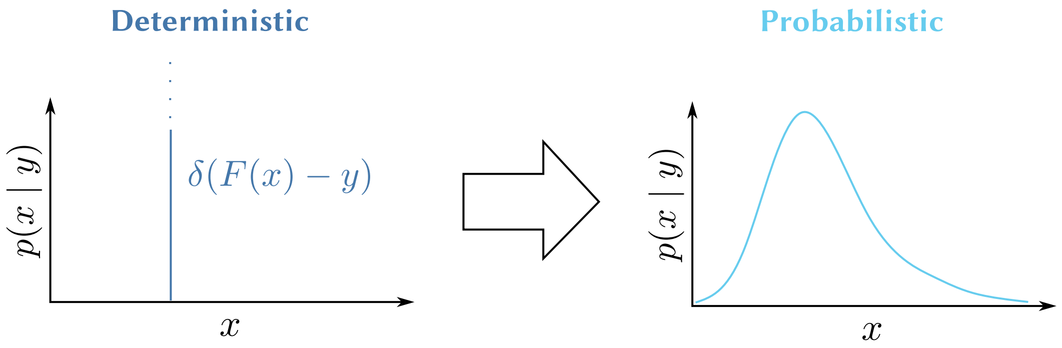


Figure 4.1.: **Conceptual shift from the deterministic to the probabilistic setting.** The deterministic inverse problem setting can be considered a special, degenerate case of the probabilistic setting, where the distribution is given by a Dirac δ distribution.

optimization task. We try to find

$$\hat{x} := \operatorname{argmin}_{x \in X} \|F(x) - y\|_Y^2. \tag{4.2}$$

The advantage of this formulation is that there are many (constructive) optimization algorithms. For example, if F is linear, we have the whole branch of linear optimization, or if F is differentiable, we can hope to use first-order optimization methods like gradient descent to find $\hat{x}$. Furthermore, if the inverse problem is well-posed, then $\hat{x}$ is guaranteed to be the unique solution to the inverse problem.

However, what if the inverse problem is ill-posed? Equation (4.2) has the uniqueness assumption ingrained, in the sense that we assume that a unique $\hat{x}$ can be found. So this optimization approach will always struggle if there is more than one solution to the inverse problem. However, even if the uniqueness property is satisfied, a lack of stability together with small errors in y might lead to a wrong minimum for x and large deviations. This stability property is generally addressed using *regularization*. This means that the values of x are drawn towards sensible regions of the space X based on prior knowledge. The most common way to do this is to choose an anchor $c \in X$ and penalize large distances between x and c via $\|x - c\|_X^2$. Hence, the optimization objective is updated to look like

$$\hat{x} := \operatorname{argmin}_{x \in X} \|F(x) - y\|_Y^2 + \|x - c\|_X^2. \tag{4.3}$$

This still does not address inverse problems with ambiguous solutions. Even worse, there is so far no reasoning from first principles, why the regularization term might make sense, and there is no motivation behind the choice of norms $\|\cdot\|_Y$ and $\|\cdot\|_X$.

A solution to this conundrum can be found by focusing on ambiguous inverse problems. What would be a good representation for the multiple solutions x belonging to an observation y? The most basic answer would be the pre-image

$$F^{-1}(\{y\}) := \{x \in X \mid F(x) = y\},$$

but there is a more informative alternative: Even though there are multiple solutions, some might be more "likely" than others. Hence, it might be practical to represent the solutions for y as a probability distribution $\mathbb{P}$ on the space X. We will assume that this distribution (which is conditioned on y) has a probability density $p(x \mid y)$[5]. In this *probabilistic setting* solving the inverse problem would correspond to finding the density $p(x \mid y)$ for a given y or equivalently (and often preferred because of computational reasons) the negative log-density $-\log p(x \mid y)$.

To see how this point of view can help our understanding of inverse problems, let us try to recover equation (4.3). For this, we need to invoke *Bayes' theorem* [Bay63; Kle13], which for densities reads

$$p(x \mid y) = \frac{p(y \mid x) \cdot p(x)}{p(y)} \propto p(y \mid x) \cdot p(x).$$

For the negative log-density, multiplication and division transforms to addition and subtraction, which yields

$$-\log p(x \mid y) = -\log p(y \mid x) - \log p(x) + \log p(x) = -\log p(y \mid x) - \log p(x) + \text{const}, \tag{4.4}$$

where the constant only depends on y, but not on x. We can interpret the quantities in equation (4.4):

- $p(x \mid y)$ is called the *posterior (distribution)*. It has been introduced above and encodes the solution to the inverse problem.
- $p(y \mid x)$ is called the *likelihood*. It represents our forward process, i. e. how the parameters x influence the observables y.
- $p(x)$ is called the *prior (distribution)*. It encodes our prior knowledge and beliefs about the parameter x.

As an example, let us assume that our forward process is "deterministic given by F with some noise" then the likelihood could be represented by a normal distribution such that $y \mid x \sim \mathcal{N}(F(x),\ \mathrm{id}_Y)$, where $\mathcal{N}(\mu,\ \Sigma)$ denotes a normal distribution with mean μ and covariance matrix Σ. Furthermore, let us assume that we expect, because of prior knowledge about the process, that parameters should be centered around $c \in X$, which could also be modeled by a normal distribution via $x \sim \mathcal{N}(c,\ \mathrm{id}_X)$. Plugging the density of a normal distribution into equation (4.4) yields

$$-\log p(x \mid y) = \frac{1}{2}\|F(x) - y\|_Y^2 + \frac{1}{2}\|x - c\|_X^2 + \text{const}. \tag{4.5}$$

[5] This might not be the case. In fact, in some of our settings, the data distribution is restricted to a low-dimensional sub-manifold of $\mathbb{R}^n$ so that such a density cannot exist.

We see that equation (4.5) is identical to equation (4.3) up to an additive constant and multiplication by a factor of $\frac{1}{2}$, which does not influence the location of the minimum of x. Taking this together, we can interpret equation (4.3) as

$$\hat{x} = \operatorname{argmin}_{x \in X} - \log p(x \mid y) = \operatorname{argmax}_{x \in X} \log p(x \mid y) = \operatorname{argmax} p(x \mid y)$$

or in other words, $\hat{x}$ is the maximum a posteriori (MAP) estimate of $p(x \mid y)$ given our choice of likelihood and prior. Hence, we have reproduced our optimization objective, but each term is now interpretable as a prior and a likelihood. So the regularization term and the norms $\|\cdot\|_Y$ and $\|\cdot\|_X$ are dictated by the chosen distribution and likelihood. Even more, we can quickly construct new optimization objectives by considering other distributions for the forward process or the prior (e. g. one could replace the Gaussian distribution with a Laplace distribution in the prior, which would change the L2 regularization to an L1 regularization).

Overall, we see that the probabilistic point of view can recover the optimization approach motivated by the well-posed inverse problem case, but it is more flexible. Indeed, we are not limited to parametric choices for the right-hand side, but we can also try to estimate $p(x \mid y)$ directly with a suitably flexible family of functions. By doing this, we can hope to represent truly ambiguous inverse problems that have multiple perceptively different solutions for a given y. One of the two main parts of this thesis lies in the exploration of such a family of functions (the INNs introduced in section 4.4) and the validation of the posteriors $p(x \mid y)$ that they generate.

4.2. Biophotonic Imaging Modalities

Biophotonic imaging is a family of imaging modalities that are based on the interaction of non-ionizing electromagnetic radiation in or close to the visible spectrum[6] with biological tissue. The common factor of all these modalities is that the tissue composition influences the light-tissue interaction, which in turn leads to a change in the spectrum of the incident light. This change is then measured by different means and used to gauge the state of the tissue. In light of the previous sections, it might be apparent that we are faced with an inverse problem. Our parameter of interest x is the tissue state, while our observable y is a spectrum. The exact nature of the spectrum depends on the biophotonic imaging modality, as introduced in the following paragraphs. In what follows, we will introduce the forward model, i. e. how the tissue causes the spectrum. Our approach to solving the inverse problem is deferred to section 4.4.2.

In section 4.2.1, we will introduce the general framework of light tissue interaction and highlight how the tissue composition leads to a spectral fingerprint. In sections 4.2.2 and 4.2.3, we will introduce the two main biophotonic imaging modalities of this work,

[6] Also known as light.

namely spectral imaging (SI) and photoacoustic imaging (PAI). The first modality, SI, uses (diffusely) reflected light to construct the spectrum, while the second modality, PAI, uses ultrasound waves generated by the absorbed light to construct the spectrum.

4.2.1. Light Tissue Interaction

In this section, we will introduce three processes by which light interacts with tissue: total/specular reflection, scattering, and absorption. These processes are sufficient for a top-level understanding of the image generation process in biophotonic imaging.

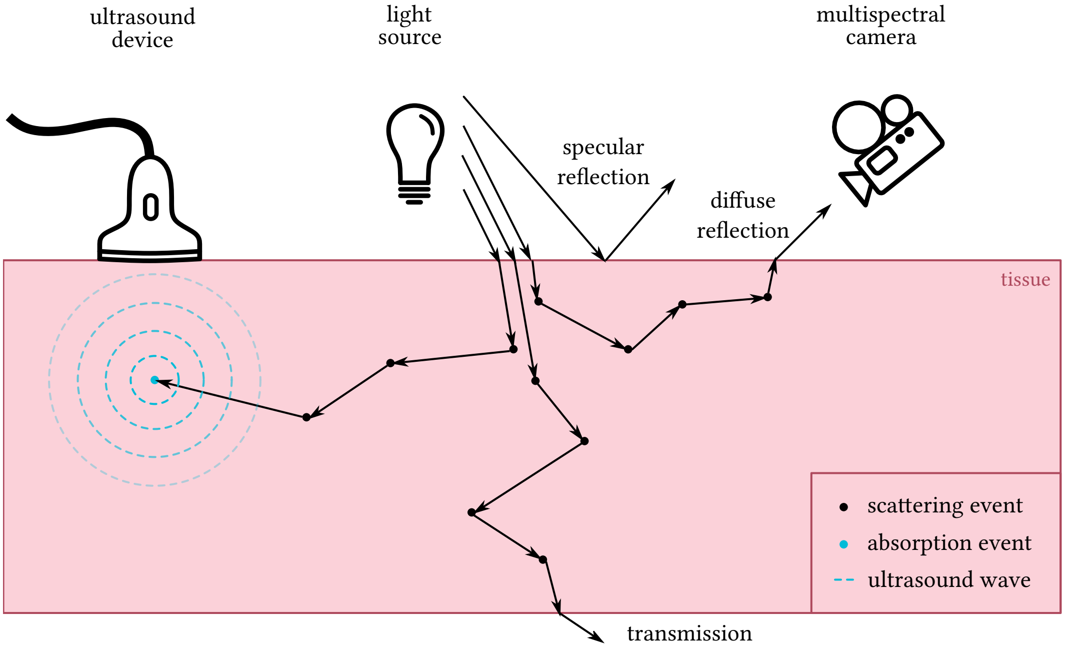


Figure 4.2.: **Schematic view of light tissue interaction.** The left-hand side shows how the interaction generates the PAI signal, while the right-hand side shows the MSI signal generation process. MSI: multispectral imaging, PAI: photoacoustic imaging.

If there is an interface between media with different refractive indices, then there is a critical angle, and if the incidence angle of the light is larger than this angle, the light will be reflected. This is called total or specular reflection, and this process is, in the best case, uninteresting and, in the worst case, a nuisance for our application. Indeed, the reflected light is not changed by the interaction with the tissue[7] and hence does not carry any information about it. Still, it is an effect that can be observed frequently in the surgical

[7] At least as a first approximation. The refractive index does depend on the wavelength of the light.

setting, and for our methods, we need to ensure to only work in regions without these specular reflections.

The other two processes, absorption and scattering can change the light spectrum and depend on the exact tissue composition. Therefore, we are most interested in light undergoing these two processes for our imaging modalities. In both scattering and absorption, a photon interacts with matter. The difference between the two is that after an absorption event, the photon is destroyed, whereas, after a scattering event, the photon remains, but its direction and/or its wavelength have changed from the original. The probability of a photon being scattered or absorbed depends on its color (i. e. its wavelength). This is encoded in so-called *cross sections* $\sigma_s(\lambda)$ for scattering and $\sigma_a(\lambda)$ for absorption (unit: m^2), where λ denotes the wavelength (unit: m). Both quantities vary with the tissue composition.

An overview of the previous discussion of the light tissue interaction can be found in figure 4.2. Multiple scattering events can lead to light leaving the tissue either on the same or another side. Hence, this cumulative process is either called *diffuse reflection* or *transmission.* If enough light is deposited in the tissue and subsequently absorbed, the stored energy can lead to a *thermo-elastic expansion* [XYW14], which relaxes as an ultrasound wave. This effect is used for PAI and will be handled in detail in section 4.2.3.

The above qualitative discussion can be made more quantitative using the radiative transfer equation (RTE) as derived in Wang and Wu [WW12]:

$$\frac{1}{c}\frac{\partial L(r,s,t)}{\partial t} = -s\cdot\nabla L(r,s,t) - \mu_t L + \mu_s \int_{S^2} L(r,s',t)\cdot p(s\cdot s')\,\mathrm{d}\,A(s') + S(r,s,t). \quad (4.6)$$

This differo-integral equation relates the *radiance* L (unit: Wm^{-2}) with other optical properties like the total absorbance μ_t and the scattering coefficient μ_s (units: m^{-1}) which might be functions of space r and time t and are themselves related to σ_a and σ_s. The variable $s \in S^2$ denotes a direction in 3D space and S (unit: Wm^{-2}) is a source term, due to light emitted in the tissue. Lastly, c denotes the speed of light and $p(s \cdot s')$ is a probability density function describing the probability that light from direction s' is scattered in direction s. All the above quantities can be wavelength λ dependent, but this dependence has been dropped in favor of a clearer exposition.

Under mild modeling assumptions, the RTE describes the light tissue interaction very well. However, in general, it is difficult to solve, and further simplifications or alternative methods like MC simulations are necessary to retrieve (approximate) solutions for L. Nevertheless, let us have a look at a strongly simplified model to ascertain that the tissue properties are, in fact, encoded in the light spectrum. To that end, we will make the following assumptions:

- We are in an equilibrium state, i. e. $\frac{\partial L}{\partial t} = 0$.
- Our medium is non-scattering, i. e. $\mu_s = 0$.
- There are no source terms, i. e. $S = 0$.
- We are only interested in the x direction, i. e. $s = \begin{pmatrix}1 & 0 & 0\end{pmatrix}^T$.

Plugging all these assumptions into equation (4.6) leads to the following much simpler equation:

$$L' = -\mu_t L, \tag{4.7}$$

where the prime denotes differentiation with regard to x. This first-order linear ordinary differential equation is straightforward to solve, and the solution is

$$L(x) = L_0 \cdot \exp(-\mu_t \cdot x). \tag{4.8}$$

We see that we have recovered a Lambert-Beer type law [Bou29; Bee52; Lam92] from the RTE which describes an exponential decrease of the radiance while the light travels through the absorbing medium. A common further assumption is that μ_t linearly depends on the absorption cross section σ_a^i times the concentrations n_i (unit: m^{-3}) of the absorbers i making up the medium. This can be expressed as

$$\mu_t(\lambda) = \sum_i \sigma_a^i(\lambda) \cdot n_i, \tag{4.9}$$

where we have reintroduced the wavelength dependence of μ_t and the σ_a^i. Taking the logarithm of equation (4.8), we get

$$\log\left(\frac{L(x,\lambda)}{L_0}\right) = \sum_i x \cdot \sigma_a^i(\lambda) \cdot n_i$$

which is a linear equation in the absorber concentrations n_i. If we know the absorption cross sections of our absorbers (or equivalently molar absorption coefficients as in figure 4.3 for (deoxy-)hemoglobin (Hb) and oxyhemoglobin ($\mathrm{HbO_2}$)) and can measure $L(x,\lambda)$ for multiple wavelengths $\lambda_1, \ldots, \lambda_k$, we get a system of linear equations, which we can hope to solve for the concentrations n_i. These, in turn, can be used to estimate medically interesting physiological parameters like blood oxygenation ($\mathrm{sO_2}$)

$$\mathrm{sO_2} \coloneqq \frac{n(\mathrm{HbO_2})}{n(\mathrm{HbO_2}) + n(\mathrm{Hb})} \tag{4.10}$$

or blood volume fraction (vHb)

$$\mathrm{vHb} \coloneqq n(\mathrm{HbO_2}) + n(\mathrm{Hb}), \tag{4.11}$$

where we replaced the n_i by $n(\cdot)$ to denote the concentration of the respective absorber.

We see that in this highly simplified setting, the radiance L (i. e. the light) contains information about the tissue. This holds true for the general RTE too, but recovering the information might be more difficult. We would like to point out that the Lambert-Beer law is regularly applied to regress tissue parameters [Bak+14]. However, we see that we needed strong assumptions to derive the equation[8]. This motivates our research of ML methods to tackle the regression task, as we hope to avoid some of these modeling assumptions.

With this rough understanding of light tissue interaction, let us introduce SI and PAI.

[8] Some of the assumptions were not strictly necessary, but e. g. a low-scattering medium is required.

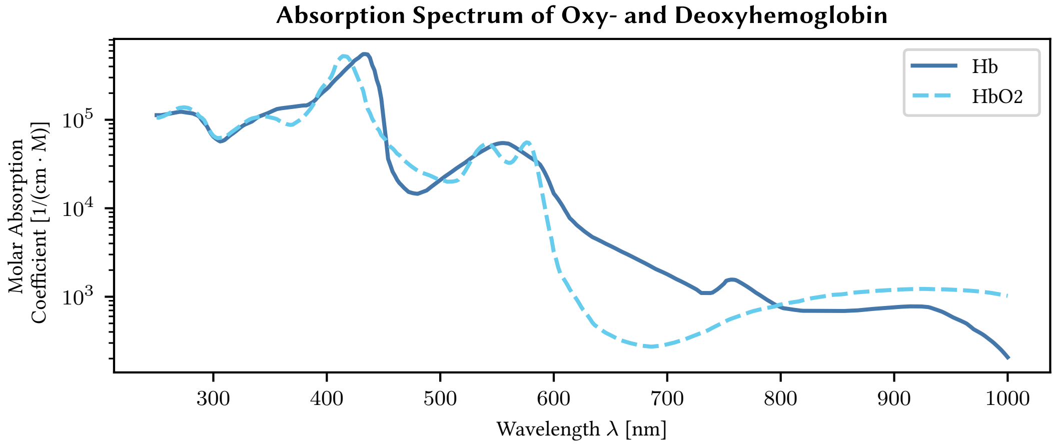


Figure 4.3.: **Differences in the molar absorption coefficient of Hb and HbO_2.** In the visible and near-infrared light spectrum, there are regions where (deoxy-)hemoglobin (Hb) and oxyhemoglobin (HbO_2) are well separated. Please note the logarithmic y-axis. The data for this figure was provided through [Pra98].

4.2.2. Spectral Imaging

In the previous section, we have convinced ourselves that the light spectrum contains tissue information. In this section, we will discuss how that spectrum is recorded in the case of SI.

Not all the light will be absorbed in the tissue, and since biological tissue is, in fact, highly scattering, we can hope for a large portion to leave the tissue and even on the same side as the light source. This process is called *diffuse reflection* and is depicted on the right-hand side of figure 4.2. In principle, our task is as simple as pointing a camera at the tissue. The question is: Which camera?

If a high spectral resolution is required, but spatial resolution can be neglected, we can use a spectrometer. In the opposite case, we can use so-called *multi-* or *hyperspectral cameras*. These cameras are similar to regular RGB (red, green, blue) cameras (see figure 4.4). They have a sensor with multiple pixels, which allow for spatial resolution. The cameras differ from RGB in that they have a higher spectral resolution. There are multiple approaches to collecting spectral information ranging from filter wheels to sub-pixels sensitive in narrow spectral bands. For more details regarding the hardware setup of multispectral cameras, we refer to [Cla+20]. The terms multi- and hyperspectral imaging are differentiated based on the number of bands. In the range of ten to 100 bands, the method is called MSI. Above this range, it is called hyperspectral imaging (HSI). An example for spectral responses of MSI cameras can be found in the experimental section 7.1 in figure 7.2.

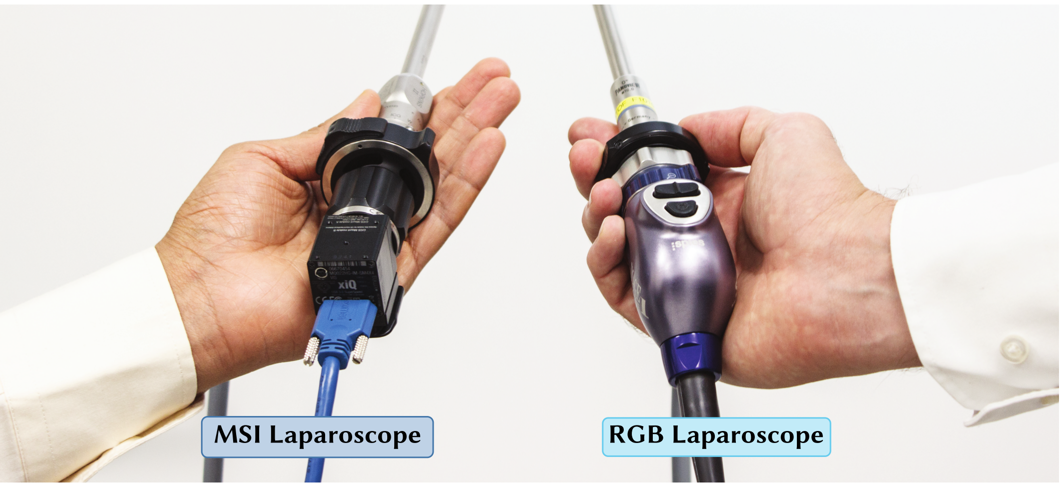


Figure 4.4.: **Comparison of an MSI camera and an RGB camera.** We can see that the form factor of the multispectral imaging (MSI) and RGB (red, green, blue) camera for laparoscopic surgery is virtually identical. The figure has been adapted from [Aya+22].

Simulation of SI data

Our overarching goal is to solve the inverse problem of recovering tissue information from the spectral information in SI recordings in a *data-driven* way. This requires training data which is impossible to collect *in-vivo*. To date, there is no gold standard to determine e. g. the oxygenation of an organ surface in a spatially resolved manner. Therefore, we have to resort to simulation. As introduced in section 4.2.1, we could try to (approximately) solve the RTE, but this approach is not very stable. Instead, we use MC simulation.

Qualitatively, we create a digital representation of our tissue with all necessary optical and physiological parameters. Then, we simulate an ensemble of photons that travel probabilistically through that tissue. In the end, we collect the fraction of photons that were diffusely reflected to determine the emitted spectrum.

The actual implementation is based on the Monte Carlo Multi-Layered (MCML) framework by Wang and Jacques [WJ92] and Monte Carlo eXtreme (MCX) framework by Fang and Boas [FB09]. Implementation details and the exact tissue model can be found in [Wir+16; Aya+22].

4.2.3. Photoacoustic Imaging

In SI, we collect the diffusely reflected light, i. e. the light that is not absorbed by the tissue. In PAI, we do the exact opposite. The absorbed light leads to a local change in temperature, which in turn leads to a change in pressure[9]. This pressure is relaxed as an ultrasound (US) wave which can be recorded using a standard US transducer. The effect which leads to this US wave is called *thermo-elastic expansion.*

The RTE in equation (4.6) describes the light distribution in tissue. In this section, we will explore the effects once the light is absorbed by the tissue. To this end, we follow and adapt the exposition by Xia et al. [XYW14].

The change in tissue volume $\mathrm{d}V$ due to heating can be expressed as

$$\frac{\mathrm{d}V}{V} = -\kappa p(r) + \beta T(r), \tag{4.12}$$

where κ is the isothermal compressibility (unit: Pa^{-1}) and β is the thermal coefficient of volume expansion (unit: K^{-1}). p and T describe the space-dependent pressure and temperature, respectively.

In PAI, irradiation is commonly achieved by very short laser pulses. Because of this and the inertia of the tissue, we can neglect the change in volume, i. e. $\mathrm{d}V = 0$, which allows us to solve equation (4.12) for the pressure

$$p(r) = \frac{\beta}{\kappa} T(r). \tag{4.13}$$

In a last step, we can relate the temperature to the absorbed light via

$$p(r) = \frac{\beta}{\kappa \rho C_V} \cdot \mu_a \cdot \Phi(r) =: \Gamma \cdot \mu_a \cdot \Phi(r), \tag{4.14}$$

where ρ is the mass density of the medium (unit: kgm^{-3}), C_V is the specific heat capacity at constant volume (unit: $\mathrm{Jkg}^{-1}\mathrm{K}^{-1}$), μ_a is the absorption coefficient of the volume (unit: m^{-1}), and Φ is the optical fluence (unit: Wm^{-2}), which relates to the radiance via

$$\Phi(r) = \int_{S^2} L(r, s) \, \mathrm{d}A(s),$$

where we dropped the time dependence for simplicity. Customarily, all parameters except the absorption coefficient and the fluence are collected in the *Grüneisen parameter* Γ (unit: Ks^{-1}). At a first glance, we seem to be in a similar position as when we derived the Lambert-Beer law in section 4.2.1. The parameter μ_a relates to the tissue composition via equation (4.9), and if we reintroduce the wavelength dependence λ, it seems that we can build a system of linear equations which we can hope to solve for tissue parameters.

There are two problems with this. First, the Grüneisen parameter Γ is not as constant as equation 4.14 makes it appear to be. Second and more importantly, if we look again at the RTE (4.6), we see that Φ also depends on μ_a. Hence, we have a non-trivial, non-linear relationship between μ_a and the pressure distribution p.

[9] This process requires that enough photons are deposited in the tissue in a very short period of time.

While so-called linear unmixing methods[10] are still prevalent in the PAI community, we can already see that their accuracy has to be limited [Hoc+19].

So far, we have neglected to ask ourselves how we can hope to measure the pressure distribution $p(r)$, and indeed this is impossible. Instead, this initial pressure is relaxed as an ultrasound wave following the wave equation

$$\frac{1}{c_s^2}\partial_t^2 p(r,t) = \triangle p(r,t) \tag{4.15}$$

with the initial condition $p(r,0) = p(r)$ and c_s the speed of sound of the tissue (unit: ms^{-1}). In the end, we record the traveling pressure wave at the skin or organ surface using an US transducer.

We see that the forward problem in PAI consists of two forward problems. The first is called the *optical problem* and describes how the tissue parameters[11] cause the initial pressure p. The second is called the *acoustic problem* and describes how the initial pressure distribution leads to the acquired US signal.

In theory, the acoustic inverse problem is well-posed because the wave equation is well-behaved. However, this neglects the fact that even in this case, we are missing information, in the sense that c_s can vary within the tissue, but there is no easy way to determine it. Furthermore, we have only a finite amount of transducer elements which, in the case of handheld devices, only measure the US wave from a limited view. All this negatively impacts the well-posedness. Still, there exist established methods to revert the US signal back to an initial pressure distribution like time reversal [XW05], delay-and-sum (DAS) [GJ82; Kim+16], and many derivatives [Mat+14; Kir+18]. Hence, many researchers focus on inverting the optical problem exclusively.

Simulation of PAI data

As in the SI case (see section 4.2.2), we would like to solve the PAI inverse problem(s) using *data-driven* methods. Hence, we need training data that relates the spectral information to the tissue parameters. Again as in the SI case, it is hard to solve the RTE, and so we resort to MC simulation for the optical problem.

The simulation is built on top of the MCX framework [FB09] which was wrapped in a python framework dedicated to PAI called Simulation and Image Processing for Photonics and Acoustics (SIMPA) [Grö+22]. Additionally, SIMPA wraps the k-Wave framework [Cox+04], which is used to solve the wave equation in the acoustic forward problem. For more details on how tissue is represented in the framework and how the simulation is carried out, we refer to [Nöl+21; Grö+22].

[10] This is the community-developed term for solving the system of linear equations implied by equation (4.14).

[11] And the light source characteristics.

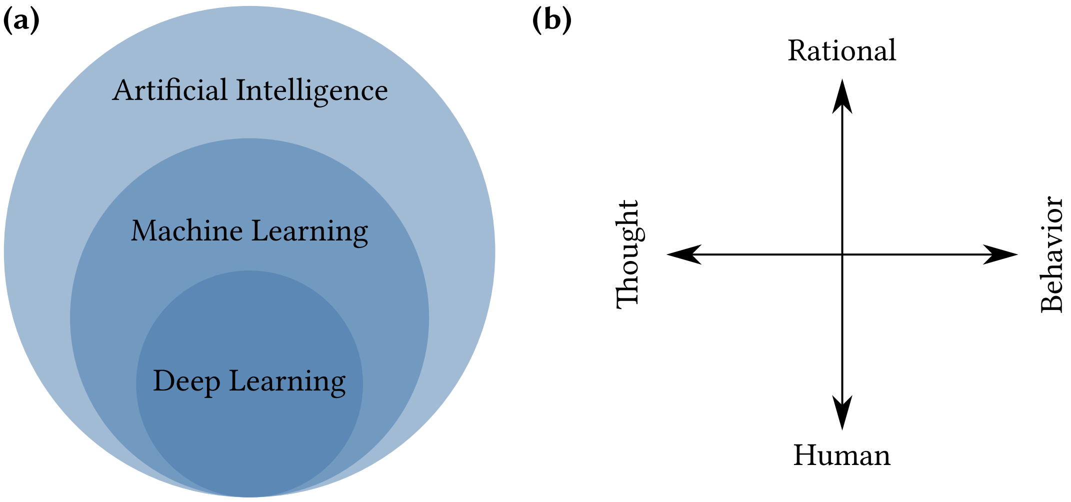


Figure 4.5.: **Outline of artificial intelligence. (a)** Relationship between AI, ML, and DL depicted as a Venn diagram. **(b)** Proposed axis to delineate and structure the definition of artificial intelligence as proposed in [RN20].

4.3. Machine Learning

Machine learning (ML) is a branch of the field of artificial intelligence (AI) (see figure 4.5), but what is AI? There is some debate about this question. All camps have in common that a machine or a computer is supposed to exhibit some sort of intelligence. Russell and Norvig [RN20], in their popular AI textbook, delineate this discussion along two dimensions

- human vs. rational and
- thought vs. behavior.

The first dimension asks the question of whether the AI algorithm in question should be judged based on its similarity to how a human would approach a problem or whether the algorithm should be judged on how rational it is. The advantage of the rational option is that it is more straightforward to formalize. The second dimension is concerned with the question of whether intelligence is "a property of internal thought processes and reasoning" [RN20] as opposed to intelligence being only detectable in "intelligent behavior" [RN20] which is external. Russell and Norvig [RN20] continue to argue that the rational-behavior quadrant is the most fruitful, i. e. that AI is concerned with understanding and/or creating agents that *act rationally* on the information and objectives given.

This definition is still broad. Let us look at an example to better understand the difference between AI and ML. One of the earliest general-purpose electronic computers, the ENIAC, was developed for the US Army and ready for work in 1945 [BB81]. It was mostly used for ballistics, i. e. computing the trajectory of missiles and bombs. In the very broad definition introduced above, this can be thought of as a very narrow AI. Given some input parameters (e. g. the speed and shape of the missile, the launch angle, etc.), the ENIAC would produce a trajectory. Assuming the programming was correct, the computer would act on the input data and produce a physically sound result. In this sense, one could say that the ENIAC acted rationally in its scope. Clearly, there is valid criticism calling the ENIAC an AI. The most central for us is that the programming was static. If the trajectory predictions turned out to be insufficient, there was no direct feedback loop to update the predictions. Instead, humans needed to locate inadequacies in their modeling of the process, update the model and then transfer it back onto the computer.

The ability to update its own prediction is a defining trait of ML, which sets it apart from other AI methods. ML introduces another layer of abstraction. In the above example, the program directly solved the problem at hand (finding a trajectory). In ML, the program has the potential to solve many problems. To get the program to solve a specific problem, it is taught what correct solutions look like. This requires so-called *training data* containing example solutions. So in ML, there is a two-stage process. First, we train the program, which is often called an *ML model*[12], on the training data. Then, we can ask for predictions on new input data.

For example, consider the conversion from the Celsius to the Fahrenheit temperature scale. We know that there is a simple conversion formula, and we could look it up and implement a program to do the conversion. This would be the classical approach from above. If we were determined to apply ML to the problem, we would first collect some data which would consist of pairs (x_i, y_i), where x_i is a temperature measurement in °C and y_i a corresponding measurement in °F. Then we would need an ML model, which we would like to train to represent this transformation. The most classic approach is called *linear regression*, which assumes a linear[13] relationship between the two quantities in question. In other words, the model takes the form

$$y = m \cdot x + b,$$

where m is the slope and b is the bias of our model, and the model is trained by choosing "optimal" m and b. If we assume there is some error in our measurements which is distributed according to a normal distribution, and if we avoid regularization, there is a closed form formula for optimal m, and b[14] [Has+09]. Afterward, we can plug in new temperatures x in °C and get predictions for the values in °F. We have chosen a very specific example where we know the underlying transformation, which is, in fact, linear. Hence, if the measurement errors are small, we can expect a good performance of our model for arbitrary values. In less simplified cases, this cannot be guaranteed, and in fact, another big

[12] Not to be confused with e. g. a physical model describing some physical process.

[13] More precisely, an affine relationship, but by abuse of notation, we will stay with the widespread use of linear.

[14] In general, finding these parameters is more involved and makes up a large portion of ML engineering.

problem an ML engineer is confronted with is to gauge the *generalization error* of the model, i. e. how much will the performance of the model drop when switching from training to unseen test data?

The temperature example was an instance of a *supervised learning* problem. We had input and output examples and wanted to learn a mapping between the two. In section 4.3.1, we will introduce the difference between *supervised* and *unsupervised learning* as we will see examples of both branches in the remainder of this thesis. In section 4.3.2, we will introduce out-of-distribution (OoD) detection, which is an unsupervised learning task that we use to explore the performance of our ML methods and which we apply in the setting of novelty detection. Although ML is a special case of AI already, it is still a large field. In the last ten years, the subfield of deep learning (DL) has received a lot of attention because of its great successes in e. g. computer vision, natural language processing, playing games, and protein folding [Goo+14; RFB15; Dev+18; Bro+20; Sen+20; Jum+21; Ram+22]. Most of the applied ML methods in this thesis belong to the category of DL, which will be introduced in detail in section 4.3.3. Since we are very much interested in *uncertainty quantification*, we will turn towards the errors that ML methods inevitably make. Is it possible for a model to know what it does not know? This subfield is called uncertainty quantification, and it is of special importance for safety-critical applications as we face them in a medical setting. It will be introduced in section 4.3.4.

4.3.1. Supervised and Unsupervised Learning

In the previous paragraphs, we have discussed learning based on data as an important defining trait of ML. In this section, we will delineate the field of ML based on the properties of the available data. This exposition has been inspired by [Jam+13; RN20].

We separate between supervised and unsupervised learning via the existence or lack of a response or target variable denoted by $y \in Y$. In supervised learning, there exists a source, predictor, or independent variable $x \in X$, the target variable y and we assume the existence of an (unknown) mapping $f\colon X \to Y$ that, together with possibly some noise ε, translates between the two, i. e. $y = f(x) + \varepsilon$. The task is to find an approximation for the unknown mapping f. In unsupervised learning, there is only the source variable x, but no y, and we are interested in understanding the structure of the distribution of x. Hence, unsupervised learning is often more exploratory in nature. Three applications that fall under the umbrella of unsupervised learning are low-dimensional projections/visualizations, clustering, and density estimation. A special case that arises from density estimation is OoD detection. That is the detection of samples that do not belong to the training distribution.

In the following, we will give a more mathematical introduction to supervised and unsupervised learning.

Supervised Learning

Let $X \subset \mathbb{R}^n$ be the source domain and $Y \subset \mathbb{R}^m$ be the target domain, and $f\colon X \to Y$ a continuous map. Furthermore, let

$$\mathcal{D} := \{(x_i, y_i) \in X \times Y \mid y_i = f(x_i) + \varepsilon_i,\ i = 1, \ldots, N\} \subset X \times Y,$$

be the *training set*, where ε_i is a noise variable[15] with zero mean and finite variance. In addition, we need a *hypothesis set*

$$\mathcal{H} := \{h\colon X \to Y \mid h \text{ satisfying some property}\}$$

which is a family of functions with which we hope to approximate f. Finally, there is a loss or cost function

$$l\colon Y \times Y \to \mathbb{R}$$

which we assume to be continuous and bounded from below. Then mathematically, the task of supervised learning corresponds to solving the optimization problem

$$\min_{h \in \mathcal{H}} \sum_{i=1}^{N} l(h(x_i),\ y_i). \tag{4.16}$$

In other words, we look for the hypothesis function $\hat{h} \in \mathcal{H}$ which minimizes the loss l between the prediction $\hat{h}(x_i)$ and y_i summed over all data points.

We can take some key observations from this

1. Equation (4.16) looks very similar to equation (4.2) from section 4.1. This highlights the connection between supervised learning and inverse problems.
2. Equation (4.16) is the case of "vanilla" supervised learning. Analogously to the inverse problem settings, we could augment the minimization with a regularization term r.
3. In principle, the family of hypotheses $\mathcal{H}$ could consist of all functions from X to Y, so with an empty property. However, this would make the optimization intractable. Hence, we need to choose properties that allow for efficient solution computation and properties that reflect prior knowledge about the function f. In the introduction of this section, we have seen the linear regression example. In this case $\mathcal{H}$ would consist of all linear functions from X to Y. This is a special case of a *parametric* family $\mathcal{H}$ because the whole family is parameterized by a set of parameters, which are m and b in the 1D case. Parametric families are quite common because they are easy to represent. In this case, we write $\mathcal{H} = (h_\Theta)_{\Theta \in \mathbb{R}^p}$, where $\Theta \in \mathbb{R}^p$ is the parameter vector of the family. Another common property is that the functions h_Θ depend differentiably on Θ. This allows for efficient optimization because first-order

[15] Assumed to be independently and identically distributed.

optimization methods like gradient descent are available [Cur44]. As an example of a non-parametric family, one can think of nearest-neighbor approaches like nearest-neighbor classification or regression. We omit to specify a family of functions but pay for this flexibility with the need to keep (part of) the training data in memory.

4. The loss function l is often given by an Lp-norm, i. e.

$$l(y, y') = \|y - y'\|_p^p = \sum_{j=1}^{m} |y_j - y'_j|^p,$$

but there are other common choices like binary cross entropy for classification, structural similarity for images, etc.

Supervised learning problems are commonly further divided depending on the target space Y. If the target variable is *discrete*, i. e. there is a discrete (often even finite) set of values that y can take, the learning problem is called a *classification task*. On the other hand, if the target variable is *continuous*, i. e. there is an infinite, non-discrete set of values which y can take[16] then the learning problem is called a *regression task*. For the inverse problems that we will consider, we will mostly be concerned with regression tasks. There are further subdivisions available (like segmentation and object detection). We refer the interested reader to [Mai+22b].

Unsupervised Learning

The big difference between unsupervised learning compared to supervised learning is that we only have a source domain $X \subset \mathbb{R}^n$, but no target domain Y and also no function f relating between the two. Instead, we have a *data distribution*[17] $\mathbb{P}_X$, and generally speaking, we would like to explore the properties of this distribution. To this end, we assume to have a data set

$$\mathcal{D} := \{x_i \in X \mid x_i \overset{\text{iid}}{\sim} \mathbb{P}_X,\ i = 1, \dots, N\},$$

where iid stands for identically and independently distributed according to the data distribution $\mathbb{P}_X$. An unsupervised learning problem is then given by a hypothesis family

$$\mathcal{H} := \{h \colon X \to \mathbb{R}^m \mid h \text{ satisfying some properties}\},$$

where m depends on the exact problem and a loss term

$$l \colon X \times \mathbb{R}^m \to \mathbb{R}$$

[16] More precisely, we should say that the distribution of y is absolutely continuous relative to the Lebesgue measure on $\mathbb{R}^m$ (that is where the *continuous* comes from), but that would take us too far.

[17] We could have talked about a data distribution in the supervised setting, too, but there it was optional. In the unsupervised setting, it is mandatory.

assumed continuous and bounded from below. The corresponding optimization problem is then given by

$$\min_{h \in \mathcal{H}} \sum_{i=1}^{N} l(x_i, h(x_i)). \tag{4.17}$$

Superficially, this setup looks very similar to the supervised case. We only drop all the y dependencies. Hence, many of the supervised observations directly translate to the unsupervised case (e. g. the use of parametric hypothesis families). Let us now turn toward specific unsupervised learning tasks and how they fit in this framework.

Low-dimensional Projection/Visualization In this case, m is often chosen to be 2 for planar visualizations. The most iconic example in this task group is the principal component analysis (PCA). For PCA, the family $\mathcal{H}$ is given by linear maps, and the loss l is chosen in such a way that the variance of the projected data contains the most variance of the original data distribution. We see that this loss definition is rather involved, which hints at the complications we face in unsupervised learning.

Density Estimation Here, we assume that the data distribution $\mathbb{P}_X$ has a density p_X, which we would like to approximate. Hence, $m = 1$ and the properties for $\mathcal{H}$ are that each function needs to be a density. So necessary conditions would be $h \geq 0$ and $\int_X h(x)\,\mathrm{d}x = 1$. Then we could maximize the (log-)likelihood of the data or, equivalently, minimize

$$\min_{h \in \mathcal{H}} \left[-\sum_{i=1}^{N} \log h(x_i) \right]. \tag{4.18}$$

Depending on the exact family $\mathcal{H}$, we might require more involved loss terms, as seen in the next paragraph.

Clustering Clustering is not as straightforward to fit into this framework. A common approach is to reduce it to a density estimation problem where we choose a family of densities $\mathcal{H}$, which can exhibit the required number of modes (which then translate to the clusters). A common example would be to choose a Gaussian mixture model (GMM), a distribution given by the sum of multiple Gaussians. This complicates the loss term, and the minimization is not as simple as depicted in equation (4.18). Instead, the expectation-maximization (EM) algorithm has to be applied, where, in one step, each data point x_i is assigned to its closest mode. Then the mode center and shape are updated based on the current members. These two steps are iterated until convergence.

In this thesis, we will encounter unsupervised learning mainly in the form of clustering and density estimation.

4.3.2. Out-of-Distribution Detection

In the community, there are many interrelated terms in the space of out-of-distribution (OoD) detection: novelty detection, anomaly detection, open set recognition, and outlier detection. Yang et al. [Yan+21] developed the framework of *generalized OoD detection* to define and explain the similarities and differences between these terms. In this section, we will adapt their work [Yan+21] to introduce the central notions for this thesis.

The starting point of the discussion is the existence of an in-distribution (ID) and out-of-distribution domain. Both domains are represented by a respective data distribution $\mathbb{P}_{\mathrm{ID}}(X \times Y)$ and $\mathbb{P}_{\mathrm{OoD}}(X \times Y)$ being a joint distribution in the input space X and a possible target space Y. Both distributions induce marginal distributions $\mathbb{P}_{\cdot}(X)$ and $\mathbb{P}_{\cdot}(Y)$. A shift in the input distribution, i. e. $\mathbb{P}_{\mathrm{ID}}(X) \neq \mathbb{P}_{\mathrm{OoD}}(X)$, is called a *covariate* or *sensory* shift, while a shift in the target distribution, i. e. $\mathbb{P}_{\mathrm{ID}}(Y) \neq \mathbb{P}_{\mathrm{OoD}}(Y)$, is called a *semantic* shift. Another dimension is whether OoD and ID data are available at the same time. The co-existence of OoD and ID data during training defines the outlier detection task. In all other cases, OoD data is detected in a test set after training. Detecting a covariate shift in the data is denoted as *sensory anomaly detection*, while detecting a semantic shift is classified as *classical OoD detection.* Depending on whether the ID data consists of a single class or multiple classes and whether a classification of these multiple classes is desired, the category of OoD detection is further split into semantic anomaly detection/single-class novelty detection, multi-class novelty detection, and open set recognition.

This thesis is mostly concerned with sensory and semantic anomaly detection. While they are philosophically different, the methods to tackle them are often quite similar. For example, having simulated spectra as ID data, detecting that *in-vivo* data looks different would be a sensory anomaly detection task. While collecting spectra from perfused organs as ID data and detecting spectra from unperfused (or ischemic) organs would be a type of semantic anomaly detection. Both are handled in this thesis (sensory shift: section 7.3, semantic shift: section 7.4).

By the definition of generalized OoD, it seems like a supervised learning task, but we have repeatedly introduced it as an unsupervised learning task. The discrepancy arises because in our setting during training, we have access to ID data only. Hence, we lack a supervision signal provided by a non-constant target space Y. If OoD data is available during training, the task is indeed supervised[18]. Whether OoD data is available during training or not, during evaluation, we have a regular binary classification problem, such that all classification metrics can be used to gauge the performance of the anomaly detection method.

There is some variety concerning OoD detection methods or, more precisely, anomaly detection methods. In the following, we will highlight three families that are commonly used to solve OoD tasks (see figure 4.6 for a visual representation).

[18] If there is a large imbalance in the availability of ID and OoD data, the task is sometimes referred to as *semi-supervised learning.*

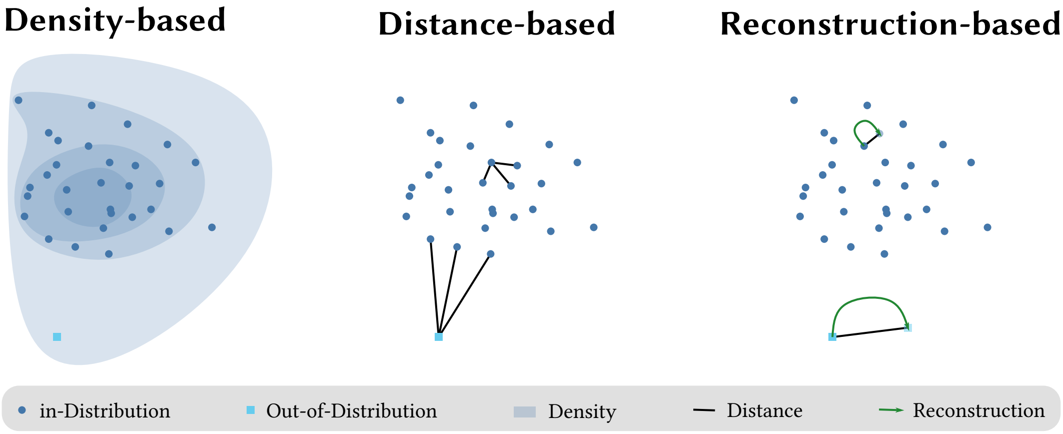


Figure 4.6.: **Visualization of different out-of-distribution (OoD) detection algorithm families. Left:** The density of the in-distribution (ID) data is estimated and used as an OoD score. **Middle:** A suitable distance to the ID data is used as an OoD score. **Right:** A data compression method is trained on the ID data and the reconstruction error is used as an OoD score.

Density-based Methods These methods are most closely related to unsupervised learning, as introduced in the previous section. The idea is to estimate the density of $\mathbb{P}_{\mathrm{ID}}$ and use it to detect OoD data points. The underlying assumption is that OoD data points should have a lower likelihood under the ID distribution than ID data points. While this assumption seems reasonable, there is empirical evidence, especially in high-dimensional spaces (think images), that it is not always satisfied [CJA18; Nal+19b; KIW20]. An alternative explanation could be that it is very hard to estimate densities in high-dimensional spaces because of the curse of dimensionality. To counteract this problem, new methods were developed that are based on density estimates but either post-process the density or project the input space to lower dimensions before they estimate the density. Examples are likelihood ratios [Ren+19], the test of typicality [Nal+19a], and WAIC [Wat13; CJA18], which will be a major focus of this thesis.

Distance-based Methods These methods are generally non-parametric and require keeping some representation of the training date, i. e. the ID distribution, in memory. The idea is to first transform the ID distribution to a semantically meaningful representation with a useful distance measure[19]. During inference, the distance of new data points to the ID distribution is computed. This often takes the form of a k-nearest neighbor (k-NN) distance computation. The underlying assumption of this approach is that OoD data should

[19] In lack of that, we can try our luck with the input space and the Euclidean distance.

be further away from the ID distribution than new ID data points.

Interestingly, while the inspiration for distance-based methods is quite different, it turns out that k-NN estimation can be rephrased as a density estimation task [BN06, chapter 2.5.2], such that, at least theoretically, there is little distinction between the two categories. Nevertheless, the conceptual shift makes it worth keeping both categories in mind. This observation also implies that we should expect failure cases as in the density-based setting.

Reconstruction-based Methods These methods are best interpreted as a learned compression algorithm. We try to find a compression algorithm on the ID distribution that optimally reconstructs the data. During inference, we can compress and decompress (or encode and decode) new data points. The underlying assumption is that the performance of this process should deteriorate for OoD data, which can be measured by the reconstruction error.

Reconstruction-based methods turn out to be density-based methods in disguise too. During the training of the compression algorithm, the reconstruction loss is often an approximation of the log-likelihood of the data. For example, if we use a variational auto-encoder (VAE), the loss term is the evidence lower bound, approximating the log-likelihood after convergence. Again, it is worth keeping the category separate because of the conceptual shift involved. Still, we should expect failure cases, just as for regular density-based methods.

4.3.3. Deep Learning

In section 4.3.1, which introduced supervised and unsupervised learning, we saw that their mathematical formulation depended on a family of hypothesis functions $\mathcal{H}$. The idea is to find a family that is flexible enough to approximate the unknown function describing our problem. Furthermore, we noted that these families are often parametric, i. e. $\mathcal{H} = (h_\Theta)_{\Theta \in \mathbb{R}^p}$ where $\Theta \in \mathbb{R}^p$ is the parameter vector. Simply put, deep learning (DL) chooses a very specific parametric family $(h_\Theta)_\Theta$, which turned out to be very flexible and powerful. Hence, DL is a special case of ML as depicted in figure 4.5. The hypothesis family used in deep learning consists of *neural networks (NNs)* or *neural nets* for short. Of course, there is far more nuance involved than simply specifying the hypothesis family, as we will see in the following. This treatise was inspired by [RN20].

Superficially, NNs are inspired by nerve cells, also known as neurons. These neurons are connected, and a previous neuron can change the electric potential of a successor. If this change is large enough, the successor will, in turn, "start to fire" and change the potential of the next neuron. This idea has been abstracted in the notion of a *perceptron.* A perceptron can either be active or inactive, encoded by 0 and 1. Each perceptron can have a different connection strength to neighboring neurons, which is indicated by a weight matrix W and a bias b. If x denotes the input states for our perceptron (i. e. the states of the perceptron that are connected to it), then the state of the current perceptron is given by the following rule:

$$\text{perceptron}(x) = \begin{cases} 1 & \text{if } W \cdot x + b > 0, \\ 0 & \text{else.} \end{cases}$$

So a single perceptron is a binary classifier that is based on a linear decision boundary. By definition, a perceptron can only work for linearly separable data sets, which clearly limits its expressive power. To remedy this, there was a research movement that tried to stack perceptrons. The result is called a multi-layer perceptron (MLP). This is the point of origin of the modern NN. However, there have been further adaptations that led to the performance we experience today.

First, note that a perceptron can be divided into two parts: 1. a linear map given by W and b and 2. a non-linear function applied to the output. This second function in today's terminology is an *activation (function).* In the classical perceptron, this activation was the Heaviside step function, which is impractical for stacked networks. As we will see in a second, NNs are trained using gradient descent. This implies that we need to be able to differentiate the operations in our network to update the parameters[20]. The step function is differentiable almost everywhere, but the derivative is 0. This renders gradient descent impossible. Therefore, the Heaviside function has been replaced by other activations. In fact, nowadays, there is a figurative zoo of activation functions with specific advantages and disadvantages. One of the most common activation functions is the rectified linear unit (ReLU), which is given by $\max(x, 0)$. At a first glance, this seems counter-intuitive because the ReLU is "almost linear" and the idea behind the activation is to make the whole perceptron non-linear, but it has been proven that NNs with non-polynomial activations and depth at least two[21] can approximate arbitrary continuous functions [Cyb89; HSW89; Pin99]. So the family of NNs is very expressive.

Contemporary NNs are very deep[22], and at least empirically, this seems to correlate with the good performance results. With growing depth, our parameter vector Θ grows larger too, and we need to adapt all these parameters in the training process. By construction, NNs are differentiable (almost everywhere) and hence it is natural to use gradient descent [Nes03] to update the parameters. For a single perception, it is easy to derive an equation for the derivatives, but for deeper NNs, this becomes rather tedious. Furthermore, this manual approach would make the whole approach very stiff because each change in network architecture and loss term would require a recomputation. This circumstance slowed the progress towards deeper NNs until, in the 60s, the backpropagation formula [Kel60; Bry61] was found and promptly forgotten until it was rediscovered in the 80s [RHW85; RHW86]. Slightly simplified, the backpropagation formula is the application of the chain rule of calculus to the special case of NNs. This leads to a set of equations that can be implemented in code to automate the computation of the derivatives. This, together with new software

[20] $\Theta = (W, b)$ in the perceptron case.

[21] The depth counts the number of stacked perceptron layers.

[22] There are ResNets [He+16] with more than 100 layers.

frameworks supporting automatic differentiation and more efficient hardware (like GPUs), allowed the training of deeper NNs.

Still, it took until 2012 before deep learning really took off. One reason for this might lie in the great flexibility of the model family: NNs can approximate a very large function space. How can we be sure that we found the right function in this huge space? Or another way to put it: If our data is noisy, how can we be sure that we actually learn the deterministic relation in the data and not the noise? The problem of fitting noise in the data is called *overfitting*, and because of their flexibility, NNs are very susceptible to it. While regularization terms and data augmentation can prevent overfitting to some extent, the most straightforward approach to counteract this tendency is more training data. Such large data sets were not available in the 80s but quickly became so with the advent of the internet. It is in this large data set regime where NNs really shine.

Another reason for their recent success lies in the ease with which NNs can be adapted to specific problem classes. This is often called introducing an *inductive bias* into the NN architecture. A simple example can be found in image classification. Multiple images of the same class, for cultural reasons, let us stick with cat images, might contain the object in question at different locations. The cat might be at the lower boundary or the left boundary, or in the middle. Still, all these images show a cat, and the NN should ignore the shift or translation of the cat. The observation that the classification result should be translation invariant led to the introduction of the convolutional neural network (CNN) [FM82; LeC+89; LeC+98]. These restrict the type of linear maps that are allowed to convolutions, which are inherently translation invariant. This introduction of prior knowledge should simplify the optimization task. While this turned out to be true, there is another reason why CNNs were so successful. Early NNs suffered from an explosion in the number of trainable parameters because the number of weights grew quadratically with the width of the NN and linearly with the depth. This quickly became prohibitively expensive. Convolutions solved this problem, too, because they drastically reduced the number of parameters.

There are more success stories like the CNN. To work with sequential data like time series, recurrent neural networks (RNNs) were introduced, propagating a state along the sequence dimension to capture that structure. Lately, transformers have beaten the RNNs [Hop82; HS97; Cho+14] on sequence tasks, especially in the domain of natural language processing but also in the domain of protein folding. Additionally, there is active debate, if transformers can also beat CNNs for computer vision tasks [Dos+21]. Transformers [Vas+17; Dev+18; Bro+20] introduce an attention mechanism that allows the NN to focus on specific parts of the sequence (or regions of the image in the computer vision case [Dos+21]).

In the last decade, NNs have led to astounding progress in the fields of computer vision [KSH12; Goo+14; RFB15], natural language processing [Dev+18; Bro+20], winning games [Sil+16; Sil+17; Sil+18; Vin+19], protein folding [Sen+20; Jum+21], and many more [Ram+22]. In recent years, we have seen how deep learning has found its way into the domain of medical imaging and computer-assisted interventions [Suz17; Zho+21]. However, the domain comes with its own challenges, so we are still missing the huge success stories we have seen in other areas [Mai+22a]. Still, deep learning is a very promising direction. One specialty of the medical domain is its high safety requirements. Hence, it is important for us to understand when and where our NNs fail. An important building block in this endeavor is uncertainty and its quantification, which will be introduced in the next section.

4.3.4. Uncertainty in Machine Learning

If we perform the exact same measurement twice, it is anecdotal wisdom that we cannot expect the same result. There is a grain of truth to this. Oftentimes, there is noise in the measurement process, such that repeated recordings lead to (slightly) different results. We have to cope with such uncertainties in the empirical sciences. However, not all uncertainties are the same. In the ML community [HW21] and medical statistics community [IM17], uncertainty is classified as either *aleatoric/statistical uncertainty* or *epistemic/systematic uncertainty*. This section aims to introduce both notions. This exposition was heavily influenced by Andrew Gelman's blog post [Gel22] and an article by Tony O'Hagan [Oha04].

Aleatoric uncertainty describes an intrinsic variability of a process. The classic example is the coin toss. Even if we know that the probability for heads is exactly $\Theta = \frac{1}{2}$, there will always be uncertainty in predicting the next toss. Mathematically, if $C \sim \text{Bernoulli}(\Theta)$ is the random variable representing the coin, then the aleatoric uncertainty can be expressed by the non-vanishing variance $\mathbb{V}\text{ar}[C] = \Theta \cdot (1 - \Theta) = \frac{1}{4}$.

On the other side, there is epistemic uncertainty which describes uncertainty due to our ignorance or lack of knowledge of the process. Another way of putting it is that we do not know the exact parameters of our model. A central distinction between the two types of uncertainties is that epistemic uncertainty is reducible by additional data, while aleatoric uncertainty is not.

To highlight the differences between aleatoric and epistemic uncertainty, let us modify the coin toss example. Let us assume that we know nothing about the coin, i. e. we do not know the probability for heads, but that $\Theta \in [0, 1]$. The Bayesian approach is to assume a prior distribution for Θ, and as we know nothing about the coin, we choose the uniform distribution $\mathcal{U}[0, 1]$ on the interval $[0, 1]$[23]. We can ask again what the probability that the coin toss is heads will be. This computes to

$$p(C = \text{head}) = \int_0^1 p(C = \text{head} \mid \Theta) \cdot p(\Theta) \, \mathrm{d}\Theta = \int_0^1 \Theta \cdot 1 \ \mathrm{d}\Theta = \frac{1}{2}.$$

The first observation is that the probability of heads for the (first) coin toss did not change, although we seem to know way less than in the first example because of the uncertainty in the parameter Θ. We could quantify this new uncertainty as in the previous example via the variance $\mathbb{V}\text{ar}[\Theta] = \frac{1}{12}$. Now, let us start collecting the coin toss results in a data set

$$\mathcal{D} = (h, h, h, t, t, t, h, t, h, h),$$

[23] The example does not depend on the specific prior as long as it's not degenerate (e. g. a point-mass).

i.e. $k = 6$ heads in $n = 10$ tosses. Using Bayes' theorem [Bay63; Kle13], we can compute a posterior distribution for Θ given $\mathcal{D}$

$$p(\Theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \Theta) \cdot p(\Theta) = \binom{n}{k} \cdot \Theta^k \cdot (1 - \Theta)^{n-k}.$$

We can identify this as the density of a beta distribution with parameters $\alpha = k = 6$ and $\beta = n - k = 4$. So $\Theta \mid \mathcal{D}$ is now distributed according to a beta distribution[24]. If we compute the variance, we obtain

$$\mathbb{V}\mathrm{ar}[\Theta \mid \mathcal{D}] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = \frac{6}{275} < \frac{1}{12} = \mathbb{V}\mathrm{ar}[\Theta].$$

Hence, the uncertainty in our parameter Θ has diminished. In other words, additional data has reduced the epistemic uncertainty we had in the parameter of our model. At the same time, there is still aleatoric uncertainty in the result of a coin toss, which no amount of data can reduce. In a symbolic formula, we can think of aleatoric and epistemic uncertainty via

$$\underbrace{p(C \mid \mathcal{D})}_{\text{total}} = \int_0^1 \underbrace{p(C \mid \Theta)}_{\text{aleatoric}} \cdot \underbrace{p(\Theta \mid \mathcal{D})}_{\text{epistemic}} \, \mathrm{d}\Theta,$$

where we assumed conditional independence $p(C \mid \Theta,\, \mathcal{D}) = p(C \mid \Theta)$.

Oftentimes (at least for me), discussions about aleatoric and epistemic uncertainties lead to confusion because what does it mean that aleatoric uncertainty is inherent and irreducible? For example, if we let a machine toss the coin with a very fixed setup and knew which side was up before the toss, etc. we could potentially reduce the uncertainty about the next toss. So is it really an epistemic uncertainty? The answer to this question is context. Aleatoric and epistemic uncertainties do not live in a vacuum but depend on the current context in the form of the experimental or model setup. In the above examples, we assumed we knew nothing about the coin's position before the flip. We could introduce this information, but that would change the model. In this new model, the uncertainty of the next coin toss could be reducible, which would transform it from aleatoric to epistemic uncertainty.

Uncertainty quantification in biophotonic imaging modalities is the goal of this thesis. We will encounter aleatoric uncertainty in the guise of ill-posed inverse problems (see section 4.4) and actually exploit epistemic uncertainty for more reliable OoD detection (see section 6.3).

[24] In fact, Θ was distributed according to a beta distribution already, since the uniform distribution is a beta distribution with $\alpha = \beta = 1$.

4.4. Invertible Neural Networks

This section will introduce the central deep learning architecture of this thesis: the invertible neural network (INN). It is well-suited to estimate complex, high-dimensional densities, and we will see how we can apply this to inverse problems and OoD detection tasks.

Similar to residual NNs and other deep learning architectures, there are standardized blocks that make up an INN. We will understand these blocks and see how they lead to important INN properties in section 4.4.1. The original INN architecture is mostly suited for unconditional density estimation. Hence, we will see how we can tweak the architecture for conditional approximation. This refined architecture is then called a conditional invertible neural network (cINN) and can be found in section 4.4.2. If we work on images, we need an invertible way to change the image resolution. Haar downsampling is introduced in section 4.4.3 to address this problem. In section 4.4.4, we will introduce z-score normalization, which is good practice before applying INNs to a data set. Lastly, we will introduce the concept of calibration for cINNs in section 4.4.5.

4.4.1. Building Blocks

Many tasks in ML can be phrased as having a complicated data distribution, and we would like to estimate its density function. If we achieve this, we can compute the likelihood for new samples to belong to the distribution or even try to generate new data points from it. Probability theory hands us a tool that can potentially accomplish the estimation of this distribution: the transformation theorem for probability densities.

Consider, we have a random variable on some space $X = \mathbb{R}^n$ with a probability density $p_X : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ and a diffeomorphism[25] $f\colon X = \mathbb{R}^n \to Z = \mathbb{R}^n$. Note that the dimensions of X and Z need to be the same as explained in section 4.1.1. This leads to a new random variable on Z, and the transformation rule tells us what the density of this new variable is. Let $x \in X$ then

$$p_X(x) = p_Z(f(x)) \cdot |\det Df(x)|, \tag{4.19}$$

where Df is the total derivative of f, $\det Df(x)$ is the Jacobi determinant of f at point x and $|\cdot|$ is the absolute value. Oftentimes, it is more practical to work with log-probabilities. In this case, the formula reads

$$\log p_X(x) = \log p_Z(f(x)) + \log |\det Df(x)|. \tag{4.20}$$

To return to our motivation of estimating complicated densities, we assume that p_X is our data distribution. We know how it transforms under general diffeomorphisms. So the problem of estimating p_X can be reformulated as the task of finding a suitable diffeomorphism f such that p_Z becomes a simple distribution like a Gaussian or a uniform

[25] This is a bijective, differentiable map with differentiable inverse.

distribution. This is where we enter ML territory. First, we need to find a parametric family of diffeomorphisms f_Θ, where Θ is the parameter vector of that family. Then, instead of using equation 4.20 to compute p_Z, we prescribe the distribution we want and use the equation as a loss-term to find the optimal Θ. In this thesis, we will only use Gaussians with mean $\mu = 0$ and covariance matrix $\Sigma = \mathrm{id}$ as p_Z. Let $(x_i)_{i=1}^N$ be our data set (i. e. a sample from the unknown training distribution). Then using the Gaussian assumption, the optimization task becomes

$$\max_\Theta \sum_{i=1}^N \log p_X(x_i) = \sum_{i=1}^N \left[-\frac{1}{2} \| f_\Theta(x_i) \|^2 + \log |\det D f_\Theta(x_i)| + \text{const.} \right] . \tag{4.21}$$

The loss given by equation (4.21) is called *maximum-likelihood loss* and using it for training is called *maximum-likelihood training*. Many optimization frameworks are implemented to minimize instead of maximize. If this is the case, we minimize the negative log-likelihood.

The last ingredient to estimate the density is the family of diffeomorphisms f_Θ, and this family is given by the INN architecture. From the above discussion, we can collect three necessary properties for INNs:

1. f_Θ needs to be invertible,
2. $\det D f_\Theta$ needs to be efficiently computable, and
3. f_Θ needs to be flexible enough to approximate a large set of distributions.

Especially the second point needs some thought because computing the determinant of an arbitrary matrix is computationally costly. This problem is addressed by *affine coupling blocks*. They were first introduced in a limited form in [DKB14] and then refined in [DSB16; Ard+18b; KD18]. Architectures that are based on these coupling blocks are often called *normalizing flow* architectures. This is motivated by the transformation rule because each step through the network slowly "normalizes" the input distribution toward a simpler distribution. In the setting of inverse problems, the name INN is more prevalent. A schematic overview can be found in figure 4.7.

An input $x \in \mathbb{R}^n$ is split into $x_0 \in \mathbb{R}^{n_0}$ and $x_1 \in \mathbb{R}^{n_1}$ such that $n_0 + n_1 = n$ ($n_i > 0$) and x_0 operates on x_1, while staying unchanged itself. Because x_0 is conserved, we can recover the transformations which were performed on x_1, and the whole process becomes invertible. In more detail, the affine coupling block consists of two functions $s \colon \mathbb{R}^{n_0} \to \mathbb{R}^{n_1}_{\neq 0}$ and $t \colon \mathbb{R}^{n_0} \to \mathbb{R}^{n_1}$. Both functions can be parameterized by arbitrary NNs. In particular, they do not need to be invertible. There is only the restriction that s needs to be nowhere zero (i. e. no single dimension is allowed to be zero at any time). This is easily achieved by exponentiating the output of a NN, but other alternatives exist. Let y_1 be the transformed x_1 then in formulas we have

$$y_1 = s(x_0) \odot x_1 + t(x_0), \tag{4.22}$$

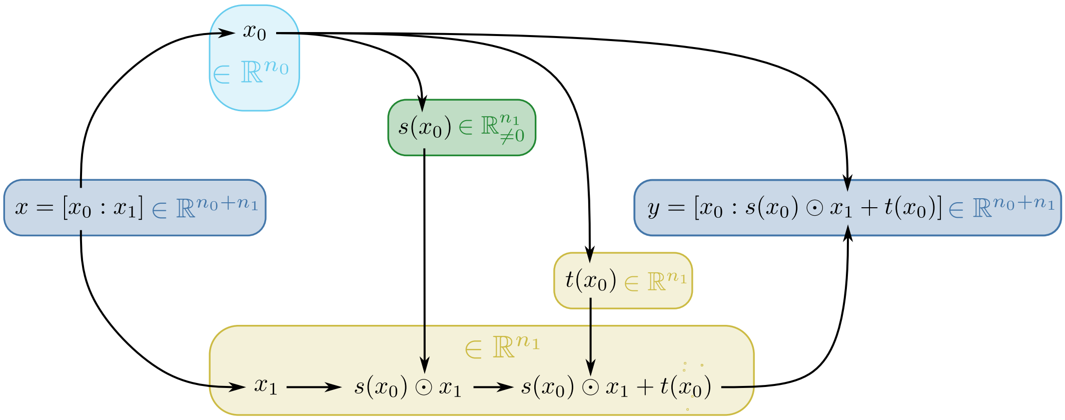

x_0
$\in \mathbb{R}^{n_0}$
$s(x_0) \in \mathbb{R}^{n_1}_{\neq 0}$
$x = [x_0 : x_1] \in \mathbb{R}^{n_0+n_1}$
$y = [x_0 : s(x_0) \odot x_1 + t(x_0)] \in \mathbb{R}^{n_0+n_1}$
$t(x_0) \in \mathbb{R}^{n_1}$
$\in \mathbb{R}^{n_1}$
x_1
$s(x_0) \odot x_1$
$s(x_0) \odot x_1 + t(x_0)$

Figure 4.7.: **Schematic depiction of an affine coupling block.** The color encodes the space on which each operation is performed. s and t are functions which can be represented by arbitrary neural networks, with the restriction that the output of s is non-negative in each component. $x_1 \in \mathbb{R}^{n_1}_{\neq 0}$ if and only if $x_{1,i} \neq 0$ for all $i = 1, \dots, n_1$.

where $\odot$ denotes the Hadamard product (i. e. component-wise multiplication of the two vectors). We see that equation (4.22) is an affine transformation which is parameterized by x_0. It is invertible because we never multiply by 0, and hence we can recover the original x_1 via

$$x_1 = (y_1 - t(x_0)) \oslash s(x_0), \tag{4.23}$$

where $\oslash$ denotes component-wise division.

More importantly, if we consider the whole coupling block as

$$\mathrm{cb}\colon \mathbb{R}^n \to \mathbb{R}^n, [x_0 : x_1] \mapsto [x_0 : y_1]$$

with y_1 given by equation (4.22) then the derivative $D\mathrm{cb}(x)$ of cb takes the following block form

$$D\mathrm{cb}(x) = \begin{pmatrix} \mathrm{id}_{n_0} & 0 \\ * & \mathrm{diag}(s(x_0)) \end{pmatrix} = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & 0 & \\ & & 1 & & & \\ & & & s_1(x_0) & & \\ & * & & & \ddots & \\ & & & & & s_{n_1}(x_0) \end{pmatrix}. \tag{4.24}$$

Indeed, since x_0 is left unchanged the first block is simply the identity id_{n_0} and the second block is 0, because x_0 does not depend on x_1. The third block can become complicated as the derivatives of s and t with regard to x_0 are non-trivial. However, as we will see in a second, we do not need to compute them. This is because the fourth block is diagonal once more given by $s(x_0)$ because the derivative of an affine transformation is just the multiplication part. Taking all this together, we see that the matrix is lower-triangular (i. e. there are only zeros above the diagonal), and the determinant of such a matrix is given by multiplying all diagonal elements. Hence,

$$\det D\text{cb}(x) = \prod_{i=1}^{n_1} s_i(x_0)$$

and with the absolute value and the logarithm, we can express this as a sum

$$\log|\det D\text{cb}(x)| = \sum_{i=1}^{n_1} \log|s_i(x_0)|, \tag{4.25}$$

which is efficiently computable as long as s is.

The affine coupling block satisfies the first two properties for our family f_Θ, but in isolation, they are too stiff in some regards. First, we note that we need to be able to split the input for the whole architecture to work. This leads to a minimum input dimension of $n = 2$. Furthermore, empirically it turns out that if the dimensional split is carried out identically each time, then the information of the separate dimensions does not *mix well*. This is counteracted by adding a fixed but random permutation in front of every coupling block. A permutation leads to a Jacobi-determinant of ± 1, which can be ignored after taking absolute values. Recently, this approach has been extended. Permutations are a subgroup of a group called the group of orthogonal transformations. They are represented by matrices $A \in \mathbb{R}^{n\times n}$ with $A^t \cdot A = \text{id}_n$, i. e. their transpose is their inverse. Hence, the inverse is efficiently computable, and again the Jacobi-determinant is ± 1. It is hard to parameterize the whole group of orthogonal transformations, so instead, only certain subgroups are considered, which are parameterized using reflections along randomly initialized hyperplanes in $\mathbb{R}^n$ [TW16]. However, there is no empirical evidence that these more flexible transformations improve performance, so we stick to fixed permutations in this thesis.

There are further caveats with the affine coupling blocks. As mentioned previously, s needs to be nowhere zero, and this is generally achieved via component-wise exponentiation. However, exponential functions easily lead to exploding or vanishing gradients. To counteract this tendency, we employ a clamping scheme. If $\tilde{s}$ is the NN output before exponentiation, we use the area hyperbolic tangent with a constant $c > 0$ via

$$\bar{s}(x_0) = c \cdot \text{artanh}\left(\frac{\tilde{s}(x_0)}{c}\right).$$

This enforces $-c < \bar{s}(x_0) < c$ and then setting $s(x_0) := \exp(\bar{s}(x_0))$ bounds s as well. This slightly restricts the expressibility of the coupling blocks, but this is no disadvantage, as stacking more blocks on top of each other will diminish the influence of the clamping. At the same time, at the beginning of training, this can still lead to instabilities. The weights are randomly initialized, and even if exponentiation bounds the block output by $\exp(c)$, a network with d blocks can still end up with $\exp(dc)$ as the order of magnitude for the network output. This can lead to floating point overflows. After a couple of iterations, this is no longer an issue, but we can encounter NaNs in the beginning, such that the training needs to be restarted with a new seed. In [KD18; Mac+21], a remedy was suggested. As we have seen above, affine transformations are well suited for invertible operations. So the idea was to add an affine transformation after each affine coupling block. The difference to the coupling block above is that the new transformations do not depend on the input. Hence, they were named *global transformations*. The exact form of this global transformation is

$$\mathrm{gt}(x) := \frac{1}{c}\log(1+\exp(c\cdot\sigma))\odot x+\tau = \mathrm{softplus}(\sigma;c)\odot x+\tau, \tag{4.26}$$

where $\sigma, \tau \in \mathbb{R}^n$ are parameters of the transformation, which are trained during optimization and $c > 0$ is a hyperparameter. The softplus function is always positive. Hence, the transformation is invertible, and the log-Jacobi-determinant can be computed analogously to equation (4.25). At initialization, σ is set to all 1s and τ to all 0s. If c is chosen smaller than 1, the transformation will squeeze the output of the coupling blocks. This prevents overflows. At the same time, the optimization can update σ and τ, and this can revert the initial squeezing by c. So once the coupling blocks have learned to transform the data roughly, the influence of the global transformations will slowly diminish. Hence, they do not negatively impact the expressiveness of the architecture.

The affine coupling blocks, together with the permutations, the clamping, and the global transformation, is the basic building block of the INN. Similarly to other architectures, like residual NNs, we can build very expressive INNs by stacking these blocks on top of one another. In fact, we still have not addressed the third property of our invertible family f_Θ. Ishikawa et al. [Ish+22] have shown that INNs are universal diffeomorphism approximators. This is a very general class of functions and basically the best we could hope for. So it turns out that INNs are very flexible, and we can hope to approximate many practical distributions.

Lastly, we would like to mention that the INN architecture is not the only possible parameterized family f_Θ. For a good comparison of different invertible architectures with their advantages and disadvantages, we refer to [Kru+21].

4.4.2. Conditional Invertible Neural Networks

In the previous section, we have seen how INNs can be used to estimate a distribution with density p_X. This is already useful for OoD detection (cf. section 4.3.2). However, in the setting of inverse problems, we need to estimate conditional probability densities, i. e. $p(x \mid y)$. In this section, we will see how the INN architecture can be adapted to achieve this.

We will stick to the convention introduced in section 4.1.1. We have a forward problem $F\colon X \to Y, x \mapsto y$ and we would like to train an INN to learn the inverse problem in a probabilistic manner (cf. section 4.1.2) via $p(x \mid y)$.

The original idea [Ard+18b] was to use INNs to learn the forward and inverse problem jointly. However, as presented in section 4.1.1, this is only possible for well-posed inverse problems where $\dim X = \dim Y$. For the overwhelming majority of real-life inverse problems, this is simply not the case. Hence, the introduction of padding dimensions became necessary. Two latent spaces Z_X and Z_Y were introduced and appended to X and Y, respectively, to guarantee

$$\dim X + \dim Z_X = \dim Y + \dim Z_Y.$$

Z_X is only necessary if $\dim Y > \dim X$, but Z_Y is necessary independently of any dimensional restrictions as this latent space is supposed to encode all ambiguities which an ill-posed inverse problem might have. The intuition behind this is, if there are two $x_0, x_1 \in X$ with $x_0 \neq x_1$, but $F(x_0) = F(x_1)$ then the trained INN would also collapse the two points x_0 and x_1, but that is not possible due to the intrinsic invertibility. Hence, the lost information has to be encoded in Z_Y.

Practically, this whole approach leads to many problems during training. The maximum-likelihood training as introduced in equation (4.21) is no longer possible because the dummy dimensions due to Z_X and Z_Y would lead to vanishing or exploding Jacobi-determinants. In its place, four loss terms are necessary. Let

$$f_\Theta : X \times Z_X \to Y \times Z_Y,\ (x, z_x) \mapsto (y, z_y)$$

be the INN. For a data point (x, y) and latent space samples z_x and z_y, we denote by $(\hat{y}, \hat{z}_y) = f_\Theta(x, z_x)$ the forward prediction of the INN and by $(\hat{x}, \hat{z}_x) = f_\Theta^{-1}(y, z_y)$ the backward prediction. Then the four loss terms can be described as follows:

1. There is a forward loss which draws $\hat{y}$ to y. This loss depends on the application. A common choice is the L2 loss.
2. We would like to shape the distribution on Z_Y as a standard Gaussian distribution so that we can sample from it. As such, we enforce that the $\hat{z}_y$ are distributed accordingly. This is achieved using maximum mean discrepancy (MMD), which is a metric on distributions and can be computed on a sample. Details about MMD can be found in section 6.2.3 and [Gre+12].
3. We assume that $p(x \mid y)$ varies smoothly in y. So we disturb $\hat{y}$ and $\hat{z}_y$ and compute the inverse f_Θ^{-1} and assume that $\hat{x}$ should still be close to x. This backward loss acts as a regularizer and can be implemented as another L2 loss.
4. Given y and latent space samples z_y drawn from a standard Gaussian $\mathcal{N}(0, \mathrm{id})$, we assume that $(\hat{x}, \hat{z}_x)$ should have a joint distribution given by the training data x and another Gaussian $\mathcal{N}(0, \sigma^2\, \mathrm{id})$ on Z_X. This is enforced using another MMD loss term.

The four loss terms have to be balanced against each other, and the MMD terms introduce further hyperparameters. Hence, the question arose whether there was no better way to represent a conditional probability density in this framework.

We are interested in $p(x \mid y)$ or equivalently in $\log p(x \mid y)$. Looking at the transformation rule in equation (4.20), there are two natural places where we could hope to integrate the conditioning y:

$$\log p(x \mid y) = \log p_Z(f(x;y)) + \log |\det Df(x;y)| \quad \text{or} \tag{4.27}$$

$$\log p(x \mid y) = \log p_{Z|Y}(f(x) \mid y) + \log |\det Df(x)|. \tag{4.28}$$

In equation (4.27), we have a density on the space Z which is independent of y. Instead, the transformation f changes with y. So depending on the conditioning y, the transformation looks different[26]. Equation (4.28) goes the opposite direction. In it, the transformation is independent of Y, but the target distribution $p_{Z|Y}$ now depends on the conditioning. Both approaches have been used. The first leads to an architecture that is called a conditional invertible neural network (cINN) and was introduced in [Ard+19]. The second is often implemented with a mixture of Gaussians in the latent space. This works well for categorical Y, where each mixture component corresponds to one of the classes. It has been introduced in [Ard+20] and needs further tweaking of the maximum-likelihood loss. Since in our inverse problems setting, we are almost exclusively faced with continuous Y, equation (4.28) does not suit our needs. So in the remainder, we will focus on the cINN architecture.

The building block of the INN is the affine coupling block (see equation (4.22)). We use NNs s and t to operate on part of the input and transform the other part. In order to make this coupling block conditional on y, we can simply make s and t conditional on y, i. e.

$$\begin{aligned} s(x_0) &\mapsto s(x_0, y) \quad \text{and} \\ t(x_0) &\mapsto t(x_0, y). \end{aligned}$$

In theory, this is all we need to move from an INN to a cINN. However, there are some details worth mentioning. First, we have to be careful if our data has a certain shape. So far, we have only talked about a general $\mathbb{R}^n$, but often times the different dimensions come with meaning and are collected accordingly in axes like $\mathbb{R}^{n_c \times n_w \times n_h}$ for an image, where n_c denotes the channel dimensions and n_w and n_h the width and height of the image respectively. Respecting this semantic information and incorporating it into the architecture is often beneficial. For INNs and cINNs, this means that the splitting in the affine coupling blocks is generally performed along the channel dimension while keeping the spatial dimensions intact. The practical impact is that the easiest way to incorporate the conditional information y is if it has the same spatial dimensions as the input x_0. Theoretically, it is possible to work around it, but practically this is a “golden rule”.

[26] Please note, that the derivative Df is only computed with regard to x and also the invertibility is only guaranteed with regard to x.

Besides this restriction, there is also a big advantage to cINNs. Compared to the original architecture, the network is no longer invertible towards y. This allows for much more flexibility regarding possible preprocessing of the conditioning input y. Especially for high-dimensional y, one can employ a so-called *conditioning network*, which can be an arbitrary NN, to transform the conditioning to a better representation for the actual cINN. Furthermore, this conditioning network can be pre-trained on other tasks to simplify the actual training. Lastly, the conditioning network is not restricted to a single output head but could have multiple heads specialized for each coupling block. This allows for a lot of flexibility.

Let us close with the remark that cINNs remove the need for the zero-padding (Z_X and Z_Y) in the original architecture, and we are left with a single latent space that has the exact same dimensions as the input space X. With that, maximum-likelihood training becomes available again, removing the need for multiple loss terms and drastically stabilizing the training procedure.

4.4.3. Invertible Downsampling

Many contemporary NN architectures use downsampling operations to reduce the spatial resolution of images and similar structures, but common choices like max or mean pooling are not invertible. However, we would like to be able to operate our INNs or cINNs on multiple resolutions whenever we work with image data. Hence, we need an invertible downsampling operation. There are multiple options for that. The trick is that spatial resolution is transformed into channel resolution, i. e. if the spatial dimensions of the images are halved, the channel dimension is quadrupled. We know that this is a necessary condition to keep the dimensions constant. In this thesis, we will use the *Haar downsampling* operation [Haa09] as suggested by [Ard+19].

The downsampling operation can be implemented as a 2×2 convolution with four output channels per input channel and fixed kernels given by the four matrices

$$\frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}, \text{ and } \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{4.29}$$

Hence, each $1 \times 2 \times 2$ input patch is transformed to a $4 \times 1 \times 1$ output patch. The inverse is given by four 1×4 matrices applied to the $4 \times 1 \times 1$ output patch. If the four input patch members are ordered in "reading order", then the matrix for each patch member is given by

$$\frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} 1 & -1 & 1 & -1 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} 1 & 1 & -1 & -1 \end{pmatrix}, \text{ and } \frac{1}{2}\begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix}, \tag{4.30}$$

which are simply the flattened versions of the convolution kernels given in equation (4.29). Hence, the inverse can also be computed with standard NN building blocks like a 1×1 convolution followed by reshaping the tensor.

4.4.4. Data Normalization

The affine coupling blocks with the exponential clamping can change the volume of the distribution only by a finite amount. This is counteracted to some extent by the global affine transformation. Nevertheless, if the volume of the input data distribution differs strongly from the latent space distribution, which has volume 1 by design, the INN might not be able to learn the transformation. Even if it is capable, we would "waste" capacity of the INN on simply squeezing or stretching the distribution, while we could as well normalize the volume of the input distribution. This has the added benefit that we do not need to adapt the noise augmentation to the specific scales of the input data. We use *z-score normalization*, which affinely transforms the data set to have zero mean and unit variance. This normalization is applied along each feature dimension separately, and if $(x_i)_{i=1}^n$ is such a feature of the data set, the normalized feature set $(y_i)_{i=1}^n$ is computed as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.31}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2} \tag{4.32}$$

$$y_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}. \tag{4.33}$$

One has to be careful not to introduce data leakage through this type of normalization. To this end, the mean and standard deviation is only computed on the training set and afterward applied to the validation and test set(s). This might lead to non-zero means and non-unit variance on these validation and test sets.

While z-score normalization is not so important for the conditioning input, having all features on the same scale and simplifying noise augmentation is still beneficial. Therefore, we also apply z-score normalization to the conditioning input throughout this thesis.

4.4.5. Calibration Curve and Error

We have seen how we can use cINNs to generate posterior distributions. While, in general, the maximum-likelihood training ensures that the posterior relates to the ground truth values, it might be that its width does not correspond to the model's uncertainty. This can happen e. g. because the noise augmentation is used too aggressively. In this case, the posteriors would be too wide because the model would take up on the added uncertainty due to the augmentation. In this light, it seems appropriate to ascertain that the shape of the posteriors is indeed reliable. The tool for this task is the *calibration curve* as seen in figure 4.8.

We need a test set with ground truth values to construct the calibration curve. For each data point, we create a posterior, then we can determine the quantile of the ground

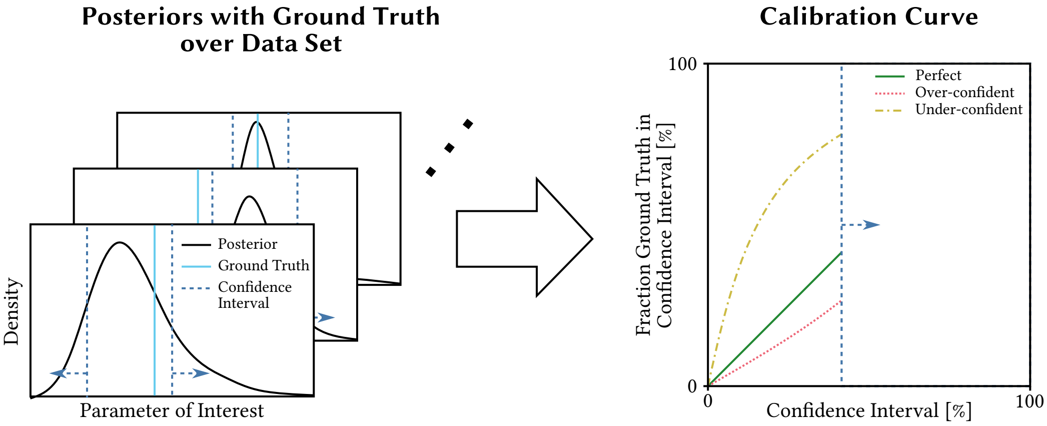


Figure 4.8.: **Schematic view of the calibration curve construction.** We scan all confidence interval (CI) levels and count the fraction of ground truth values within the corresponding CIs. The identity corresponds to a perfect calibration, while a curve above indicates under-confident posteriors and a curve below over-confident posteriors.

truth within that posterior. Based on the quantile, we know what the smallest confidence interval (CI)centered around the median is that contains the ground truth. Then, over the whole data set, we scan all possible levels for CIs and count the fraction of ground truths in the CI of that level. By definition, the CI to level 0 is empty and hence contains no ground truth values, and the CI to level 1 is the whole real line, so it contains all ground truths. Hence, we obtain a calibration curve $\mathrm{CC} : [0, 1] \to [0, 1]$ with $\mathrm{CC}(0) = 0$ and $\mathrm{CC}(1) = 1$.

The underlying assumption behind the above construction is that for the posteriors to be well-calibrated, the confidence level of the interval should correspond to the fraction of ground truths within this CI. So, for example, the CI to the confidence level of 0.3 should contain 30 % of the ground truths. This implies that a perfect calibration would be represented by the identity. A curve that runs below the identity implies that the CIs contain fewer ground truths than expected. This means that the posteriors are too narrow or, in other words, that the model is *over-confident*. On the other hand, if the curve runs above the identity, there are too many ground truths in a given CI. Hence, the posteriors are too wide or, alternatively, the model is *under-confident*. In some cases, the calibration curve crosses the identity one or more times in the interior of $[0, 1]$. If so, the model is over- or under-confident on the respective sub-intervals.

Often, it is more practical to look at the deviation from the identity. This quantity is called the *calibration error*

$$\mathrm{CE}\colon [0,1] \to [-1,1],\ \mathrm{CE}(x) := \mathrm{CC}(x) - x, \tag{4.34}$$

where CC denotes the calibration curve.

This still leaves us with a curve to interpret. If we want to aggregate the information further, we can use the expected calibration error (ECE) which is given by

$$\mathrm{ECE} := \int_0^1 |\,\mathrm{CE}(x)|\,\mathrm{d}x \approx \frac{1}{n+1}\sum_{i=0}^{n}\left|\mathrm{CE}\left(\frac{i}{n}\right)\right|. \tag{4.35}$$

While well suited for a quick first check, the ECE has the disadvantage that we discard the sign information. Hence, it can only indicate that the calibration is off, but not whether the model is over- or under-confident.

5. Related Work

This section aims at giving an overview of current work within the scope of this thesis. As INNs are still rather niche, and a good portion of ML focuses on classification(-derived) tasks, while we focus on regression problems, it is hard to find work that is directly applicable to our setting. We will discuss the properties of the related work and what parts are better or worse suited for our applications.

Section 5.1 will introduce related work to uncertainty quantification. We will first talk about general, foundational work in ML and then give examples of how this work was applied in a medical setting. For this section, we interpret uncertainty quantification as any method of gauging the uncertainty of the actual prediction of the network, i. e. uncertainty quantification is an auxiliary task to gauge the reliability of the main task. This section is aimed at covering related work regarding research question T.1, which is concerned with representing ambiguous solutions to inverse problems for data-driven analysis. As this field is still rather small, we extended the scope of the related work section to include general uncertainty quantification methods. Section 5.2 will be concerned with related work regarding posterior validation. We will introduce commonly applied metrics and validation strategies of posteriors for inverse problems in different application fields. In particular, we will focus on how suitable the metrics are with regard to analyzing the well-posedness of inverse problems. This section covers related work with regard to research question T.2, which is concerned with the validation of (multimodal) posteriors. Section 5.3 will address related work regarding OoD detection. We will introduce work based on reconstruction algorithms and representation learning and how these approaches are applied in medical imaging. Furthermore, we will outline the current discussion about the scope of OoD detection and a lack of clear benchmarks to gauge progress in the field. This section covers related work with regard to research question T.3, which is concerned with OoD detection methods for biophotonic imaging. Section 5.4 contains a conclusion of the related work discussion and their impact on our research questions.

5.1. Uncertainty Quantification

Uncertainty quantification is an active field of research in ML in general, but also in medical imaging. In fact, “[...] medical AI, especially in its modern data-rich deep learning guise, needs to develop a principled and formal uncertainty quantification (UQ) discipline [...]” [BBK19]. Many tasks in medical imaging can be phrased as classification tasks or classification-derived tasks, like segmentation or object detection. This is most likely one of the reasons why uncertainty quantification in a medical context has mostly focused on these areas [Zho+21]. This circumstance has the downside that most uncertainty quantification work is hard to transfer to our tasks, as we are mostly concerned with inverse problems in the guise of regression tasks. Hence, we will start with general DL uncertainty quantification methods that can also be applied to regression tasks and afterward list uncertainty quantification approaches in medical imaging.

Uncertainty Quantification in General Machine Learning One of the first uncertainty quantification approaches for NNs was based on dropout [Sri+14]. While originally introduced to regularize the training and avoid overfitting, it was soon applied during inference time too. This practice took the name MC dropout [GG16]. Dropout deactivates some edges in the NN by setting the corresponding weights to 0. Which edges are chosen is based on a Bernoulli distribution with (dropout) probability p. The probability might be constant for all edges or change based on the layer or other environmental factors. For MC dropout, the NN operates multiple times on the same input but with a newly sampled dropout mask each time. This leads to multiple predictions, which are generally aggregated using the mean. The standard deviation can then be used as an uncertainty measure. The advantage of the approach is that it is very easy to implement, as most modern architectures include dropout. The disadvantage is that multiple passes are necessary for a prediction. Furthermore, it is hard to calibrate the uncertainty, i. e. since the loss term is unaware of dropout and the use as a distribution, there is no signal to make sure that the spread of the dropout distribution corresponds to a real physical quantity. Lastly, while dropout could, in principle, lead to multimodal posteriors, first results in MSI [Ard+18b] suggest that dropout tends to produce unimodal posteriors. Hence, dropout is a sub-optimal approach to analyzing the ambiguity of inverse problems.

A second widely spread approach to uncertainty quantification was introduced by Kendall and Gal [KG17]. This approach builds on the MC dropout approach and uses dropout as one factor to gauge the uncertainty. In addition, they suggest how common loss terms like the L2 loss or binary cross-entropy can be changed to include uncertainty. For the L2 loss, this is achieved via the observation that minimizing L2 is equivalent to maximizing the log-likelihood of a Gaussian distribution with unit variance. Hence, the loss term is updated to include a variance term, which is also dependent on the input. To achieve this, the network architecture has to be modified to double the outputs. The first half is interpreted as the original prediction[1], and the second half is interpreted as the variance[2]. The variance term, together with the MC dropout, is then used as an uncertainty score. As with MC dropout, this approach is minimally invasive in that it only needs a doubling of the output channels. The downside is the multiple passes necessary for the dropout and the parametric assumptions in constructing the uncertainty score. Indeed, the construction

[1] Or the mean of the Gaussian

[2] In practice, the log-variance is chosen due to numerical stability.

of the uncertainty score was based on an identification of the loss term with a parametric family of probability distributions. Hence, the approach is only applicable to cases where this identification is possible. As most parametric families are unimodal, this approach is again disqualified for ambiguous inverse problems, where we want to encode ambiguous solutions as multimodal distributions.

A classic approach to boost ML performance, even before the DL boom, was ensembling. However, next to increased predictive performance, ensembles can also be used to quantify uncertainty. As in the previous approaches, this is mostly achieved using the variance or standard deviation between the ensemble predictions as an uncertainty score. For NNs in computer vision this approach has been applied successfully, e. g. in [LPB17; SG18]. In these works, the same network architecture was chosen but trained multiple times, leading to different parameters for each ensemble member. In other works, the ensemble members do not share the same architecture [Tra+22; Yam+22]. Again, this approach's beauty is its simplicity, requiring no modification to the underlying model. However, this simplicity comes at the cost of increased computation time, as more models have to be evaluated. Furthermore, the ensemble size is generally restricted to single digits. Hence, working with the posterior generated by the ensemble does not make sense because we only have such a small sample size. Thus, we are restricted to the above-mentioned uncertainty scores, making this approach unsuitable for detecting ambiguities in inverse problems.

A last, more theoretic approach to uncertainty quantification is approximate Bayesian computation (ABC). These approaches build a posterior distribution, as do the cINNs, but they require the possibility of simulating new samples. Different flavors of ABC can be found in [Wil13]. The underlying idea is rejection sampling. A small ball around the input, spectra in our case, is chosen, then new data points (tissue parameters) are drawn based on the prior distribution[3]. The new samples are used to simulate spectra. If the spectra fall into the ball around the input spectrum, the generated samples are kept; otherwise, they are rejected. This process is repeated until a threshold of kept data points is reached. The kept data points make up a representation of the posterior. The clear advantage of this approach is the generation of a full posterior distribution, which is exactly what we want to detect ambiguities in inverse problems. However, we need to be able to simulate new data to create it, and we perform rejection sampling, which can be very inefficient depending on the prior distribution. Overall, this leads to a high computational burden, which can become too costly for high-dimensional data or intricate simulation pipelines. If simulation is impossible, we can try to apply ABC if our data set is large. In that case, instead of simulating new data, we choose all spectra of the training data set that are within the ball surrounding the input spectrum and build the posteriors using the associated tissue parameters. This approach will break down for high dimensional data, where the data set will be too sparse to find enough samples in a small ball.

[3] Or some clever modification of the prior distribution.

Uncertainty Quantification in Medical Imaging As mentioned previously, the uncertainty quantification work in medical imaging is mostly focused on classification(-derived) tasks [BBK19].

For segmentation tasks, labeling uncertainty has been a major focus. There is a strong inter-rater variability depending on how precise the annotation instructions are. It would be desirable for segmentation models to incorporate and visualize this uncertainty at inference time. This challenge has been addressed in two recent works [Koh+18; Mon+20]. Both are based on the U-Net architecture [RFB15], but modify it with additional latent spaces. At inference time, multiple samples are drawn from the latent space leading to multiple predicted segmentations. These can then be aggregated, and uncertainty scores can be derived. The approach has been adapted to include cINNs on the latent space level of the encoder-decoder architectures [Sel+20], but the sampling process to construct the posteriors does not include the cINNs. Instead, it only uses the latent space of the encoder-decoder architecture. While they were first developed for segmentation, these approaches would lend themselves well to be adapted to the ambiguous inverse problem setting, as they are based on multiple samples drawn from a latent space. However, the architectures are generally more involved than cINNs and require more loss-terms. This affects the amount of hyperparameter tuning necessary to acquire well-calibrated posterior distributions. In addition, the encoders and decoders would need to be adapted to the new setting. In particular, when we operate on single spectra compared to complete images as in the segmentation cases above.

In MSI, there is previous work on uncertainty-aware tissue classification [Moc+18]. In this work, tissue is classified using superpixels and a support vector machine (SVM) [CV95] classifier. Uncertainty is quantified using the entropy and the Gini coefficient [Gin36]. This approach allows for filtering uncertain classification results and thus increasing accuracy. This method could easily be adapted to modern NN classifiers. However, the regression regime makes the entropy and Gini coefficient harder to compute. As such, they do not adapt well to our ambiguous inverse problem setting. Overall, while MSI is a growing topic in medical imaging, there is little dedicated work towards uncertainty quantification. Another field where MSI is applied is remote sensing via satellite imagery. There, uncertainty quantification has been explored more thoroughly [SSS19; Son+21]. However, the main task seems to consist in classifying the ground in the image. Hence, the developed uncertainty quantification methods are particular to classification(-derived) tasks.

One uncertainty quantification branch in medical imaging, where regression problems are addressed, is PAI. A review of the current state of DL in PAI with a section about uncertainty quantification can be found in [Grö+21]. Gröhl et al. [Grö+18] employed a second model to estimate the prediction errors of a first U-Net, which was used to predict the oxygenation of tissue using photoacoustic initial pressure spectra. This second model was implemented as a U-Net, too, and received the initial pressure as well as the prediction network output as input. The embracing network predictions were then interpreted as an uncertainty score. In contrast to all other previously introduced methods, this approach introduces a single additional network and requires one extra evaluation. However, similar to the other approaches, we are left with a confidence score but not a complete probability distribution. While we can use this score to filter low-confidence predictions, we cannot detect inherent ambiguities in the inversion process. Model-based Bayesian approaches to uncertainty quantification in PAI have been developed in [Tar+13; TPT16; TPT19; Sah+20]. These models are built on strong error distribution assumptions, making it possible to invert the problem and arrive at a tractable posterior. The advantage is that these methods require far fewer data to fit and are less prone to overfitting. However, the modeling assumptions make it hard to detect ambiguities.

5.2. Posterior Validation

One key property of cINNs is their capability of generating posterior distributions. However, while a posterior promises access to more fine-grained information about the analyzed inverse problem (e. g. ambiguous solutions), exhausting that potential seems very hard.

In the original INN publication [Ard+18b], the width of the posterior was validated using the calibration curve as introduced in section 4.4.5. In addition, the MAP was determined and used as a point estimate for classical regression metrics like the root-mean-squared error. As the paper dealt with a simulation setting instead of validating the posteriors directly, the reconstructed parameters were used to re-simulate the conditioning input, and regression metrics were applied to this problem. Lastly, a qualitative comparison of the 1D marginals of the posteriors was performed comparing the INN posteriors to baseline posteriors.

This kind of analysis is a common trend when applying cINNs to inverse problems over a wide range of applications like high energy physics [Bie+21; But+21], astrophysics [Kso+20; Kan+22; NAB22], remote sensing [Mar+22], stochastic processes [HHS21], and epidemiology [Rad+20]. All these approaches have in common that the posteriors are mostly used to extract an uncertainty score based on a measure of variability like the standard deviation but without explicit mode detection and analysis of the well-posedness of the inverse problem.

In settings where a reference posterior can be generated, there are rare cases where the posteriors are not only compared using visual inspection but metrics to gauge the difference between the distributions. Mokrov et al. [Mok+21] use cINNs to learn to represent stochastic processes and use the symmetric Kullback-Leibler (KL) divergence [KL51] to gauge the fit of the posteriors. Haldemann et al. [Hal+22] uses the Hellinger distance [Hel09] to compute the distance between the cINN posteriors and posteriors generated using MC methods.

cINNs are also applied in computer vision, where they are used to conditionally generate images [Ard+19]. This is a very high-dimensional setting such that the sample drawn from the posterior is generally sparse. This complicates the posterior validation as e. g. clustering a sparse data set is hard. In addition, in this generative setting, the focus is often shifted.

The exact shape of the posterior is not so important, but the quality and diversity of the generated images are. This is reflected in the metrics like the structural similarity index measure [Wan+04], the peak signal-to-noise ratio, the Frechét inception distance [Heu+17], or the variance as in [Ard+19; Chá22].

Overall, we see that there is a lack in the prior art regarding the identification and validation of multimodal posteriors. However, reliable (multimodal) posteriors are central to our application in analyzing the well-posedness of inverse problems. Furthermore, there is not yet a community consensus on proper metrics for different settings or even a metric selection framework as the one for classification(-derived) tasks in [Mai+22b].

5.3. Out-of-Distribution Detection

As introduced in section 4.3.2, there is not yet a community-wide accepted definition of OoD detection. Hence, while there is a lot of prior work on OoD detection, the exact scope varies from manuscript to manuscript. In the following, we will introduce common patterns and how they relate to our setup.

One major branch of unsupervised OoD detection is based on reconstruction (cf. section 4.3.2). In the age of DL, the reconstruction algorithm often takes the form of a VAE as in e. g. [Zim+19] in the medical domain. The auto-encoder is trained on ID data with the assumption that the reconstruction loss for OoD data points should be higher than for ID data points after convergence. This vanilla approach can be modified by introducing suitable data augmentation during training or adapting the exact loss term used as OoD score (e. g. derivatives of the network could be taken into account). Another modification to this approach concerns the absence of OoD data during training. The unsupervised learning task can be transformed into a semi-supervised learning task by allowing a small portion of OoD data during training. The signal of the OoD data can help to better define the boundary of the ID data. This modification of the VAE approach was successfully applied in the medical domain in [Tia+20].

A second major branch is based on representation learning. The idea behind this approach is that the pixel-space is not a good representation for OoD detection, e. g. because of the manifold hypothesis, which implies that the images live on a low-dimensional submanifold of the pixel-space. Hence, an unsupervised representation learning model is trained to approximate a better representation of the data. Then classical OoD detection algorithms like k-NN, SVM, kernel density estimation (KDE), or GMM can be used on this lower dimensional representation [GAR19; Soh+21; KM22]. A widely-spread approach to representation learning in computer vision is *contrastive learning* [Che+20]. A central ingredient is access to a data augmentation scheme that preserves the semantics of the data. For example, let us assume the data set consists of images. Then one image is augmented twice, and it is assumed that both augmentations do not change the semantics of the image (e. g. both images still show a cat), so the network is trained to embed both instances close in the representation space. At the same time, any pair of images from the data set is assumed to represent different things, so they are pushed away from one another. This scheme has been very successful for images and, more precisely, image classification, where it is straightforward to construct randomized data augmentations that do not change the semantics. However, our tasks are often regression-based, and we commonly work on single spectra instead of whole images. This impacts the applicability of representation learning at two sites. First, the set of data augmentations is far less diverse. In fact, in this thesis, we use Gaussian noise augmentation almost exclusively. Second, it is far less obvious if our augmentations preserve semantics. If the noise level is chosen too high, the spectrum could potentially switch from a perfused to an ischemic spectrum. Hence, representation learning does not transfer easily to our setting. Nevertheless, this paradigm has been successfully applied in the medical imaging domain with the example of colonoscopy [Tia+21].

In [Nal+19a], a density-based approach to OoD detection which is task-agnostic was proposed. A model is trained to learn the density of the ID data distribution. However, instead of using the log-likelihood as an OoD score directly, they proposed a score based on the *typical set* of a distribution. While the mode is the point with the highest density, the region around this point (in high dimensions) generally has a low volume such that few data points are drawn from that region. In fact, if we look for the region from which we expect to draw most data points, we have to look at the mass of the distribution, which is, intuitively speaking, related to the product of the density and the volume. Roughly, this region with the highest mass is called the typical set. Its location can be determined empirically using the entropy of an ID validation set. For new data points, the absolute value of the entropy to the reference entropy is computed, which estimates the distance of the new data point from the typical set, and is used as an OoD score. A disadvantage of the approach is that it seems to require batches of new data points instead of working on single data points due to the entropy estimation.

Another difficulty in OoD detection is a lack of a clear definition of what OoD means. In [Win+20], this question is discussed and the notion of near and far OoD is proposed. The idea is that OoD is a spectrum and that it would be beneficial to quantify the "OoD-ness" of data points during the evaluation of OoD algorithms. For example, if the ID data consists of color images, then grayscale images are "far" off, which should simplify their detection as OoD. On the other hand, if the ID data contains cat images and the new data consists of lynxes, then the new data is "close", which might make it harder to detect. They propose the confusion log probability as a score to measure how far OoD new data is and build on contrastive learning as OoD detection methodology. While the general direction of quantifying OoD is very interesting, the exact methodology is tailored toward image classification tasks, and in particular, the confusion log probability assumes a discrete label distribution. This hinders direct application to regression tasks.

A further consequence due to the "soft" definition of OoD is that results are hard to reproduce and compare between tasks, data sets, and other settings. In fact, there are first results questioning whether any of the new DL-based approaches truly is a new state of the art [Taj+21]. Public benchmarks and challenges are often useful for a community to converge on an accepted definition and can simplify progress in the field. For the medical domain, the Medical Out-of-Distribution Analysis Challenge (MOOD) was introduced as part of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2020. The scope is to benchmark the performance of OoD detection and localization in CT and magnetic resonance imaging (MRI) of the brain and thorax. The community accepted the challenge, and MOOD has been a part of subsequent MICCAI conferences.

5.4. Conclusion

With regard to research question T.1, we see that while there is prior art toward uncertainty scores to gauge the reliability of a point estimate, there is little work encoding and analyzing solutions to inverse problems as posterior distributions. One exception is work based on ABC, but this is restricted to settings where (fast) simulation is available. Hence, we find that the proposed methodology based on cINNs validated with the example of biophotonic imaging addresses a gap in the literature.

Analogously, with regard to research question T.2, we find that while cINNs have recently been applied to inverse problems over a diverse set of domains, there is not yet a consensus regarding their proper validation. Generally, the width of the posteriors is validated using calibration curves, but afterward, the posteriors are aggregated by a measure of centrality (mean, median, or MAP) and a measure of variability (standard deviation, variance, or IQR). In particular, there is little work toward detecting and validating (multiple) modes of the posterior. Furthermore, there is not yet a discussion (even less agreement) on suitable metrics for a multimodal setting. This observation motivated our development of a posterior validation framework.

With regard to research question T.3, we see that OoD detection is an active field of research with many recent contributions. However, many of the computer vision-based OoD detection methods are not easily transferable to biophotonic imaging as we oftentimes have to work on single spectra (due to data sparsity and limitations in the simulation). Furthermore, OoD detection is often incorporated in the model addressing the main task, which is regularly classification-based. As physiological tissue parameter estimation is better described in terms of regression, we cannot accept this restriction. Hence, our contribution proposes a task-agnostic OoD detection methodology adapted to the needs of biophotonic imaging.

Part III.

Contributions

6. Framework for Uncertainty Handling in Biophotonics

This chapter introduces a framework for uncertainty handling in biophotonic imaging using INNs. We will address the methodological challenges introduced in the technical research question T.1 – T.3 in section 2. As such, the chapter is split into three sections, with each section focussing on one research question. An overview of the proposed framework can be found in figure 6.1.

In section 6.1, we will discuss how cINNs can be used to generate posteriors that represent solutions to inverse problems. In addition, we will explain how we can detect the modes of a posterior in an automated fashion, as this is central to judging the ill- or well-posedness of the inverse problems. The content of this section is the foundation on which the remainder of the posterior validation framework rests. The contributions in this section address research question T.1, which was concerned with how to suitably encode solutions to inverse problems. Section 6.2 will introduce a posterior validation framework to evaluate potentially multimodal posteriors in an inverse problem setting. We will showcase edge cases where classical metrics might fail, observe parallels between ambiguous inverse problems and object detection, and, finally, build on these observations to suggest the posterior validation framework. The contributions in this section address research question T.2, which was concerned with the validation of the generated posteriors. Section 6.3.1 presents WAIC, how INNs can be used to estimate it, and how we can use the criterion as an OoD score. In addition, we will elaborate on how OoD detectors can be used as personalized models to track physiological tissue parameter changes while circumventing confounders like inter-patient variability. The contributions in this section address research question T.3, which was concerned with identifying biophotonic imaging-specific OoD detection methods.

6.1. Representation of Inverse Problem Solutions as Posteriors

Figure 6.2 (a) shows a schematic view of the motivating scenario of our quest to represent solutions to inverse problems by posteriors. Light-tissue interaction is complex (as outlined in section 4.2). Hence, a priori, we do not know if markedly different tissue parameter configurations lead to similar or virtually identical spectra. This is indicated in the figure

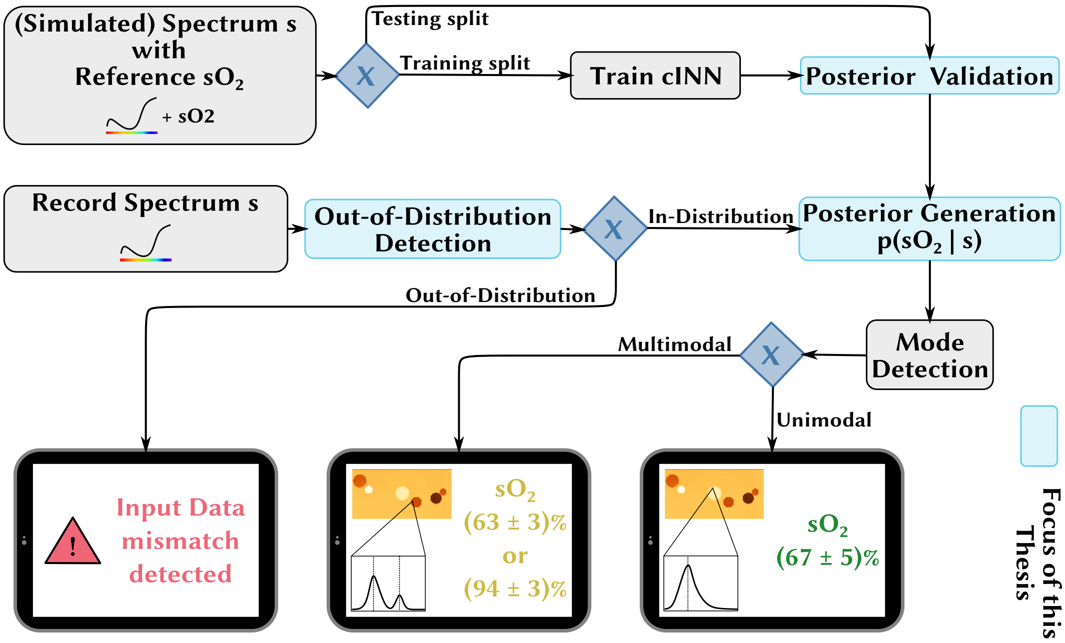


Figure 6.1.: **Framework for addressing uncertainty in machine learning systems for biophotonic imaging.** Annotated data (spectra + tissue parameters) is used to train and validate the posterior generating methods (cINNs in our case). Afterward, newly collected data (spectra) is filtered for out-of-distribution data on which prediction performance cannot be guaranteed. For the actual prediction task, we use the validated cINN, which is capable of generating full posteriors. These posteriors can then be used to analyze ambiguities in the inversion process. This thesis's focus is color-coded in turquoise. cINN: conditional invertible neural network, sO_2: blood oxygenation.

using the tissue parameter sO_2 as an example. A posterior representing the solution to the inverse problem could encode such a scenario as multiple modes. Figure 6.2 (b) shows a common setup of a cINN for solving such an inverse problem in order to analyze the existence of ambiguities.

This section will first introduce how cINNs have to be adapted to the biophotonic setting and how the posteriors are generated in detail (section 6.1.1). In particular, we will discuss issues due to the very low-dimensional nature of the tissue parameter space. By themselves, cINNs give us access to a density function to evaluate the posterior and a way to draw samples from the posterior. However, there is no intrinsic way to identify the modes of the posterior. Hence, we require post-processing tools to find them (section 6.1.2). Whereas the low-dimensionality caused challenges for the cINNs architecture, we will exploit it for mode detection, as there are algorithms that require a 1D setting.

Overall, this section will address research question T.1, which was concerned with how to suitably encode solutions to inverse problems. In particular, with encodings that allow the analysis of the well-posedness of the problem.

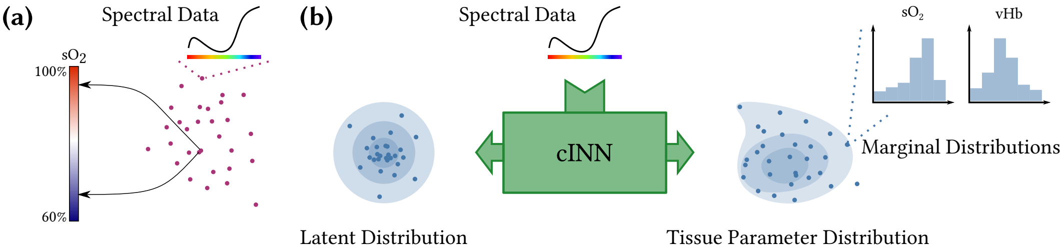


Figure 6.2.: **cINNs can be used to generate posterior distributions over tissue parameter space.** **(a)** Schematic view of an ill-posed sO_2 estimation task. **(b)** Spectral data (top) is used as conditioning input. The latent space (left) follows a Gaussian distribution. Drawing from this distribution, we can transform the latent samples into tissue parameter samples. This process builds up the posterior. For post-processing, we generally marginalize to the appropriate tissue parameter. cINN: conditional invertible neural network, sO_2: blood oxygenation, vHb: blood volume fraction.

Disclosure and Contributions Lena Maier-Hein proposed the usage of INNs to analyze biophotonic inverse problems, and she initiated the collaboration between our department and the Institute of Scientific Computing at Heidelberg University, where the architecture was developed. In particular, Lynton Ardizzone, Jakbo Kruse, Carsten Rother, and Ullrich Köthe developed the INN and cINN architecture and gave advice regarding implementation details. Lena Maier-Hein advised the whole experiment process and gave feedback throughout the method development. I adapted the cINNs architecture for use in MSI while Jan-Hinrich Nölke, in his master thesis under Lena Maier-Hein's and my supervision, adapted the architecture to PAI. Jan-Hinrich Nölke and I researched and discussed possible mode detection algorithms together. I (re-)implemented the mode detection algorithms used in this thesis.

6.1.1. Posterior Generation for cINNs

In section 4.4, we have described the construction and training process of cINNs in general. This short section will describe how INNs are applied to biophotonic imaging and how we can access the posterior distribution. Figure 6.2 (b) gives an overview of the cINN setup for tissue parameter estimation. As seen in equation 4.19, we can evaluate the density of the posterior thanks to the transformation rule. However, we generally do not work with this representation. Our aim is to detect the posterior modes, which are not directly visible from the density. We could perform gradient ascent, as the density is differentiable, but there is little prior work on good seeding locations and the number of seeds necessary to reliably localize all modes.

In addition to being discriminative models, i. e. giving access to the density, cINNs are also generative models. Hence, we can draw samples following the posterior distribution. After the training has converged, the latent space Z follows a standard Gaussian distribution $\mathcal{N}(0, \mathrm{id})$. To construct the posterior for a conditioning input c, we sample $z \sim \mathcal{N}(0, \mathrm{id})$ and compute

$$x := f_\Theta^{-1}(z, c),$$

where f_Θ denotes the cINN. This approach guarantees $x \sim p(x \mid c)$.

Repeating the process with multiple z leads to an arbitrarily large sample following $p(x \mid c)$[1] and mode detection transforms to the task of clustering as introduced in the next section. The cost is an additional hyperparameter, namely the sample size of the posterior.

Next to these theoretic considerations, there are practical implications we need to consider when working with the posterior. First, we have to de-normalize generated samples x. As all training data is z-score normalized, so is the generated posterior. However, we need to revert back to the original scale and units for interpretation purposes. This requires access to the mean and standard deviation of the training data, which need to stay accessible at inference time.

Additional post-processing steps might be necessary if the data is very low dimensional or very high dimensional. As introduced previously, the cINN architecture demands at least two input dimensions. If we work with only a single parameter of interest, we need to perform padding at training time. This could be achieved via doubling the parameter of interest or by adding a noise dimension. In both cases, the samples need to be marginalized. Additionally, in the first case, it is advisable to check the consistency of the doubled parameter dimensions. In the case of large dimensional data, like images, we run into the

[1] Assuming the training was successful.

curse of dimensionality in the shape of sparse posteriors, i. e. it is hard to draw a large enough sample to capture the shape of the posterior. This complicates the mode detection process. In this thesis, we will circumvent this problem by working on smaller regions of interest or even single pixels, i. e. we interpret and evaluate the marginalized posteriors of each pixel separately. This approach has the drawback of discarding all correlation information, but it is necessary to enable stable mode detection at all.

6.1.2. Mode Detection for cINNs

In section 6.2, we will see how we can select metrics to evaluate cINNs applied to inverse problems. A central assumption of the metric selection framework will be that the predicted posteriors have labeled modes. However, a cINN by itself only produces a posterior without any mode labeling. Hence, mode detection needs to be addressed as a post-processing step.

While we can explicitly evaluate the posterior's density, it is easier to simply draw samples and work with this representation. In this representation, mode detection is equivalent to clustering the data points. This is an unsupervised learning task as introduced in section 4.3.1 and there exist many off-the-shelf algorithms to tackle it (e. g. mean shift clustering, EM with GMMs, density-based spatial clustering of applications with noise (DBSCAN), etc.). For the inverse problems in this thesis, we do not know in advance how many modes to expect. Hence, we require our clustering algorithms to automatically detect the number of modes. This decreases the number of admissible algorithms. Conversely, the problems of this thesis are either one- or two-dimensional, and even in the 2D case, we can work on the 1D marginal distributions. This special structure of our problem gives us access to clustering algorithms that do not generalize well to higher dimensions, and we will restrict our focus to two of them.

The first is based on kernel density estimation (KDE). For 1D data, KDE can give a very good representation of the density, and we can easily evaluate it on a grid. This leads to a 1D array of values, and detecting modes reduces to finding local maxima in this array. Another advantage in our setting is that we know the support of the distribution in advance because the tissue parameters sO_2 and vHb are restricted to the interval $[0, 1]$. So we can be sure to evaluate the density in the correct region. The only hyperparameter left is the bandwidth (bw) of the kernels. If it is too large, the KDE will smooth out the structure of the posterior, if it is too small, the KDE will fit noise, and we will detect spurious modes. For this work, we decided to use Silverman's rule of thumb [Sil18] which estimates the bandwidth based on the empirical standard deviation $\hat{\sigma}$, the interquartile range (IQR), and the data set size n via

$$\mathrm{bw} = 0.9 \cdot \min\left(\hat{\sigma}, \frac{\mathrm{IQR}}{1.34}\right) \cdot n^{-\frac{1}{5}}. \tag{6.1}$$

With this method, we can detect any number of modes, as long as they are separated by roughly the bandwidth parameter.

As Silverman's rule of thumb is only a heuristic, we decided to employ a second scheme to decide whether a posterior is multimodal or not. Again the central ingredient is that we have a very good understanding of the plausible range of our parameters of interest (sO_2 and vHb) and even more on a lower bound on the required resolution of our mode detection method. That means that it is irrelevant from an application perspective if two modes closer together than e. g. 1 pp are resolved as separate or not. Based on this knowledge, we can create a bandwidth range, and for each bandwidth in that range, we check whether the corresponding KDE is multimodal or not. Afterward, we compute the fraction of multimodal KDEs over the bandwidth range, and we call this value the *KDE score* of the posterior. The score ranges in the interval $[0, 1]$, and higher values indicate that the posterior is multimodal. While this method does not produce the exact mode locations, it is interesting to scan a whole data set for the fraction of multimodal posteriors while reducing the bandwidth dependence of the method.

In addition to this handcrafted multimodality score, Hartigan's diptest [HH85] can be used to detect multimodal posteriors, which was brought to my attention by Jan-Hinrich Nölke [Nöl21a]. This method is again restricted to 1D distributions, where the empirical cumulative distribution function is compared to the closest unimodal cumulative distribution function. This difference can be used as a test statistic, such that we can compute the fraction of multimodal posteriors of a data set for a given α-level. For implementation details, we refer to the original publication [HH85].

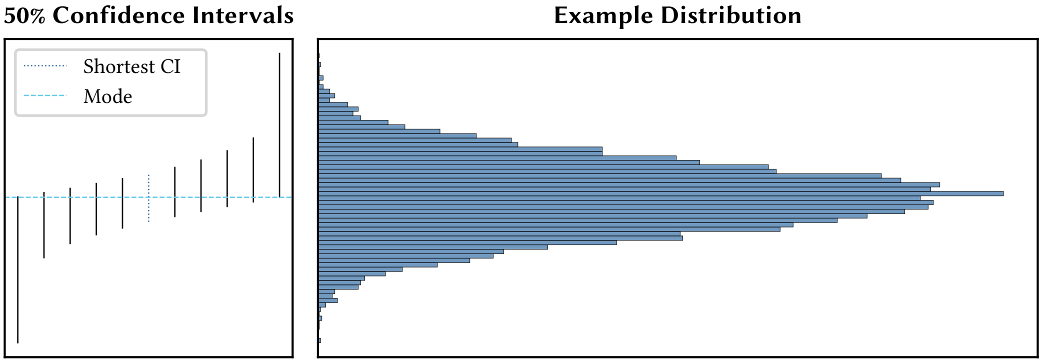


Figure 6.3.: **Heuristic of the half-sample mode (HSM) method.** The shortest confidence interval (CI) should be located around the highest mode of the distribution. Iteratively choosing the shortest 50 % confidence interval within the previous confidence interval can be used to localize the (largest) mode.

As a second method to detect the mode location of our distribution, we use the half-sample mode (HSM) method [BF06]. The intuition behind this method is depicted in figure 6.3. For 1D distributions, it is easy to construct a confidence interval to a certain confidence level. However, by itself, there is no CI for a given level. A common choice to get a unique CI is to choose the symmetric interval around the median. However, there are other choices. A practical alternative for mode detection is to choose the shortest confidence interval for the chosen confidence level. This CI should be located at the mode. The HSM method uses this observation iteratively. We start with the shortest 50 % CI, then locate the shortest 50 % CI within the first and continue in this fashion until convergence. The advantage of this method is that there are no hyperparameters to set, and it is very robust with regard to noise. The disadvantage is that we can only detect the largest mode, and as such, it is ill-equipped to detect multimodal posteriors.

In the unimodal case, we can often get away with simply using the median as a mode detector because the strength of the above-mentioned methods only comes into play if the distribution is heavily skewed. As the median is faster to compute than the actual mode, we will resort to it if we know that the inverse problem is sufficiently well-posed.

We will apply and compare the performance of these mode detection algorithms in our experiments analyzing biophotonic inverse problems. Section 7.1 will discuss the well-posedness of the estimation of sO_2 and vHb in MSI, while section 7.2 will discuss the well-posedness of the estimation of sO_2 in PAI.

6.2. Posterior Validation Framework

The biggest advantage of cINNs for inverse problems is, at the same time, their biggest problem: What are suitable metrics to evaluate a posterior distribution against a reference value? This problem is by no means trivial. How should we handle a multimodal posterior if we only have a single reference value? Is the second mode a false positive?

In section 6.2.1, we will highlight such caveats. In section 6.2.2, we will explore an analogy between the task of object detection and the task of posterior validation. In section 6.2.3, we will introduce a framework that aims at structuring and simplifying the process of choosing suitable metrics when validating the results of cINNs.

The contributions of this section will address research question T.2., which was concerned with the validation of the generated posteriors.

Disclosure and Contributions This section is heavily inspired by the metrics reloaded framework introduced in [Mai+22b]. Lena Maier-Hein noticed the object detection analogy and suggested adapting the framework to the inverse problems setting. I elaborated on the analogy, adapted the framework, and collected feedback from our collaborators. In this vein, I would like to thank and mention the valuable feedback from and discussion with Lena Maier-Hein, Jan-Hinrich Nölke, Lynton Ardizzone, Sebastian Gruber, and Ullrich Köthe.

6.2.1. Validation Pitfalls

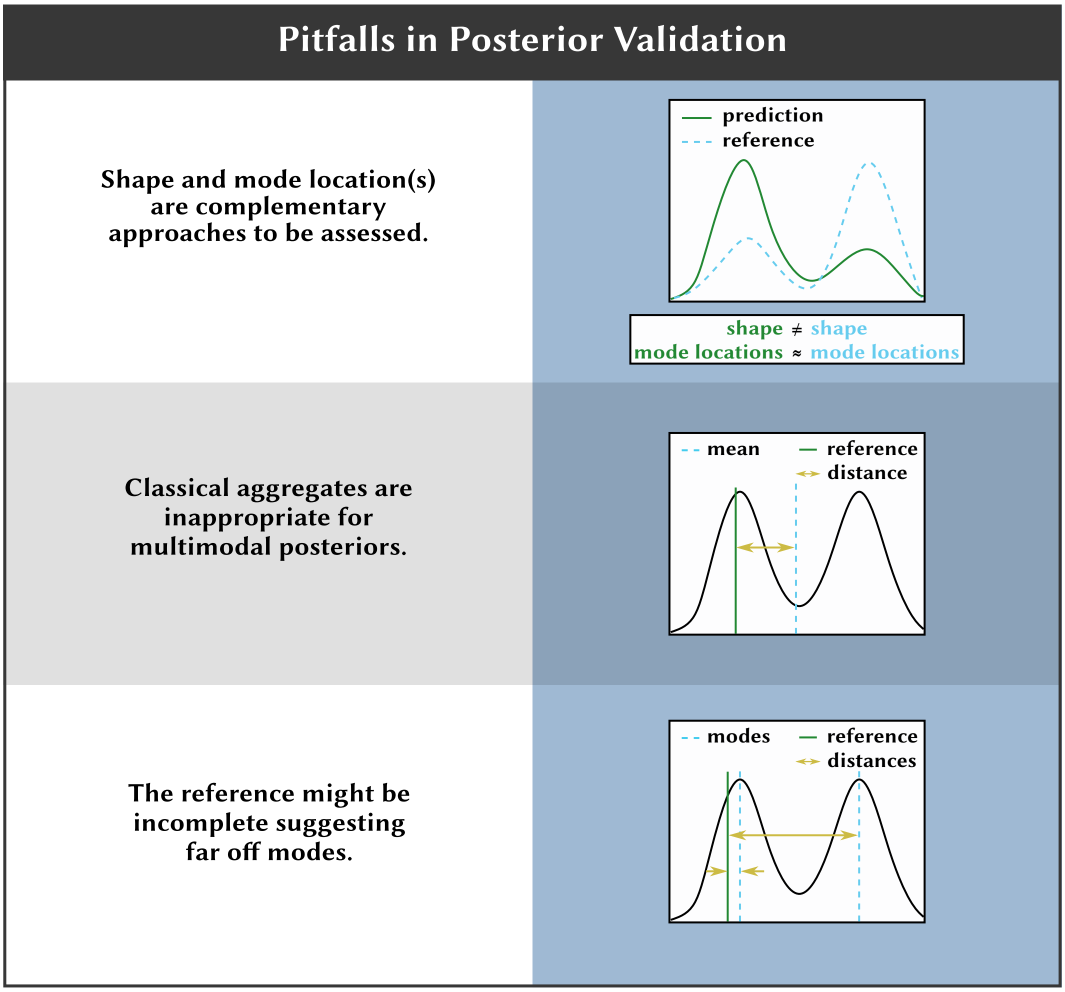
We can easily adapt most common regression metrics if we know that the inverse problem is well-posed. This is because the posterior is unimodal in these cases. Hence, most summary statistics work well to capture the shape of the distribution, and we can apply the regression metrics to these summary statistics. However, oftentimes we do not know or do not want to assume that the inverse problem is well-posed. This means we have to handle the possibility of multimodal posteriors. In this section, we will highlight pitfalls related to the validation of posteriors that we should keep in mind when developing metrics to assess the performance of our cINNs. Figure 6.4 contains three example pitfalls, which we will discuss in more detail. Naturally, this exposition cannot be exhaustive.

Mode Location vs. Shape (first row): Our model produces a posterior distribution, and let us assume that our reference is a distribution, too. From a mathematical point of view, we would aim for our posterior to follow the reference shape as closely as possible. Most metrics or divergences that operate on distributions would encourage and measure this. However, from an application point of view, the exact shape might not be important; instead, precise localization of the modes might suffice. An example of this scenario might be when the mode locations are intended as initialization for an iterative solver, where better initial localizations are important, but the shape of the posteriors might not be incorporated. If we are in this scenario, the chosen metric should reflect it. Instead of comparing the whole distribution, one could only consider the mode location and compare the distance between the predicted and the reference mode.

Classic Summary Statistics (second row): Many classic summary statistics have caveats, e. g. the mean and the standard deviation are susceptible to outliers. One caveat that is true for most of them is that they handle strongly multimodal distributions poorly. Let us assume a bimodal distribution with two well-separated modes with equal mass, as indicated in figure 6.4, second row. Then the mean as well as the median[2] fall into low-density regions. Assuming further that the distribution correctly captures an ambiguity in an inverse problem, this also implies that the mean or median is far away from the reference. Hence, any metric using the mean or median as a predictor would falsely rank the model's performance very low, even though it perfectly captured the ambiguity.

One solution to this problem is mode detection. If we can locate all the modes, we can handle each as a separate instance and compute metrics on each mode separately. This approach leads to an interesting analogy to object detection in computer vision, which will be explored in section 6.2.2.

[2] If applied component-wise in the multivariate case.



Pitfalls in Posterior Validation

Shape and mode location(s) are complementary approaches to be assessed.

Classical aggregates are inappropriate for multimodal posteriors.

The reference might be incomplete suggesting far off modes.

Figure 6.4.: **Metric edge cases due to (multimodal) posteriors. First row:** Depending on the application, it might be more important to identify the mode location than the exact posterior shape. This should be reflected in the metric. **Second row:** For downstream tasks, it might be helpful to compute summary statistics of the posterior, but classical summary statistics are often ill-equipped for multimodal posteriors. **Third row:** Commonly, inverse problem data sets consist of data points with exactly one solution to the inverse problem given an observable, even though the solution is ambiguous. If the model predicts a mode at a location without reference, it is hard to judge whether it is due to an incomplete reference or an error of the model.

Incomplete Reference (third row): There is a common pattern in how inverse problem data sets are constructed. The forward model is often accessible with manageable effort, either through simulation or experimental design. Hence, some set of parameter configurations is constructed, and the forward model is applied, leading to the observables. This construction leads to a data set where each observable has exactly one corresponding parameter configuration, even if the inverse problem is highly ambiguous. If we are lucky, there is some other data point in the data set where we chose a parameter configuration leading to the same observable so that we could hope to spot the ambiguity, but in general, we cannot assume that.

Still, our model might pick up on the ambiguity and produce a multimodal posterior through interpolation. At evaluation time, we are faced with the problem that our data point has exactly one reference parameter configuration, which should coincide with one of our modes. This leaves us with at least one other mode with no associated reference. Naively, we could compute the distance between this second mode and the one reference we have, but that would artificially downplay the model's performance. Still, for some applications, a falsely-detected mode (i. e. a false positive) might come at high costs, so it would be great if we could determine the existence or non-existence of a fitting reference value.

In some settings, this is possible. For example, if the data set is simulation-based, we can use the predicted parameters and re-simulate the observables. If they are close to the original, this validates the second mode. Alternatively, if the data set is large enough, we can try to use a nearest neighbor approach on the observables[3]. We choose an observable, allow for small deviations, and build a reference posterior by including all parameter configurations whose observable fall within the deviation ball. The re-simulation and the neighborhood approach can be interpreted as ABC.

In cases where we cannot identify all reference modes, we have to make do with a reduced set of accessible metrics. Which metrics these are will be explored in the posterior validation framework in section 6.2.3, which will cover this edge case.

6.2.2. Object Detection Analogy

In the previous section, we have seen that there are peculiarities to validating posteriors. Most of them are concerned with multimodal posteriors. This is exactly the case where we leave the classical regression regime, where there is exactly a single prediction which we then compare to the reference using a suitable metric. If there is more than one mode, each one has to be compared to the reference in some way. As seen previously, this can lead to difficulties if the number of modes and the number of reference values do not match up. While this is uncommon for regression tasks, there are other branches of ML where we encounter similar problems, namely object detection.

[3] This is a more formalized version of the 'lucky approach' mentioned at the beginning of this section.

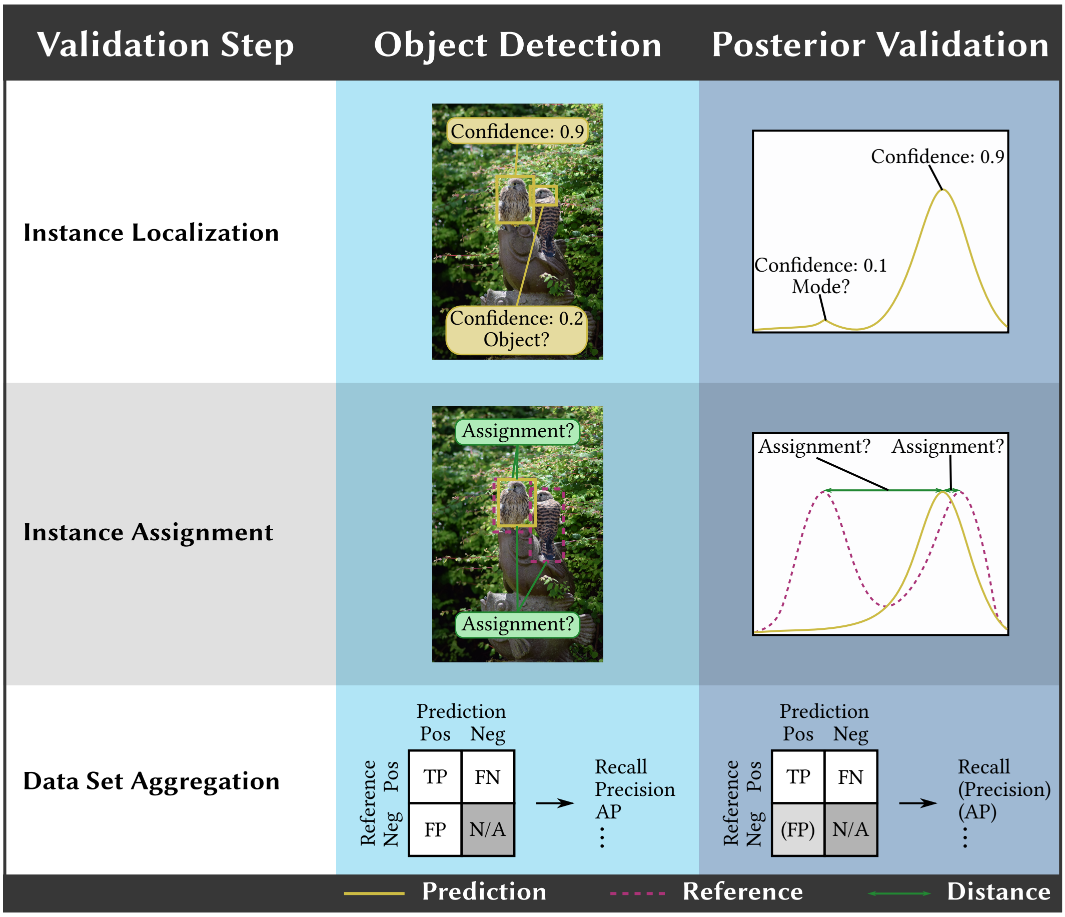


Figure 6.5.: **Analogy between object detection and posterior validation.** Each row depicts a step in the validation process. **First row:** In the localization step, predicted instances are filtered based on a confidence score, and distances to reference instances are computed. This corresponds to filtering low-mass modes and computing distances to reference modes for posterior validation. **Second row:** In the assignment step, we match predicted to reference instances using a matching algorithm. This corresponds to matching predicted modes to reference modes for posterior validation. **Third row:** In the data set aggregation step, we aggregate the per instance metrics via the confusion matrix and compute counting metrics. This can be transferred one-to-one to the posterior validation setting. In both cases, true negatives are undefined. Depending on the reference completeness, false positives might be undefined in the posterior validation case. AP: average precision, FN: false negative, FP: false positive, TP: true positive.

An image can easily contain multiple instances of the same object (e. g. the falcons in figure 6.5, middle column). Similarly, detection networks produce multiple predictions (in our example, bounding boxes). In the ambiguous inverse problems setting, the predicted instances correspond to the modes of the posterior. Hence, we can think of the reference values as the reference instances and the modes of the posterior as the predicted instances. The analogy is depicted in figure 6.5.

We will use three object detection validation steps to highlight the similarities between object detection and posterior validation. In [Mai+22b], there is a fourth validation step which is concerned with calibration, but cINNs have their own calibration methodology (cf. section 4.4.5) which is inconsistent with object detection calibration. While harmonizing the two calibration notions is future work, this does not negatively impact the analogy between the other validation steps. The first step (cf. figure 6.5, first row) concerns instance localization. Generally, object detectors produce multiple instances with a confidence score. This confidence score is then used to decide which candidates are kept for further evaluation. For the kept candidates, suitable distances (like intersection-over-union) to the reference instances are computed. In the posterior validation setting, the instances correspond to modes, and we can use e. g. the relative mass of each mode as a confidence score to decide which candidates we would like to keep. Then we can compute distances to the reference mode(s). The second step (cf. figure 6.5, second row) is instance assignment. We have the candidate instances and distances to the reference. In this step, we apply matching algorithms (e. g. the Hungarian algorithm) to find an optimal matching between predictions and references. This step translates one-to-one to the mode validation setting, where we need to assign the predicted modes to the reference mode(s). The third step (cf. figure 6.5, third row) concerns aggregation. The first two steps operate on each sample individually. For object detection, the matched instances are aggregated using a confusion matrix with the peculiarity that there are no true negatives (TNs). Then counting metrics, like recall, precision, or average precision (AP), can be computed. This step translates once more to the posterior validation setting, with the restriction that depending on the completeness of the reference (cf. figure 6.4, third row) FPs might not be defined either. This might restrict the set of available counting metrics.

We see that the analogy between object detection and posterior validation is quite extensive. This observation sparked the development of a metric recommendation framework based on [Mai+22b]. This framework is introduced in the next section.

6.2.3. Metric Recommendation

As we have seen in the previous two sections, some pitfalls can influence our choice of metric, and depending on the structure of our inverse problem, we might hope to borrow metrics from the object detection community. However, there are more important reasons for a posterior validation framework. First, such a framework can be a starting point for a discussion between communities, domains, and modalities. Instead of reinventing the wheel for each manuscript or an opaque inheritance of metrics from some prior work, we can discuss the framework's proposed branching points and refine them. Newly developed metrics can be incorporated at a central point instead of being proposed in a single manuscript which might or might not get the attention of all potential users. Second, it can increase comparability between contributions. If the framework reaches acceptance within the communities and converges to some stable recommendations, benchmarking of new methods becomes far more transparent. In this section, we would like to propose such a posterior validation framework to begin a fruitful discussion. We will formalize key properties that an inverse problem[4] might have and use this *problem blueprint* to choose the optimal set of metrics for the current situation.

Figure 6.6 contains the top-level overview of the framework. It can be separated into four building blocks: the problem blueprint, distribution-based metric selection, object detection-inspired metric selection, and data set aggregation. Each block may contain sub-processes, denoted by S1 to S7, which will be discussed in the following paragraphs. Each branching point references one of the key problem properties, which make up the problem blueprint. As the problem blueprint takes such an important role in the framework, the first sub-process (S1) is building said blueprint.

Problem Blueprint A pictorial representation of the blueprint can be found in figure 6.7. We have categorized the defining properties into three groups. First, there are reference-related properties. The granularity refers to the question of whether the reference is a probability distribution or if the reference is a discrete set[5] of mode coordinates, i. e. we only have a discrete set of solutions to the inverse problem in the reference but not a distribution. The mode labels property only comes into play if the reference is given by a distribution. The question is whether all modes of the distribution are labeled or not. For a discrete set of modes, this is automatically true. The last property in this group only comes into play for a discrete set of modes, where we need to check whether our set of modes is exhaustive. This goes back to the third edge case discussed in section 6.2.1.

The second property group relates to the representation of the prediction. Analytical distribution asks whether the posteriors are given by an explicit density function, which we can evaluate, or whether the distribution is only implicitly accessible through sampling. The dimensionality property asks whether the whole problem is 1D because the 1D case has access to a wider range of metrics and divergences. The binning property is about whether it is possible to discretize the posterior. Oftentimes there is a natural bin size available. If this is the case, some more metrics and distances become available. Lastly, the confidence property is about the existence of a confidence score for each predicted mode, i. e. whether the model can provide a score of how certain it is about each mode in its prediction. In most cases, this score should exist, as we can use the mass of each mode relative to the

[4] And the model solving it.

[5] Maybe only consisting of a single member.

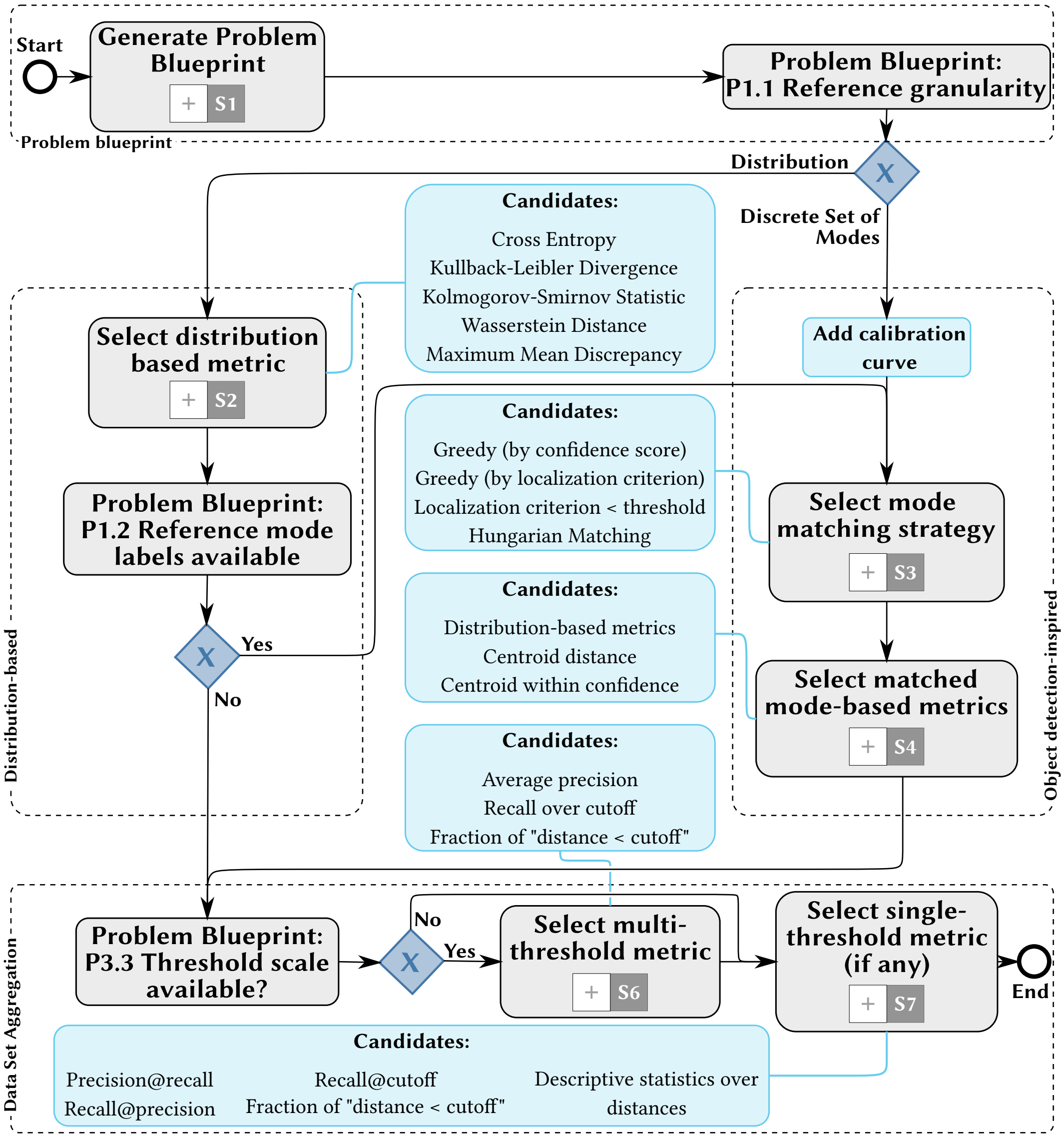

Start
Generate Problem Blueprint
S1
Problem Blueprint: P1.1 Reference granularity
Problem blueprint
Distribution
X
Discrete Set of Modes
Candidates:
Cross Entropy
Kullback-Leibler Divergence
Kolmogorov-Smirnov Statistic
Wasserstein Distance
Maximum Mean Discrepancy
Select distribution based metric
S2
Add calibration curve
Candidates:
Greedy (by confidence score)
Greedy (by localization criterion)
Localization criterion < threshold
Hungarian Matching
Select mode matching strategy
S3
Problem Blueprint: P1.2 Reference mode labels available
Candidates:
Distribution-based metrics
Centroid distance
Centroid within confidence
Select matched mode-based metrics
S4
Yes
X
No
Distribution-based
Object detection-inspired
Candidates:
Average precision
Recall over cutoff
Fraction of "distance < cutoff"
Problem Blueprint: P3.3 Threshold scale available?
X
No
Yes
Select multi-threshold metric
S6
Select single-threshold metric (if any)
S7
End
Data Set Aggregation
Candidates:
Precision@recall
Recall@precision
Recall@cutoff
Fraction of "distance < cutoff"
Descriptive statistics over distances

Figure 6.6.: **Top level view of the posterior validation framework.** The framework consists of four main parts. First, a problem blueprint is generated. This blueprint identifies key properties of the problem that allow or preclude the use of certain metrics. The metrics are selected via distribution properties or via the object detection analogy. The final part is the data set aggregation, where we aggregate the per-sample metrics over the whole data set. Sx with $x = 1, \ldots, 7$ are sub-processes that describe the current selection stage. Each sub-process has a corresponding figure in this section.

total mass (which is 1) as the confidence score. However, if this approach is infeasible, one might be in a situation without a confidence score, which has to be handled separately.

The last property group is about domain-related properties. This group is about the application and how it might influence our metric choices. The first two properties are related and aim to learn whether accurate predictions and/or accurate uncertainty are important for the task. For example, accurate prediction might be the driving force, but we can cope with uncalibrated uncertainty (e. g. over-confident predictions). On the other hand, it might acceptable to trade some prediction accuracy as long as the model knows what it does not know, i. e. the uncertainty estimation is very accurate. The third property asks for the existence of a threshold scale. This is a rather technical property relating to whether we can use threshold-independent metrics like AP. To be able to use it, we need an application-motivated range of thresholds over which to compute the aggregated metric. If we do not have it, we are restricted to single threshold or threshold-independent metrics. The next property is about double assignments. If the reference distribution is unimodal and very wide, but the predicted posterior is bimodal with both modes in the support of the reference, we could potentially assign either one of them to the single reference mode. We need to decide how to handle the second mode for this property. Depending on our application, assigning both modes to the reference mode might be adequate. In another setting, the surplus mode should be penalized because e. g. the correct estimation of the number of modes is central. The last property concerns the availability of a predefined cutoff. If a natural distance scale suggests a cutoff, we can use counting metrics to gauge the quality of our predictions. Otherwise, we will only be able to use threshold-independent metrics, like averaging the distances over a test data set.

Metric Selection With the complete problem blueprint, we can start walking through the posterior validation framework (cf. figure 6.6). The first branching point is about the reference granularity. Our prediction is always a posterior, but the reference might not be. If it is a distribution, we can enter the distribution-based metric block of the framework. If not, we jump directly to the object detection-inspired metric block. In this way, we can address commonalities and differences between the object detection and the posterior validation setting. Within the distribution-based metric block, we select a suitable metric to compare the reference and the predicted posterior (sub-process S2). Afterward, we might enter the object detection-inspired block. However, this requires that the reference distribution has its modes labeled. Otherwise, we directly enter the aggregation block.

If we enter the object detection block because the reference consists of a discrete set of modes, we need to add a calibration curve and calibration errors to our list of metrics. The calibration curve and associated metrics were introduced in section 4.4.5. This is necessary if we want to use measures of variability like standard deviation or IQR of our posterior for uncertainty quantification. Otherwise, the width of the posterior might not reflect the actual uncertainty. If the reference is also a distribution, a calibration is unnecessary

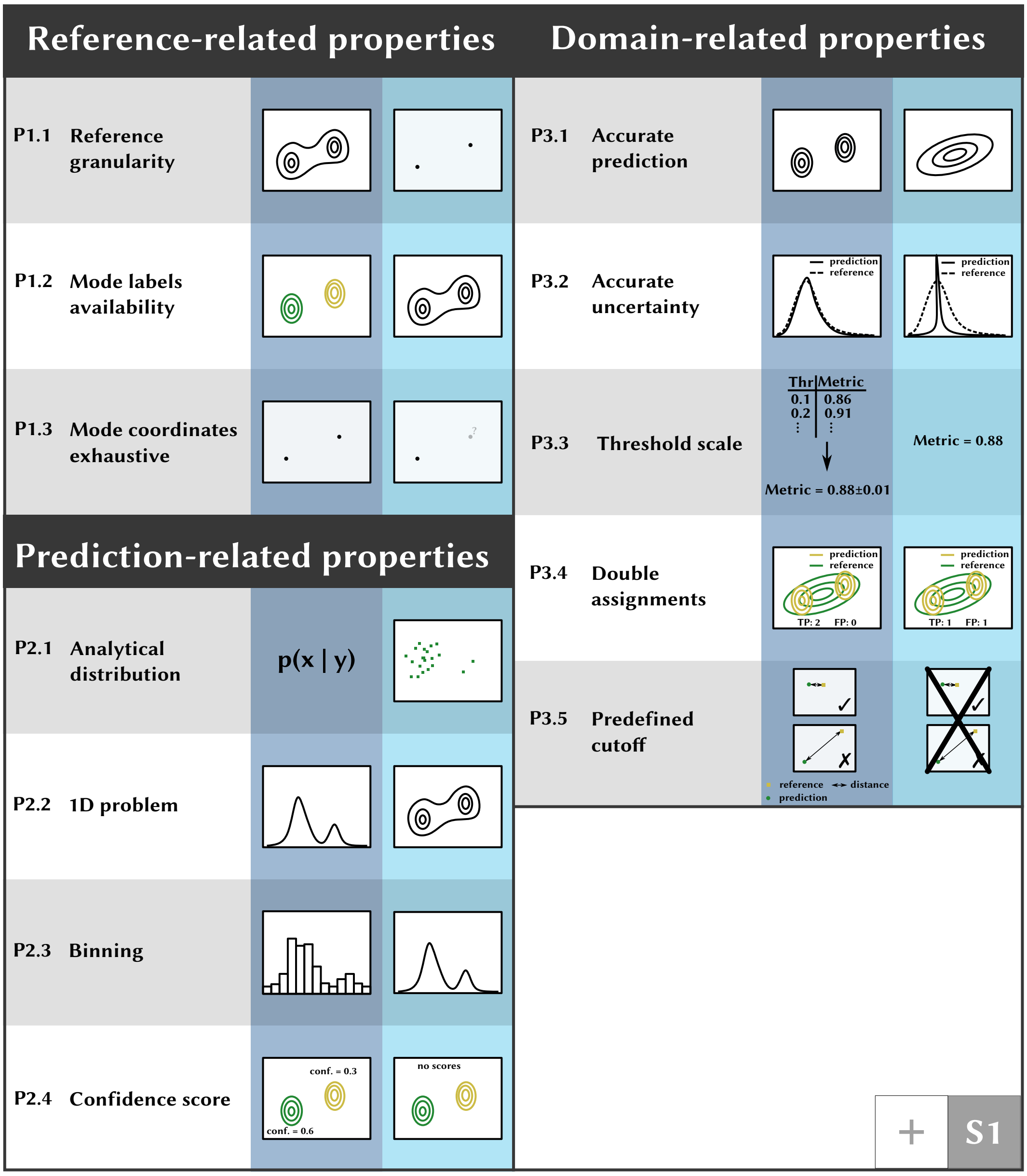


Figure 6.7.: **Pictorial representation of the problem blueprint for inverse problems.** The key properties of the problem are grouped into three categories: Properties relating to the representation of the reference, properties relating to the representation of the prediction, and domain-specific information. An in-depth description of all properties can be found in the text of section 6.2.3.

because the distribution-based metrics can cover the shape of the posterior, too, if so desired. Assuming that both the predicted and the reference modes are labeled, we can use the object detection analogy introduced in section 6.2.2. To this end, we need to match the predicted modes to the reference modes. Sub-process S3 will guide us through the process of selecting suitable metrics and algorithms. After all the modes have been assigned, we can use metrics to gauge the match of the mode pairs. These might be as simple as measuring the distance between the centroids of the modes or more involved distribution-based metrics. Sub-process S4 will guide us through this selection.

The data set aggregation block consists of two sub-processes. If a threshold scale is available, we can choose a multi-threshold metric as described in sub-process S6. Otherwise, we can only use single-threshold or threshold-independent metrics as found in sub-process S7. After all these steps, we have generated a comprehensive set of metrics to thoroughly evaluate our model for ambiguous inverse problems. The following paragraphs will explain the sub-processes in more detail.

Distribution-based Metrics If both the reference and the prediction are a (non-degenerate) distribution, we can use "distances" defined on distributions to measure the quality of the fit. Please note that we can interpret single modes as distributions in their own right and apply the distances, metrics, and divergences in this section to them too. Naturally, only if all additional assumptions of the distance are satisfied.

If the density of the prediction is explicitly computable, we can compute the (negative-)log-likelihood of the reference distribution under this density, i. e. we compute the cross entropy of the predicted distribution q relative to the reference distribution p:

$$H(p, q) \coloneqq -\mathbb{E}_p[\log q] \approx -\frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \log q(x \mid y), \tag{6.2}$$

where $\mathcal{D}$ is the data set representing the sample drawn from the reference distribution p. The advantage of this approach is that we can avoid sampling from the predicted distribution q, which would introduce errors due to the finite sample size. On the other hand, the scale of the cross entropy is hard to judge absolutely. For example, the cross entropy H takes its minimum for $p = q$, and at this point, it is equal to the entropy of p, but since the reference distribution generally is only available implicitly as a sample, this value is hard to compute. Hence, the cross entropy is a nice relative measure, but the absolute scale might be difficult to interpret.

If the application domain allows for suitable binning (e. g. because we know a certain resolution of change is sufficient, like detecting 1 pp sO$_2$ change might be enough), we can discretize both the reference and the predicted distribution. Hence, they are given by sequences $(p_i)_i$ and $(q_i)_i$ respectively. In that case, the KL divergence is given by a sum[6],

[6] Or a series if the sequences are infinite.

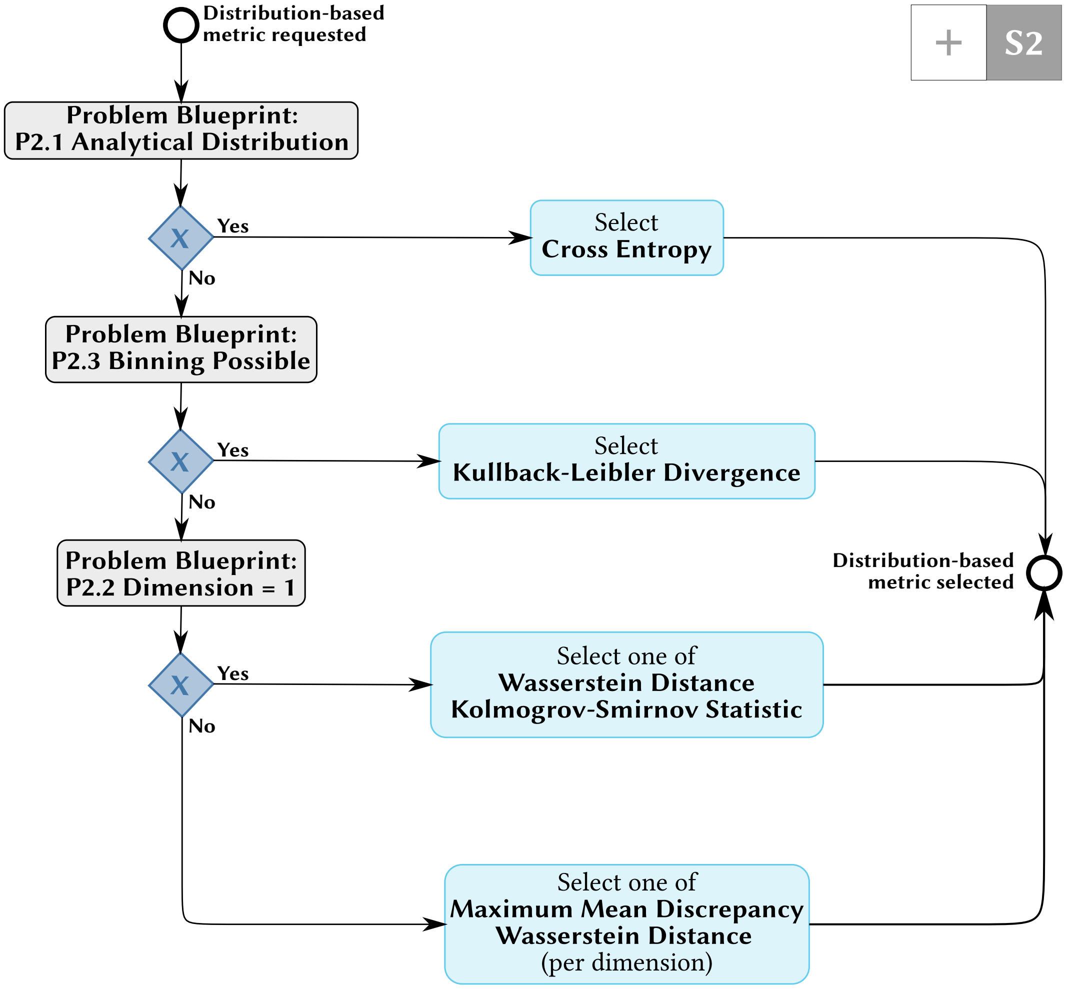

Distribution-based metric requested
+
S2
Problem Blueprint: P2.1 Analytical Distribution
X
Yes
No
Select Cross Entropy
Problem Blueprint: P2.3 Binning Possible
X
Yes
No
Select Kullback-Leibler Divergence
Problem Blueprint: P2.2 Dimension = 1
Distribution-based metric selected
X
Yes
No
Select one of Wasserstein Distance Kolmogrov-Smirnov Statistic
Select one of Maximum Mean Discrepancy Wasserstein Distance (per dimension)

Figure 6.8.: **Sub-process to select metrics or other distances to compare probability distributions.** Depending on the exact representation of the distributions and additional properties of the data, only some metrics or distances are available.

which can be computed as

$$D_{\mathrm{KL}}(p,q) \approx \sum_i p_i \log\left(\frac{p_i}{q_i}\right). \tag{6.3}$$

In the continuous case, it is hard to compute the KL divergence because an integral would need to be approximated and p is generally only available implicitly. The advantage of the KL divergence is that its scale is clearer than the cross entropy's. D_{KL} is non-negative and only equal to zero if $p = q$. However, it requires binning, which introduces discretization errors. Furthermore, binning is very susceptible to the curse of dimensionality, so we should expect the usefulness of the KL divergence to deteriorate quickly for higher dimensional distributions p and q.

If we are in the very special case, where the distributions p and q are 1D, we can access two distances. The first is based on the Kolmogorov-Smirnov (KS) test [Kol33; Smi48], which tests if two 1D distributions significantly differ from one another. The test statistic is based on the cumulative distribution functions F_p and F_q and is given by

$$d_{\mathrm{KS}}(p,q) := \sup_{x\in\mathbb{R}} |F_p(x) - F_q(x)|. \tag{6.4}$$

In principle, the test could be extended to the multivariate case by using the generalized cumulative distribution function, but this extension is not straightforward as the orientation on $\mathbb{R}^n$ for the chosen representation can influence the supremum (see [JPZ97]). Interestingly, d_{KS} is the first distance measure that is symmetric in p and q. A further advantage is that one can use the KS test to check whether the distributions differ significantly at the chosen significance level. The other methods are less amenable to constructing such a hypothesis test. Naturally, the restriction to the 1D case makes the applicability of this metric rather niche.

Another metric that is mostly used in 1D is the Wasserstein distance [Kan60]. While it can be defined in any dimension, computational costs grow fast. Hence, it is mostly applied to 1D. However, as a trick to apply it to higher dimensional data, one can apply it to each marginal distribution independently. Naturally, this comes at a cost of expressiveness of the distance. Intuitively, the Wasserstein distance measures the minimal cost of moving the mass of the first distribution to the second distribution. In 1D, this can be expressed by the inverse of the cumulative distribution functions

$$W_n(p,q) := \left(\int_0^1 |F_p^{-1}(x) - F_q^{-1}(x)|^n \,\mathrm{d}x\right)^{1/n}, \tag{6.5}$$

where $n \geq 1$. In higher dimensions, the definition becomes more involved and harder to compute. The advantage of the Wasserstein distance is that it is a metric in the mathematical sense on the space of distributions, i. e. it is non-negative, symmetric, satisfies the triangle inequality, and is 0 if and only if $p = q$. As for the KS statistic the main downside is its "restriction" to 1D.

The only remaining branch is the catch-all branch. If our distributions have no exploitable structure, we are either left with the heuristic Wasserstein distance (i. e. applying it to the marginal distributions) as introduced above or with MMD. MMD is a kernel-based method introduced in [Gre+12] in formulas it reads:

$$\mathrm{MMD}_k(p, q) := \mathbb{E}_{x,x' \sim p}[k(x, x')] + \mathbb{E}_{y,y' \sim q}[k(y, y')] - 2\mathbb{E}_{x \sim p, y \sim q}[k(x, y)], \tag{6.6}$$

where $k \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is the kernel. Depending on certain properties of the kernel (cf. [Gre+12] for more details), this defines a metric on the space of distributions. Oftentimes, it is sufficient to work with translation-invariant kernels which take the form $k(x, y) = \tilde{k}(x - y)$. Two common kernels are the Gaussian kernel

$$\tilde{k}_{\mathrm{G}}(x) := \exp\left(-\frac{\|x\|_2^2}{2\sigma^2}\right), \tag{6.7}$$

with parameter σ[7] and the inverse multiquadric kernel

$$\tilde{k}_{\mathrm{IM}}(x) := \frac{1}{\sqrt{1 + \frac{\|x\|_2^2}{\sigma^2}}} \tag{6.8}$$

with parameter σ as introduced in [Tol+17]. The inverse multiquadric kernel has the advantage of falling off slower in the tails compared to the Gaussian kernel, which is beneficial for comparing heavy-tailed distributions. The great advantage of MMD is that we can apply it to any pair of distributions as a two-sample test. This flexibility is, at the same time, its weakness: We have to choose a kernel, and each kernel comes with parameters we have to determine. Even worse, starting with two kernels, we can create a new kernel by summing them. The problem is that MMD is very susceptible to the chosen kernels and parameters. For example, if we choose the Gaussian kernel and the distributions only differ in their tails, then the MMD will suggest that the distributions are very similar because the Gaussian distribution suppresses differences in the tails. Hence, it might become necessary to perform a hyperparameter search just to find the optimal setting of the metric, which is sub-optimal. Still, MMD is a great backup option if no other metric, distance, or divergence is available.

Object Detection-inspired Metrics The object detection block consists of two main blocks. The mode matching and the computation of metrics on the matched pairs. The mode matching is summarized in sub-process S3 as seen in figure 6.9. In object detection,

[7] In principle, we are not restricted to an isotropic Gaussian but could use a more general covariance matrix Σ as parameters.

the matching step is separated into two steps: First, the localization step computes some distance between or an overlap of the instances in question. This step is described in sub-process S4 in figure 6.10. Second, the actual assignment step uses the previously computed distances to come up with the matching. This step is described in sub-process S5 in figure 6.11. We transfer the object detection matching steps one-to-one to the mode matching steps.

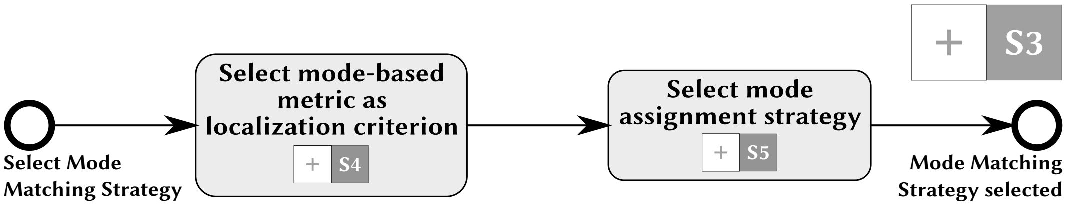

+ S3
Select Mode Matching Strategy
Select mode-based metric as localization criterion
+ S4
Select mode assignment strategy
+ S5
Mode Matching Strategy selected

Figure 6.9.: **Sub-process for matching predicted modes to reference modes.** The sub-process consists of two parts. First, the mode localization, which computes a suitable distance between all pairs of modes, and second, the assignment, where the predicted modes are matched with the reference modes based on the localization criterion.

For the mode localization sub-process S5, we distinguish once more between a reference distribution and a discrete set of reference mode coordinates. In the first case, we can use all the distribution-based metrics introduced in the previous paragraph (sub-process S2). To this end, we interpret each mode as a distribution in its own right. We have only the predicted distribution if the reference is a discrete set of modes. In this case, we can still compute the distance of the mode coordinates to the centroid of the predicted modes. Depending on the skewness of the predicted mode, a normal distribution might be a good approximation. In this case, the centroid distance can take the variance of the mode into account via the Mahalanobis distance [Mah36]. In the same vein, if an exact match of the centroids is not important, but a rough match with the mode is sufficient, a confidence ellipsoid based on the mode could be built, and the computed distance would be binary depending on whether the reference is within or outside of the confidence ellipsoid. All the metrics that work for a discrete set of mode coordinates can also be applied if the reference is a distribution by simply reducing to the mode coordinates. Whether this is advisable depends on the application properties of the blueprint.

Distribution-based metrics will always take the shape of the modes into account and might unfairly penalize a predicted mode whose location is correct but whose variance is off. Hence, if uncertainty quantification is negligible, the distribution-based metrics might not be a good proxy for the application's success. Instead, a centroid-based distance might be better suited because it only checks for the mode location, ignoring the variance completely.

Once the modes are localized, i. e. a distance between each pair of reference and predicted

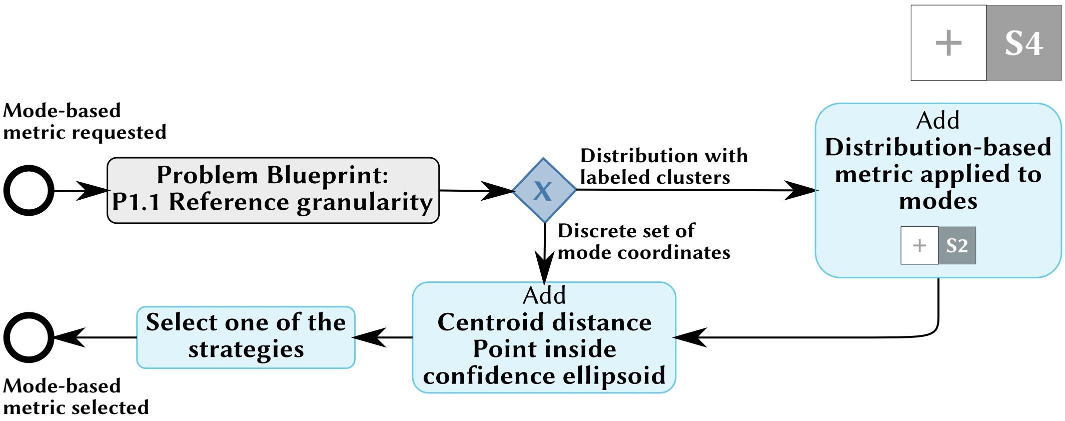

+ S4
Mode-based metric requested
Problem Blueprint: P1.1 Reference granularity
X
Distribution with labeled clusters
Add Distribution-based metric applied to modes
+ S2
Discrete set of mode coordinates
Add Centroid distance Point inside confidence ellipsoid
Select one of the strategies
Mode-based metric selected

Figure 6.10.: **Sub-process for mode localization.** Localization is the process of computing a distance between the predicted and reference mode. This sub-process aims at selecting the distance.

mode is computed, we can start with the final matching as shown in sub-process S5 (cf. figure 6.11). The first branching point concerns the existence of confidence scores for the predicted modes. If they exist, more confident modes should be assigned first (greedy). If there is no confidence score, there are three options. We can still perform a greedy assignment, in this case, based on the localization criterion. Alternatively, we can use the Hungarian matching algorithm [Kuh55], which optimizes the global assignment costs. Whether to prefer the greedy matching or the Hungarian matching is a bit of a philosophical question. The Hungarian matching considers all assignments simultaneously and minimizes global costs. For a theoretical evaluation, this might be exactly what we want. However, for many downstream tasks, we might prefer not to have to handle all modes all the time, but instead, start with the best located one and, only if that one does not fit, work our way down to modes with worse localization. This would correspond to the greedy approach. Another simple matching approach considers all pairs with the localization distance below a certain threshold a match. This approach is only advisable if predicted modes cannot overlap with multiple reference modes or if double assignments are of no concern.

In any case, as a last step in the assignment process, we need to decide how to handle double assignments. This is again driven by domain-specifics. If it is okay that two or more predicted modes are assigned to the same reference mode, we can simply ignore the double assignment. Otherwise, a common practice is only to allow a single assignment (either the one with the higher confidence or the better localization metric) and count the other mode as unassigned. During metric aggregation, we can then tabulate the matched and unmatched modes to build a confusion matrix[8] and compute classification metrics like

[8] However, without the false negative cell.

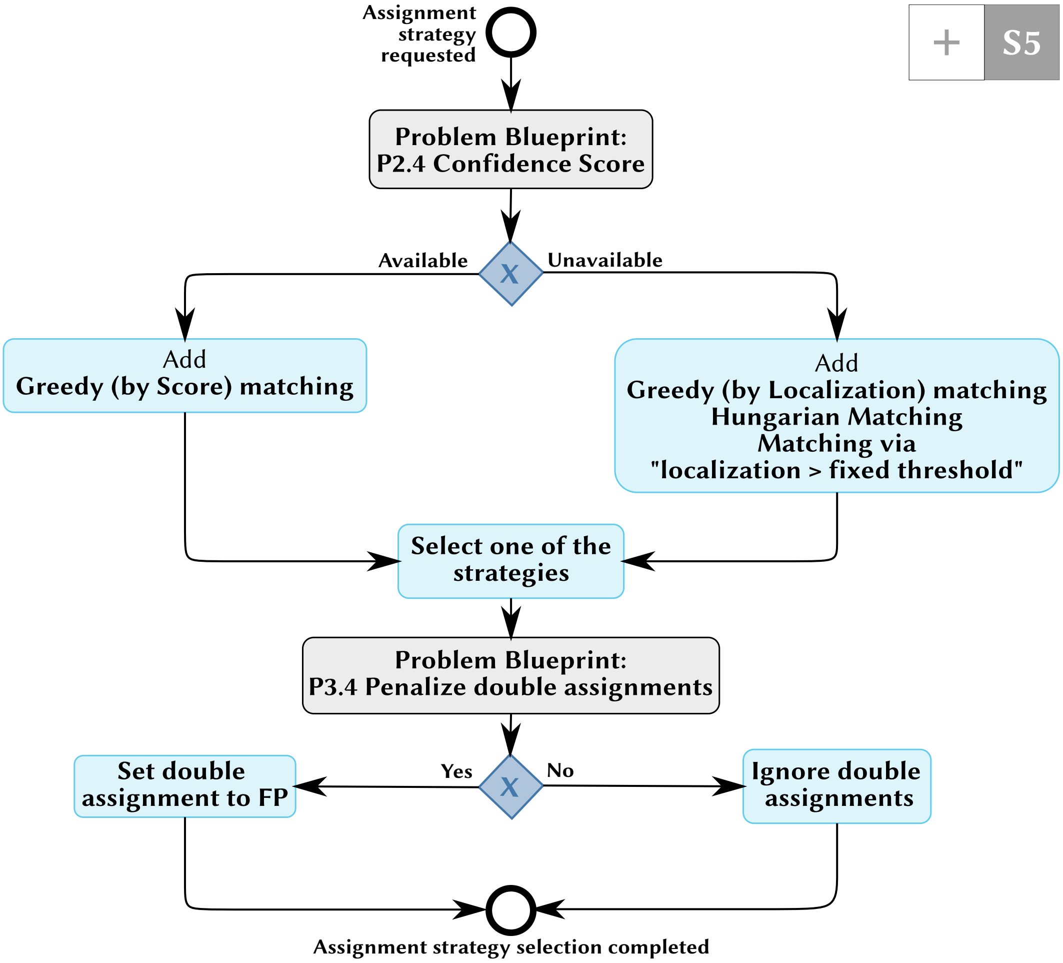

Assignment strategy requested
+
S5
Problem Blueprint:
P2.4 Confidence Score
Available
X
Unavailable
Add
Greedy (by Score) matching
Add
Greedy (by Localization) matching
Hungarian Matching
Matching via
"localization > fixed threshold"
Select one of the strategies
Problem Blueprint:
P3.4 Penalize double assignments
Yes
X
No
Set double assignment to FP
Ignore double assignments
Assignment strategy selection completed

Figure 6.11.: **Sub-process for mode assignment.** Based on the localization criterion from sub-process S4 (figure 6.10) and a possible confidence score, the reference and predicted modes are assigned to one another. As a final step, double assignments are handled.

precision and recall. These would penalize the unassigned modes.

After the modes have been matched, we can compute metrics on the matched pairs. These metrics coincide with the metrics used for mode localization introduced in sub-process S4.

Data Set Aggregation The previous paragraphs explained the process of collecting metrics on a per data point basis, i. e. comparing a single predicted posterior to the reference. In this paragraph, we will walk through the sub-processes to choose fitting aggregation strategies to get a single value per data set. The whole data set aggregation block consists of two sub-processes. One for multi-threshold metric aggregation (sub-process S6, cf. figure 6.12) and one for single-threshold and threshold-independent metric aggregation (sub-process S7, cf. figure 6.13). While S6 is only accessible with a predefined threshold range, S7 is always accessible.

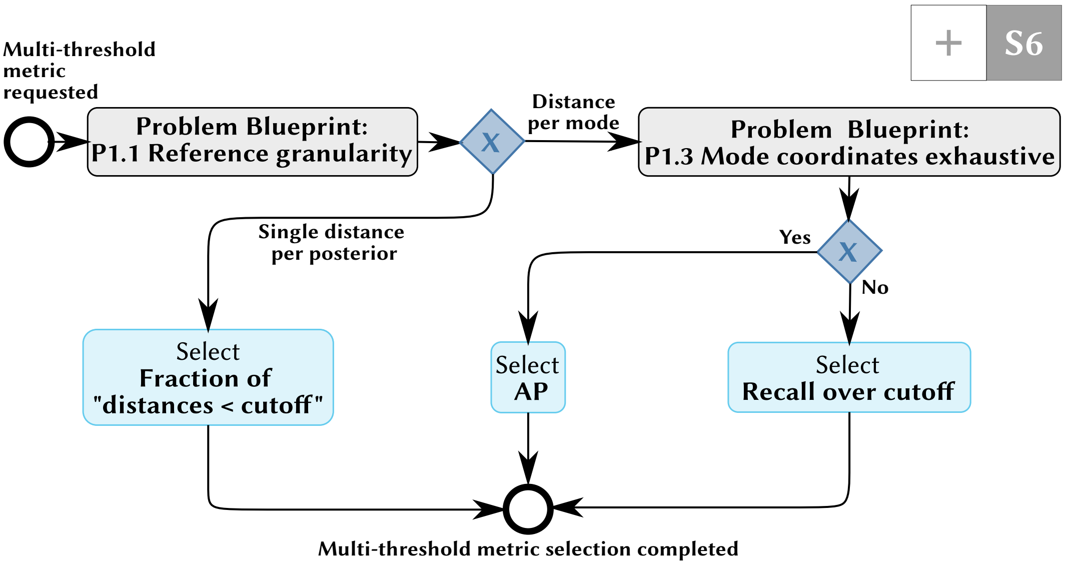

+ S6
Multi-threshold metric requested
Problem Blueprint: P1.1 Reference granularity
X
Distance per mode
Problem Blueprint: P1.3 Mode coordinates exhaustive
Single distance per posterior
Yes
X
No
Select Fraction of "distances < cutoff"
Select AP
Select Recall over cutoff
Multi-threshold metric selection completed

Figure 6.12.: **Sub-process to select appropriate multi-threshold metrics.** Depending on whether the reference modes are labeled and the reference modes are exhaustive, different metrics are available. AP: average precision.

In the multi-threshold setting, we first branch off based on the reference labeling state. If the reference has no labeled modes, we can only use distribution-based metrics chosen in S2. This setting is ill-fitting for the multi-threshold metric setting but is better suited to threshold-independent metrics like summary statistics, as introduced in S7. At the same time, if the metrics computed on the predicted and reference distribution bear meaning for the application and come with a natural range, it might be interesting to scan this range and record the proportion of data points below this cutoff. The plot might be informative for further exploration of the data set. For further aggregation, the area under curve (AUC) of this curve might be considered. If the reference modes are labeled, we can apply counting metrics based on the confusion table of a classification task. In fact, for each mode, we either have an assigned partner or the mode is unassigned. This allows us to compute the number of true positive (TP) (matched modes), false negative (FN) (unassigned reference modes), and at first glance, FP (unassigned predicted modes). However, at a second glance, we need the reference modes to be exhaustive to be able to compute the number of false positives. Otherwise, the predicted mode might be unassigned due to a missing reference mode. TNs are undefined in detection. With TP and FN, we have access to recall, also known as sensitivity, which is defined as

$$\text{recall} = \text{sensitivity} := \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{6.9}$$

If we have access to FP, we can additionally compute the precision (or positive predictive value)

$$\text{precision} := \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{6.10}$$

We can compute the precision-recall curve with precision and recall at different threshold values, a standard tool for evaluating classification tasks. The precision-recall curve can be aggregated further via the AP, which is defined as [ZZ09]:

$$\text{AP} := \sum_i (\text{recall}_i - \text{recall}_{i-1}) \cdot \text{precision}_i, \tag{6.11}$$

where i iterates over the threshold values. AP is restricted to the interval $[0, 1]$ with higher values indicating better detection performance. This makes the metric easy to interpret and widely used. If the precision is unavailable, we can only compute the recall over the threshold range and aggregate it via AUC, but this approach is rather limited in its expressiveness.

Sub-process S7 describes the single-threshold and threshold-independent metric selection. If a threshold or cutoff is provided, the available counting metrics depend on the completeness of the reference. If it is complete, we proceed completely analogously to the object detection setting, where the cutoff might either be specified as a cutoff on the recall, in which case we compute the precision at the specified recall cutoff, or as a cutoff on the precision, in which case we compute the recall at the specified precision cutoff. A recall cutoff is preferable if we want to limit the rate of undetected modes. A precision cutoff will limit the number of unassigned predicted modes. If the reference is incomplete, we are reduced to computing the recall. In this case, the cutoff cannot be defined in terms of a counting metric, as there is no second counting metric that we could use, but instead would need to be specified based e. g. on the localization criterion or the confidence score. If there is no natural choice for such a cutoff, we are reduced to the threshold-independent metrics.

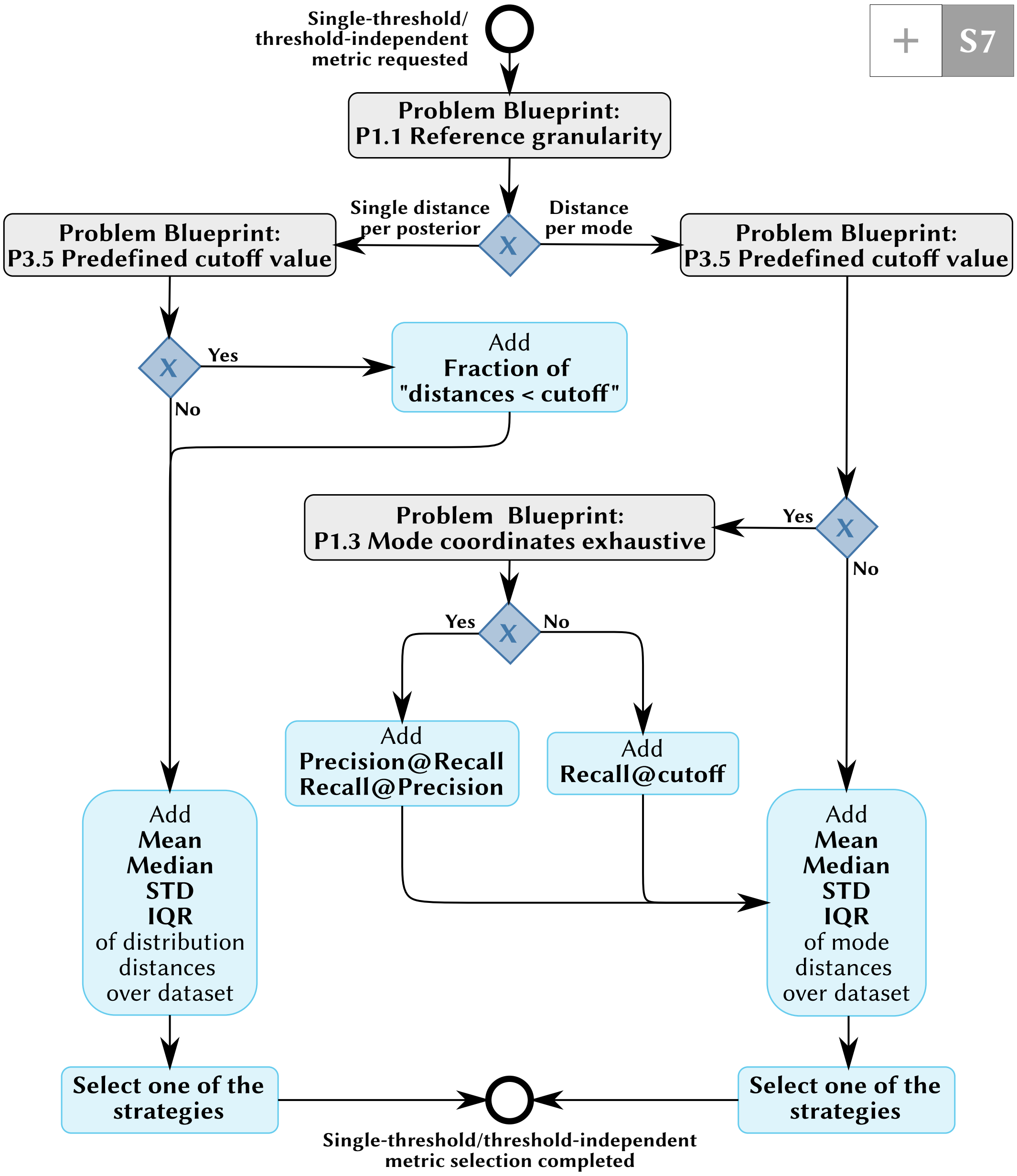

+ S7
Single-threshold/ threshold-independent metric requested
Problem Blueprint: P1.1 Reference granularity
Single distance per posterior
Distance per mode
Problem Blueprint: P3.5 Predefined cutoff value
Problem Blueprint: P3.5 Predefined cutoff value
Yes
No
Add Fraction of "distances < cutoff"
Problem Blueprint: P1.3 Mode coordinates exhaustive
Yes
No
Yes
No
Add Precision@Recall Recall@Precision
Add Recall@cutoff
Add Mean Median STD IQR of distribution distances over dataset
Add Mean Median STD IQR of mode distances over dataset
Select one of the strategies
Select one of the strategies
Single-threshold/threshold-independent metric selection completed

Figure 6.13.: **Sub-process to select single-threshold and threshold-independent metrics.** The choice of metrics depends on the existence and completeness of labeled reference modes. If a threshold value is available, the metrics correspond to the multi-threshold setting S6 (cf. figure 6.12). Threshold-independent aggregation is achieved via descriptive statistics over the appropriate distance distribution. STD: standard deviation, IQR: interquartile range.

For threshold-independent metrics, we have to consider the following cases. If the reference modes are not labeled, we have distances computed between the predicted and the reference posterior. In the other case, we have distances computed between the modes. In both cases, we can compute descriptive statistics on the distribution of these distances, e. g. the mean, median, mode, standard deviation, IQR, or many others. Which exactly will depend on the application. Generally, quantile-based statistics are more robust with regard to outliers, which might be preferable in a noisy setting. If the distances are computed on the modes, we need to be careful because we introduce a hierarchical structure in the data. The data set consists of posteriors, and each posterior consists of one or more modes. This hierarchy should be respected when aggregating the data, i. e. the mode distances for a single posterior should be aggregated first before aggregating the results for all posteriors. For more details, please refer to [WWG06].

Please note that the original metric selection framework [Mai+22b] includes a block to determine a suitable calibration metric for classification(-derived) tasks. However, the calibration term used in object detection is slightly inconsistent with the calibration term used for cINNs. While the cINN calibration is part of the proposed framework, object detection calibration has been omitted from the current version, and transferring it remains future work.

Overall, we have introduced a posterior validation framework that uses key properties of an inverse problem setting (the problem blueprint) to derive a problem-specific set of metrics. In this way, our framework can structure and increase the transparency and comparability of posterior-based inverse problem analyses.

6.3. Out-of-Distribution Detection for Biophotonic Imaging

In section 4.3.2, we gave a general introduction to OoD detection and noted some of the caveats we might encounter. One class of OoD detectors are the density-based methods. As introduced in section 4.4, INNs were developed as density estimators, so it is only natural to try to use them as OoD detectors. In this section, we will highlight a common disadvantage of density-based OoD detectors and introduce the widely applicable information criterion (WAIC) as a method to avoid it (section 6.3.1). Afterward, we will introduce our concept to use OoD detectors to monitor physiological tissue parameter changes (section 6.3.2).

This section contributes toward answering research question T.3, which was concerned with biophotonic imaging specific OoD detection methods.

Disclosure and Contributions I proposed the use of WAIC together with INN ensembles to gauge the realism of our MSI simulation framework. Lena Maier-Hein advised and guided me and proposed experimental setups. I implemented the method, performed the experiments, and collected the results. In addition, Lena Maier-Hein suggested the use of WAIC to detect physiological parameter changes as an OoD detection task.

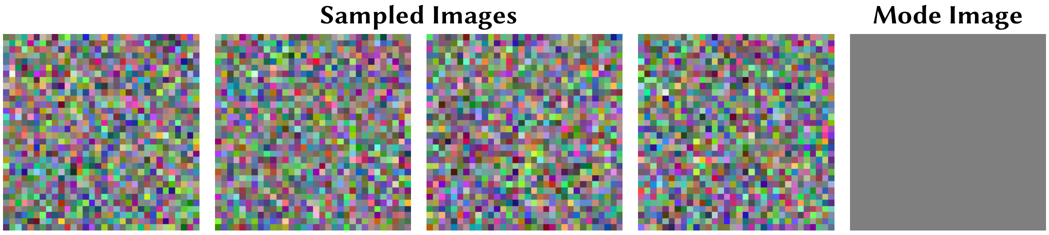


Figure 6.14.: **Comparison of four samples to the mode of a high-dimensional distribution.** The samples are drawn from a Gaussian with mean $\mu = 0.5$ and standard deviation $\sigma = 0.2$ for each pixel and color of the image. The figure was adapted from [Nal+19a].

6.3.1. The Widely Applicable Information Criterion for Out-of-Distribution Detection

High-dimensional data often behaves unexpectedly. This observation is sometimes subsumed in the term *curse of dimensionality.* Intuitively speaking, it is concerned with the fact that the volume of a neighborhood of a point shrinks with increasing dimension. More precisely, the fraction of the volume of the unit ball compared to the volume of the unit hypercube[9] goes to zero as the dimension goes to infinity. While this sounds rather abstract, it has practical implications. Consider, for illustration purposes, a standard Gaussian distribution. Its mode, i. e. its point with maximal density, is located at zero. Next, let us consider a small ball around this mode and sample points for the Gaussian distribution. In low dimensions, we can expect many points to fall within our ball, but the chances get smaller as the dimension increases. Even though the mode is the point with the highest density, we have to consider the volume surrounding it, which decreases ever further. As a result, we should not expect to draw samples close to the mode for high-dimensional data sets. They will stem from a region with lower density but larger volume. This set is called the *typical set* [Sha48; Nal+19a].

An example of this phenomenon is depicted in figure 6.14. Each image on the left-hand side is drawn from the same distribution of a Gaussian centered at $\mu = 0.5$ with standard deviation $\sigma = 0.2$ for each pixel and color. Clearly, the sampled images look very different than the "mode image", which is visible on the right-hand side.

The conclusion of this discussion is that using the likelihood of a data point as an OoD score is misleading because a very high likelihood on high-dimensional data should be taken as evidence for an outlier. There are multiple ways to address this issue [CJA18; Nal+19a]. We will introduce one that is based on the epistemic uncertainty of the model used to estimate the density.

[9] which is 1

The first observation is that the true data distribution is not accessible for virtually every realistic application. Hence, we are forced to estimate the density of the in-distribution (ID) data using a family of densities p_Θ (cf. section 4.3.2). Now, we take a Bayesian perspective and assume there is no one correct parameter vector Θ, but instead a posterior distribution $p(\Theta \mid X_{\mathrm{ID}})$ given the ID data X_{ID}. This perspective is natural in the deep learning regime, where the models are generally overparameterized[10], such that multiple parameter configurations could lead to similar predictions. The training enforces this behavior on the ID data. However, there is no training signal to enforce that different $\Theta \sim p(\Theta \mid X_{\mathrm{ID}})$ lead to similar densities $p_\Theta(x_{\mathrm{OoD}})$ for an OoD data point x_{OoD}. On the contrary, we would expect high uncertainty in this region because of the overparameterization.

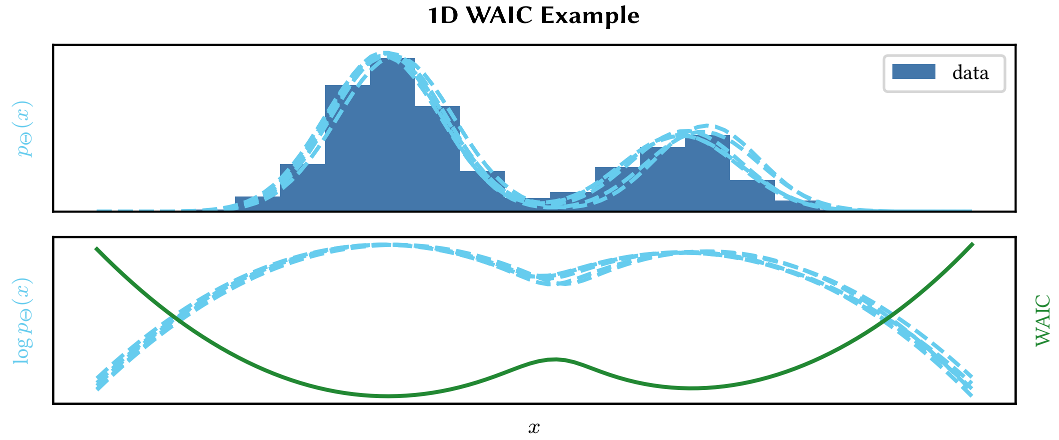


Figure 6.15.: **Illustration of the WAIC score.** The data set is represented by a 1D Gaussian mixture model with two components (dark blue). The density estimation models are depicted in dashed turquoise. The widely applicable information criterion (WAIC), computed following equation (6.12), is depicted in solid green.

Based on this theory, Choi et al. [CJA18] proposed the use of the widely applicable information criterion (WAIC) as an OoD score. WAIC was introduced by Watanabe [Wat13] and is given by

$$\begin{aligned} \mathrm{WAIC}(x) &:= \mathbb{V}\mathrm{ar}_\Theta[-\log p_\Theta(x)] + \mathbb{E}_\Theta[-\log p_\Theta(x)] && (6.12) \\ &= \mathbb{V}\mathrm{ar}_\Theta[\log p_\Theta(x)] - \mathbb{E}_\Theta[\log p_\Theta(x)], && (6.13) \end{aligned}$$

where the expectation and variance is taken over $\Theta \sim p(\Theta \mid X_{\mathrm{ID}})$. The formula consists of two terms. First, a variance term intended to measure the uncertainty as motivated

[10] Roughly speaking, this means they have too many parameters compared to the amount of available training data.

above. Secondly, the mean log-probability assures that data points from a low-density region get assigned high WAIC values whether the uncertainty is high or not. Please note, that Choi et al. [CJA18] use an opposite sign convention, such that lower WAIC values imply OoD, but we follow the original convention as introduced by Watanabe [Wat13]. A low-dimensional visualization of WAIC can be found in figure 6.15.

In order to compute WAIC, we need two ingredients. First, a way to evaluate the log-density $\log p_\Theta$. This explains the importance of INNs, as these enable us to estimate the density of the ID data X_{ID}. Second, we need to aggregate over Θ following $p(\Theta \mid X_{\mathrm{ID}})$. This is achieved using an ensemble approach. In our case, each INN is randomly initialized with a separate seed and trained until convergence. The resulting parameter vectors are interpreted as a sample from $p(\Theta \mid X_{\mathrm{ID}})$ and the expectation value and variance are approximated using the empirical mean and variance:

$$\hat{\mu}(x) := \frac{1}{n}\sum_{i=1}^{n} \log p_{\Theta_i}(x), \tag{6.14}$$

$$\hat{\sigma}^2(x) := \frac{1}{n-1}\sum_{i=1}^{n} (\log p_{\Theta_i}(x) - \hat{\mu}(x))^2, \tag{6.15}$$

$$\mathrm{WAIC}(x) \approx \hat{\sigma}^2(x) - \hat{\mu}(x). \tag{6.16}$$

All in all, INNs enable us to estimate higher dimensional densities, and WAIC counteracts the problems of density-based OoD detection due to the curse of dimensionality. Both together lead to an interesting OoD detection methodology for our biophotonic imaging setting. We will see the validation experiments in section 7.3.

6.3.2. Out-of-Distribution Detection for Personalized Medicine

Commonly, we would phrase the task of monitoring physiological tissue parameters, like sO_2 or vHb, as a supervised learning task. To stay with the example of biophotonic imaging, we would acquire a data set consisting of spectral data and corresponding labels. The labels could either be actual tissue parameter values, in which case the task would be a regression task, or a coarse description of the tissue state, like ischemic or perfused, in which case the task would be a classification task. This classic approach to perfusion monitoring based on classification can be seen in figure 6.16 in dark blue. As ML approaches in general and DL approaches, in particular, are susceptible to OoD data, we would require a large and diverse data set for this approach to work. Especially in settings where the inter-patient variability is high, we would need a large patient cohort for new patients not to be OoD. Analogous observations are true for variability due to a change in recording hardware and/or the recording site. This is one of the reasons that complicate the application of classic ML methods in medicine: It is notoriously hard to collect large, annotated data sets.

This section proposes a method that counteracts the problems due to e. g. inter-patient variability. Depending on the viewpoint, the underlying idea is either elegantly or naively

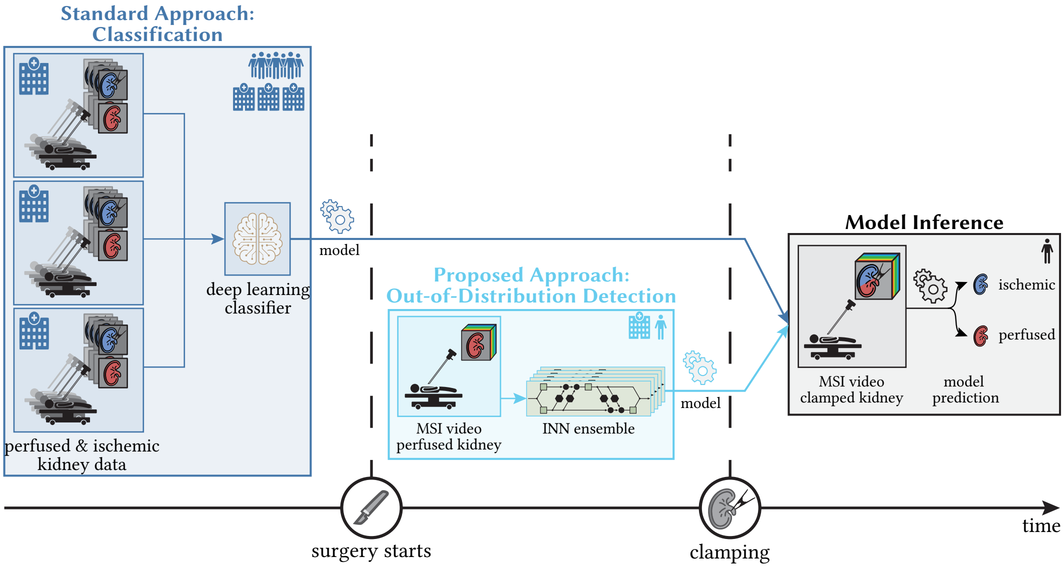


Figure 6.16.: **Proposed personalized approach to tissue parameter monitoring.** We use the example of perfusion monitoring. Classic machine learning (ML) approaches (blue) would treat this approach as a classification task that would require a large and diverse patient cohort. The proposed solution (turquoise) operates on a single patient. We use an invertible neural network (INN) ensemble as a central component. They compute the likelihood of ischemia based on a short multispectral video sequence of perfused spectra. Ischemic spectra can then be detected as out-of-distribution (OoD). With pre-training, the ensemble can be trained and perform inference fast enough for surgical applications. This figure was adapted from [Aya+22].

simple: We only have to consider the inter-patient variability if we want to apply the same ML model to multiple patients. In other words, if we can train an ML model for each patient individually, we can ignore many of the confounding effects that decrease the performance of a model that is intended to operate on multiple patients (analogously for multiple hardware setups and sites). Hence, the use of the term "personalized medicine".

Naturally, such a personalized approach introduces its own set of problems. Let us stay with the ischemia monitoring example. If we want to train a personalized model in a supervised fashion, we would require labeled training data from the patient in question. This training data would include labels (ischemic and perfused). However, that implies the existence of a reference method that created the labels. Furthermore, in their natural state, (luckily) most patients do not have ischemic organs. Hence, we would require an intervention just to get the training data. Such an intervention seems ethically questionable, and the existence of a reference method poses the question of the benefit of a second classifier. Both together underline that supervised learning would be hard, if not impossible, to use in a personalized setting.

Instead, we propose an unsupervised learning approach based on OoD detection (cf. figure 6.16, turquoise). While most patients do not have ischemic organs, we can collect perfused organ spectra with a reasonable effort at the beginning of an intervention. We define these perfused spectra as ID. Under the condition that we can train an OoD detector fast enough, we can then use it to detect ischemic spectra as OoD. This approach removes the need for a reference method[11] and the model would be trained during the main intervention, sidestepping ethical problems. We will use INN ensembles and WAIC as OoD detectors. The INNs can be pre-trained on simulated data, shortening the training time on the actual patient to a realistic range for training during the actual intervention. Additionally, the inference time of the approach is fast enough to allow video rate predictions.

For the proposed method to work reliably, we need to ensure that the detected change is due to the change in physiology and not due to a confounding factor like a change in pose or lighting. We will address this issue with the setup of our validation experiments as found in section 7.4. Overall, personalized physiological parameter monitoring is an exciting new avenue in medical ML. It has the potential to circumvent many of the common ML generalization problems due to new patients, hardware, or sites, leading to more robust medical ML systems.

[11] As long as we are sufficiently certain that the organ is perfused at the beginning of an intervention.

7. Experiments and Results

In this chapter, we will demonstrate the value of our framework for uncertainty handling in biophotonic imaging by means of four experiments. We will thereby underline the strengths of INNs and cINNs that put them at the core of our methodology. The experiments roughly follow the domain research questions D.1 – D.3 introduced in chapter 2. In sections 7.1 and 7.2, we will analyze the well-posedness of the inverse problems in MSI and PAI. To this end, we will construct cINN architectures and encode the solutions to the inverse problems as posterior distributions. The contributions in these sections address research question D.1, which was concerned with the well-posedness of inverse problems in biophotonic imaging. In section 7.3, we will explore the realism of our simulated MSI data and compare them to *in vivo* data using an OoD approach based on INNs and WAIC. The contributions in this section address research question D.2, which was concerned with the realism of *in silico* MSI data. In section 7.4, we will transfer this OoD approach to detect physiological parameter changes at the example of ischemia detection in minimally invasive kidney surgery in humans. The contributions in this section address research question D.3, which was concerned with the possibility of monitoring physiological tissue parameters as an OoD detection task.

All experimental sections follow the same structure. First, there is a short introduction to the experiment, including a paragraph about disclosures and contributions explaining where the work might have been previously published and what my contribution to the work was compared to my collaborators. Second, there is a section about method details introducing any methods particular to the experiment not introduced in the general methods section. Third, the experimental setup is described in detail, followed by the results and a short experiment-specific discussion. A longer and more exhaustive discussion considering this thesis's findings can be found in chapter 8.

7.1. Analyzing the Well-Posedness of Quantitative Multispectral Imaging

In this experiment, we try to understand how cINNs can potentially help us to design better multispectral cameras. Indeed, higher spectral resolution generally comes at the cost of slower acquisition time or lower spatial resolution. Hence, it makes sense to only record spectral bands that help in the prediction task. cINNs are good candidates to identify these bands. With a full posterior distribution, we are not restricted to simple prediction performance but can take the spread and possible multimodes of the posterior into account. As a proof of concept, we performed an *in silico* study where we used high-resolution, simulated spectra and the filter response functions of different cameras to adapt the spectra to each camera. Afterward, we trained cINNs to predict the physiological tissue parameters sO_2 and vHb.

The above formulation of the experimental goal is very application-driven, but there is an equivalent formulation that highlights the relation of our experiment to research question D.1. D.1 asks whether biophotonic inverse problems are well-posed. Here, we will use the posteriors generated by the cINNs and the different cameras to answer this question for varying spectral resolutions.

Disclosure and Contributions This work is based on previous work published at the International Conference on Information Processing in Computer-Assisted Interventions (IPCAI) 2019 [Adl+19b]. However, the results found in this thesis are a reproduction, using cINNs, whereas the original publication was based on the older INN architecture. Since INNs are very hard to train in the inverse problem setting, they have effectively been outdated by the newer cINNs. Hence, we focus on the state of the art architecture. Regarding the original publication, I would like to mention the following contributions: Lena Maier-Hein came up with the study idea and design. She supervised the whole process and provided valuable feedback and guidance. Leonardo Ayala, Anant Vemuri, and Sebastian Wirkert generated the *in silico* data set used for this study. Thomas Kirchner and Janek Gröhl discussed all preliminary results and provided valuable feedback. Lynton Ardizzone, Jakob Kruse, Carsten Rother, and Ullrich Köthe developed the INN and cINN architecture and gave advice regarding implementation details. I performed all experiments, implemented all the code, and evaluated all results.

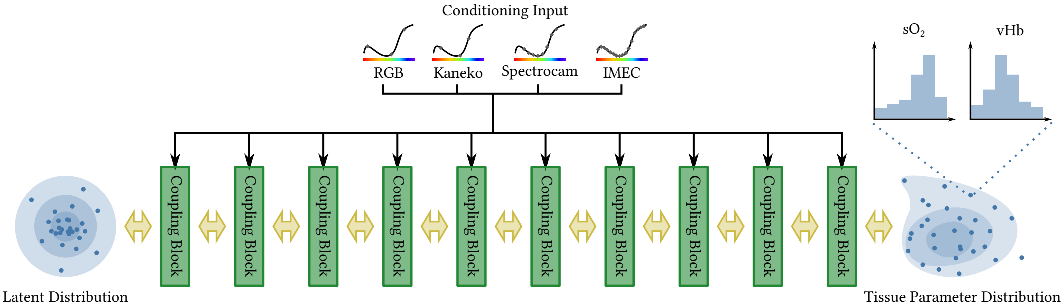


Figure 7.1.: **Schematic overview of the conditional invertible neural network (cINN) architecture.** The latent distribution is transformed by ten affine coupling blocks to the tissue parameter distribution. Each coupling block receives a camera spectrum (either RGB, Kaneko, SpectroCam, or IMEC) as a conditioning input. The figure was adapted from [Adl+19b].

7.1.1. Method details

As mentioned in the introduction of the section, we aimed to compare the performance of different cameras for medical use with the example of regressing sO_2 and vHb from the spectra generated by each camera. In this section, we will introduce the cINNs trained on the camera data sets and the posterior validation process.

cINN Architecture and Training For each camera, we trained a cINN as introduced in section 4.4.2 consisting of ten affine coupling blocks (see figure 7.1). The cINNs were implemented in PyTorch using the framework for easily invertible architectures (FrEIA) [Ard+18a]. We predicted sO_2 and vHb leading to a 2D latent space. Depending on the chosen camera, the conditioning input was either 3D, 8D, or 16D. Each affine coupling block consisted of a shallow fully-connected subnetwork with a single hidden layer of dimension 256, ReLU activations, and was initialized using the Kaiming initialization scheme [He+15]. We performed exponent clamping with 2.0 and initialized the global transformation with 0.7. For optimization, we used the maximum-likelihood loss as introduced in equation (4.21) and the AdamW optimizer [LH17] with a learning rate of $5 \cdot 10^{-4}$ and L2 weight regularization parameter of $1 \cdot 10^{-5}$. We trained each network for 100 epochs, and the batch size was 256. In addition, we added Gaussian noise to the training data with standard deviation $\sigma = 0.01$ for the spectra (conditioning) and $\sigma = 0.001$ for the tissue parameters (input) as data augmentation.

Posterior Validation We computed a full posterior distribution over the tissue parameter space for each spectrum by sampling 1024 latent samples and transforming them with the cINN. We computed the calibration curve and error for each camera and tissue parameter combination as introduced in section 4.4.5. We used the posterior median as a tissue parameter predictor and computed the absolute error distribution. Furthermore, we computed the IQR as a measure of spread for the posteriors. Lastly, we used the diptest introduced in section 6.1.2 at level $\alpha = 0.01$ to compute the proportion of multimodal posteriors.

7.1.2. Experiment description

Cameras In this experiment, we compared the performance of four virtual representations of cameras. Two of the cameras were multispectral cameras: 1.) the Pixelteq (Largo, FL, USA) SpectroCam (8 bands) and 2.) the XIMEA (Münster, Germany) MQ022HG-IM-SM4x4-VIS (16 bands) abbreviated as IMEC. In addition, we approximated a regular RGB camera using Gaussian filter responses. Lastly, we used another 3-band camera, which was suggested, for medical applications [Kan+14], which we abbreviate as Kaneko. The filter response functions of the four cameras can be found in figure 7.2.

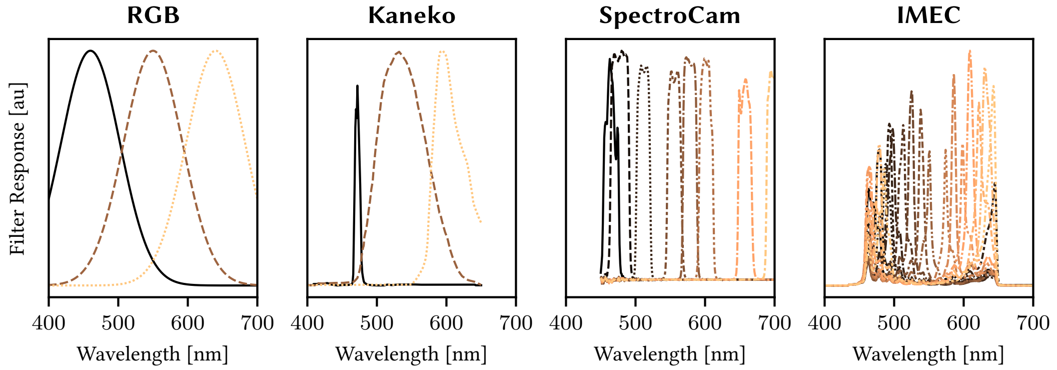


Figure 7.2.: **The filter response functions of the cameras.** Each function describes the sensitivity of the camera filter in a certain spectral range. The figure was adapted from [Adl+19b].

Data Set The data set was generated using Monte Carlo simulation with 10^6 photons per spectrum building on the MCML framework [WJ92]. The data set was simulated for previous work, and the exact implementation can be found in [Wir+17; Aya+22]. We modeled tissue as three infinitely wide layers with assigned optical and tissue parameters as found in table 7.1. The output was a high resolution spectrum (300 nm to 1000 nm in 2 nm steps). This spectrum can be adapted to different cameras using the filter response functions of each camera respectively (cf. figure 7.2). The tissue parameters for each layer were drawn independently, leading to three sO_2 values and three vHb values per sample. We aggregated these to obtain a single sO_2 and vHb value by averaging the values up to a depth of 250 nm, weighting each value by the layer thickness. The depth was chosen based on the penetration depth of light in tissue.

Tissue Parameter	vHb [%]	sO_2 [%]	a_{mie} [cm^{-1}]	b_{mie} [1]
Distribution	$\mathcal{U}(0, 30)$	$\mathcal{U}(0, 100)$	$\mathcal{U}(5, 50)$	$\mathcal{U}(0.3, 3)$
Tissue Parameter	g [1]	n [1]	d [cm]	
Distribution	$\mathcal{U}(0.8, 0.95)$	$\mathcal{U}(1.33, 154)$	$\mathcal{U}(0.002, 0.2)$	

Table 7.1.: **Tissue parameters of a single layer of the simulation framework.** The tissue parameter distributions for the three layers were identical. vHb: blood volume fraction, sO_2: blood oxygenation, a_{mie}: reduced scattering coefficient at 500 nm, b_{mie}: scattering power, g: tissue anisotropy, n: the refractive index, d: thickness of the tissue layer, $\mathcal{U}(a, b)$: Uniform distribution between a and b. This table was adapted from [Aya+22].

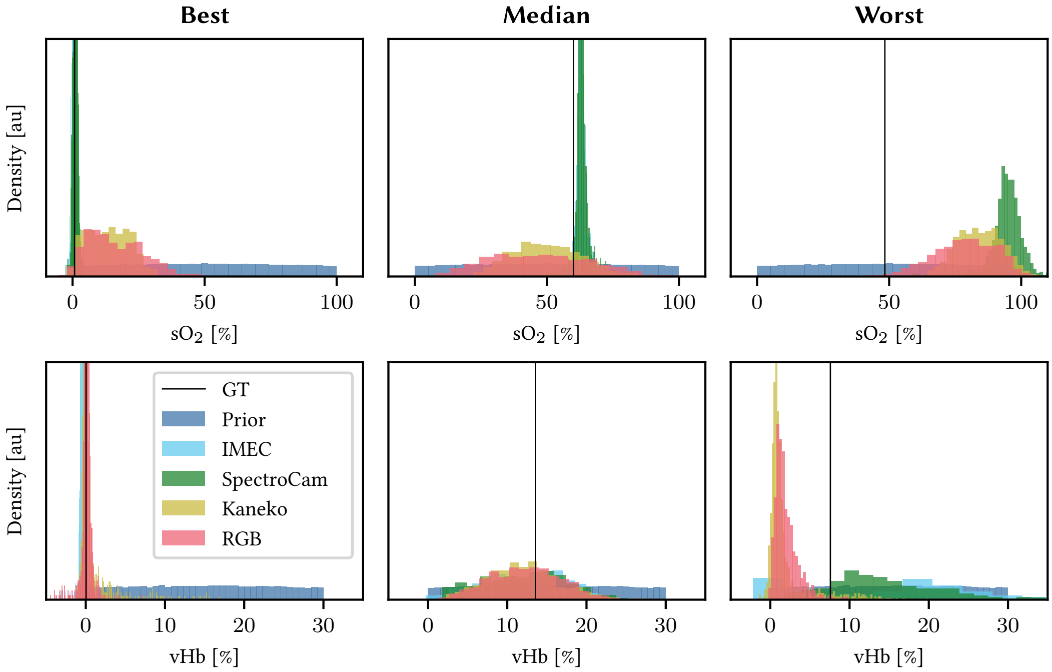


Figure 7.3.: **Example tissue parameter posteriors.** The examples were chosen based on the interquartile range (IQR) of the IMEC camera (best, median, and worst). The first row depicts blood oxygenation (sO_2) posteriors, while the second row depicts blood volume fraction (vHb) posteriors. The IMEC (turquoise) and SpectroCam (green) performance is almost indistinguishable, making the IMEC posterior nigh invisible in the plots. The figure was adapted from [Adl+19b].

Overall, we simulated 450,000 spectra for the training set, 50,000 for the validation set, and 50,000 as a test set. All reported results were computed on the test set, which was evaluated only once. The spectra were L2 normalized individually to remove the light intensity dependence. Afterwards, the data was band-wise z-score normalized (cf. section 4.4.4). The mean and standard deviation of the training set were applied to the validation and test set.

7.1.3. Results

Figure 7.4 shows the calibration curves and errors. For vHb, all cameras are well calibrated with a ECE of at most 1.3 pp. For sO_2, the RGB and Kaneko camera are still well calibrated with a slight over-confidence of the cINN (ECE $\leq$ 1.3 pp), whereas the SpectroCam and IMEC are very under-confident (ECE $\approx$ 18 pp). This implies that the posteriors for SpectroCam and IMEC are too wide.

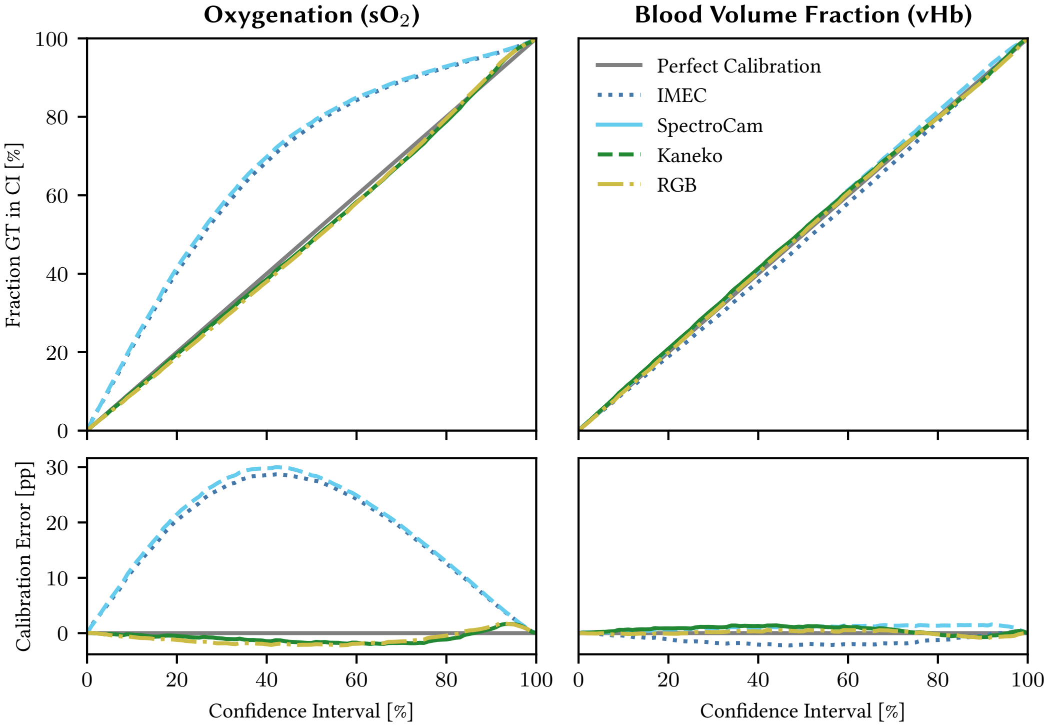


Figure 7.4.: **Models are either well calibrated or under-confident.** The calibration curve describes the fraction of ground truth (GT) values found in a certain confidence interval (CI). The calibration error is given by the fraction minus the confidence interval percentile. The figure was adapted from [Adl+19b].

Using the median as the tissue parameter predictor yields the absolute error distribution depicted in figure 7.5, first row. For sO_2, there is a deterioration in the absolute error of the Kaneko and RGB camera compared to the IMEC and SpectroCam cameras. For vHb, there is no trend in performance visible. The width of the posteriors as measured by IQR can be found in figure 7.5, second row, and mirrors the observations for the absolute error: For sO_2, SpectroCam and IMEC produce more narrow posteriors compared to the Kaneko and RGB camera. The median RGB IQR (21 pp) is noticeably higher than the Kaneko IQR (18 pp). For vHb, there is no clear trend.

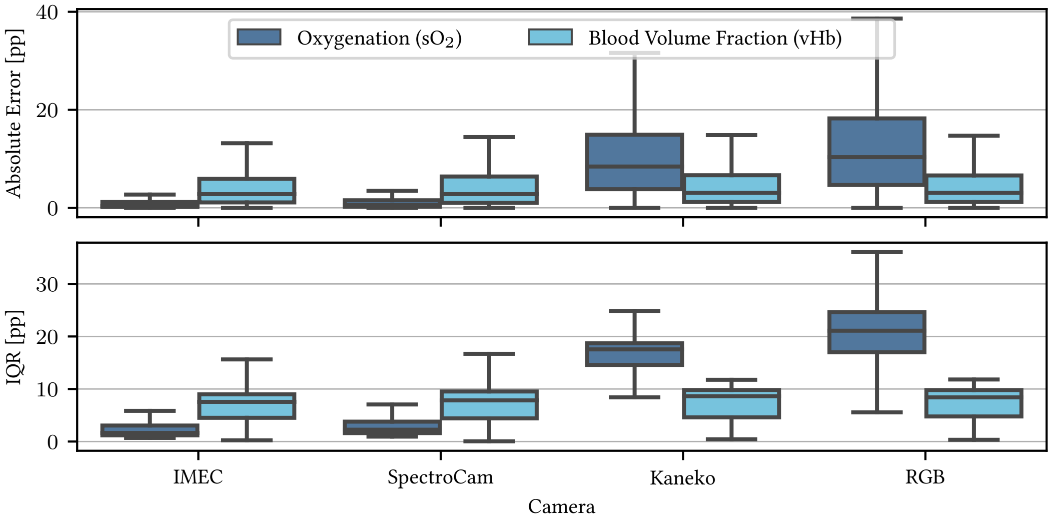


Figure 7.5.: **Comparison of the absolute error and IQR distribution over varying spectral resolution.** The absolute error distribution for sO_2 is lower for higher spectral resolution, while for vHb no effect is visible (first row). The interquartile range (IQR) of the sO_2 posterior decreases with spectral resolution, while the IQR of the vHb posterior seems constant over spectral resolution changes. The median of the posterior was used as the tissue parameter predictor.

The fraction of multimodal posteriors detected using the diptest ($\alpha = 0.01$) can be found in figure 7.6. For IMEC $4.0\,\%$ of the sO_2 posteriors are multimodal, but the median IQR of the multimodal posteriors is 0.9 pp. Kaneko is the only other camera with a considerable fraction of multimodal posteriors ($0.8\,\%$) and median IQR of 11 pp. SpectroCam is the only camera with a non-negligible fraction of multimodal vHb posteriors ($1.3\,\%$) and median multimodal IQR of 0.14 pp.

Figure 7.3 shows example posteriors with the best, median, and worst IQR based on the IMEC camera IQRs. For sO_2, the IMEC and SpectroCam posteriors are located close to the ground truth and very narrow, except for the worst-case example where the posterior is wider, indicating uncertainty, and the location is off. The RGB and Kaneko camera exhibit

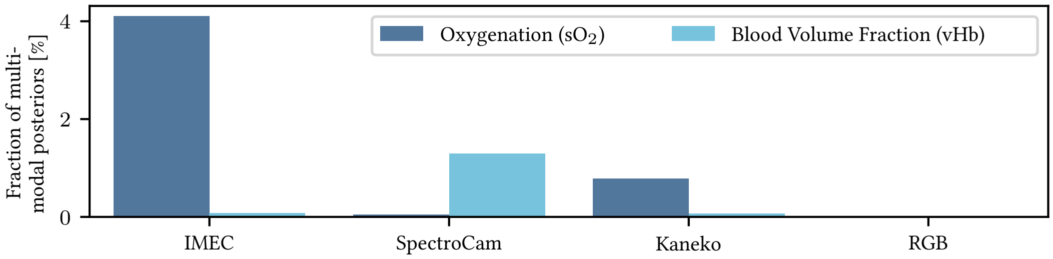


Figure 7.6.: **A low fraction of posteriors was detected as multimodal.** The diptest (cf. section 6.1.2) at level $\alpha = 0.01$ was used to identify multimodal posteriors. Please note that the y-axis stops at $4\,\%$.

wider posteriors throughout, highlighting the decreased performance in line with figure 7.5. For vHb (cf. figure 7.3, second row), we find that only the best posterior is narrow and located well around the ground truth value for all four cameras. In the median and worst case, the posteriors for IMEC and SpectroCam are very wide, approximating the prior distribution and indicating high uncertainty in the prediction. The worst example shows a failure case of the RGB and Kaneko camera, which produced narrow posteriors, but whose location does not correspond to the ground truth.

7.1.4. Discussion

In most cases, the calibration error of the posteriors is very small. However, the calibration curve shows a strong under-confidence for the two MSI cameras, SpectroCam and IMEC, in the sO_2 case. This means that the posteriors should be more narrow. There are two factors contributing to this under-confidence. First, as depicted in figure 7.5, most posteriors are already very narrow, and it seems that the network cannot concentrate the standard Gaussian distribution in the latent space to a distribution with such a small support. The 3-band cameras, Kaneko and RGB, do not suffer from this problem because their predictive performance is worse; hence, a wider posterior is appropriate. The problem worsens because the difficulty of the sO_2 prediction and vHb prediction seems very asymmetric. As seen in figure 7.5, first row, there are large vHb prediction errors for all four cameras. This is reflected in wider posteriors for vHb. However, the network actually learns a joint distribution for sO_2 and vHb. So for the MSI cameras, this joint distribution would be highly anisotropic (high variance along the vHb axis, very low variance along the sO_2 axis), and the network does not seem powerful enough to represent this strong anisotropy. Nevertheless, the results stay interpretable because an under-confidence implies that the predicted IQR is an upper bound for a well-calibrated IQR.

The first row of figure 7.5 shows the benefit of MSI for sO_2 prediction. Both IMEC and SpectroCam can predict sO_2 virtually perfectly in this simulation setting(median absolute error (MedAE) of 0.4 pp and 0.6 pp for IMEC and SpectroCam respectively). The 3-band cameras perform worse and have a MedAE of 8 pp (Kaneko) and 10 pp (RGB) which is too large for medical applications. The prediction of vHb seems more challenging (MedAE $\approx$ 3 pp for all cameras), which is in line with our previous results [Adl+19b]. In particular, higher vHb values seem hard to predict. One reason might be that higher overall concentration leads to more absorption and, hence, a decrease in reflected light, which might negatively impact the signal-to-noise ratio. Another explanation might be that our aggregation scheme for the 3-layer vHb values is too limited such that the network performance deteriorates.

Additionally, we see that the MSI inverse problem seems very well-posed with at most 4 % multimodal posteriors. Even better, these multimodal posteriors for IMEC and SpectroCam should be interpreted as artifacts of the diptest instead of true multimodal posteriors because of the low IQR. Even if they were multimodal, their different parameter predictions would be indistinguishable at the current noise level. Only the Kaneko sO_2 posteriors might be truly multimodal, which could explain the bad prediction performance of the median in this case. However, even in this case, only 1 % of the posteriors are affected.

A peculiarity of this study is that it was only performed on *in silico* data. However, this was necessary as no reference method can provide spatially resolved oxygenation maps of human tissue surfaces. Hence, simulation is the only remaining option for such a comparison study.

Overall, we see how cINNs might be useful in characterizing and choosing cameras for surgical applications. In addition to the regular predictive performance, we have access to uncertainty measures like IQR or the number of modes. This enables a more fine-grained optimization procedure. For example, predictive power could be traded off for well-calibrated uncertainty depending on the application.

7.2. Analyzing the Well-Posedness of Quantitative Photoacoustic Imaging

As established in section 4.2.3, the optical inverse problem in PAI is generally ill-posed. However, most inversion schemes regularize the problem in a way that leads to a single solution. Hence, there was little prior work empirically testing the extent of ill-posedness.

We decided to approach this question using cINNs as they have the potential to encode ambiguities in multimodal posteriors. In particular, we wanted to understand how spatial context influences the uniqueness of the solution. To this end, we performed an *in silico* study where we provided different amounts of spatial context as a conditioning input to the cINN. Because of the *in silico* setting, we have access to ground truth tissue parameter values, which is a single parameter in this study: blood oxygenation (sO_2). This allows for a quantitative comparison of the performance subject to the amount of context of the photoacoustic initial pressure distribution.

In a second step, we tried to understand the generalization capabilities of cINNs from the simulation domain to *in vivo* data. Diminishing this domain gap is an active field of research [Kar+15; FK18; Dre20], so we anticipated a considerable domain gap. However, we expected that for smaller regions or even single pixels, the simulation might be an adequate approximation, whereas on a global scale, i. e. a whole photoacoustic image, we expected a noticeable drop in performance.

The *in silico* part of the experiment addresses research question D.1 in the way that we analyze the well-posedness of the optic inverse problem of PAI using cINNs. The *in vivo* part of the experiment can be seen as a first glimpse at research question D.2, which is concerned with the realism of our *in silico* data.

Disclosure and Contributions The work presented in this section was performed by Jan-Hinrich Nölke in his master thesis [Nöl21b], which he executed under Lena Maier-Hein's and my supervision. Preliminary results of the thesis were published at Bildverarbeitung für die Medizin Workshop (BVM) 2021 [Nöl+21]. Lena Maier-Hein and I came up with the project and developed the thesis scope. Furthermore, we gave regular feedback and discussed the project status. Kris Dreher, Melanie Schellenberg, and Janek Gröhl developed the PAI simulation pipeline SIMPA and supported Jan-Hinrich Nölke during data generation. Jan-Hinrich Nölke implemented all methods, developed the evaluation pipeline, and presented all results.

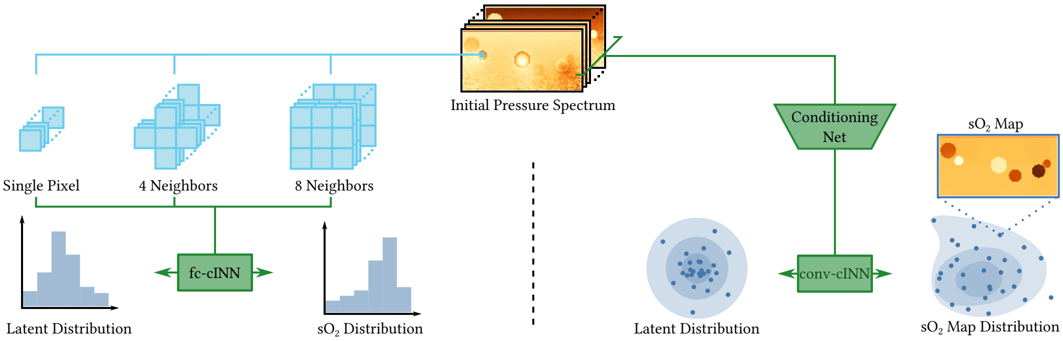


Figure 7.7.: **Experimental setup using different amounts of spatial context. Left:** A single pixel or a single pixel with a local neighborhood (four or eight neighbors) of the multispectral initial pressure image is used to predict the sO_2 posterior of the central pixel. **Right:** The whole multispectral initial pressure image is preprocessed by a conditioning network before being used to generate an sO_2 map distribution. The figure was adapted from [Nöl21b].

7.2.1. Method details

Networks and Training Two families of cINNs were applied for this experiment. An overview of the two settings can be found in figure 7.7. Both families were implemented in PyTorch using the FrEIA framework [Ard+18a].

For small spatial context (a single pixel, four neighbors, or eight neighbors), we used a fully-connected cINN (fc-cINN). We call these the *local models*. The spatial context was flattened and passed directly to the coupling blocks as conditioning input, and each coupling block used shallow fully-connected subnetworks with a single hidden layer with 512 dimensions and ReLU activations. The only tissue parameter of interest is sO_2, but by design, the cINN architecture requires at least two input dimensions. The sO_2 dimension was doubled and used twice to achieve this.

The networks were trained for 30 epochs with a learning rate of 10^{-3} using the AdamW optimizer [LH17] and L2 weight regularization factor of 0.01. The learning rate was reduced by a factor of 10 after epochs 15, 20, and 25. The batch size was chosen as 256, and the norm of the gradients of the network was clipped to 10. We used Gaussian noise augmentation with standard deviation $\sigma = 0.001$ for the normalized sO_2 and $\sigma = 0.06$ for the normalized initial spectra dimensions. In the case of the four and eight neighbors model, the conditioning input was randomly horizontally flipped with a probability of $p = 0.5$.

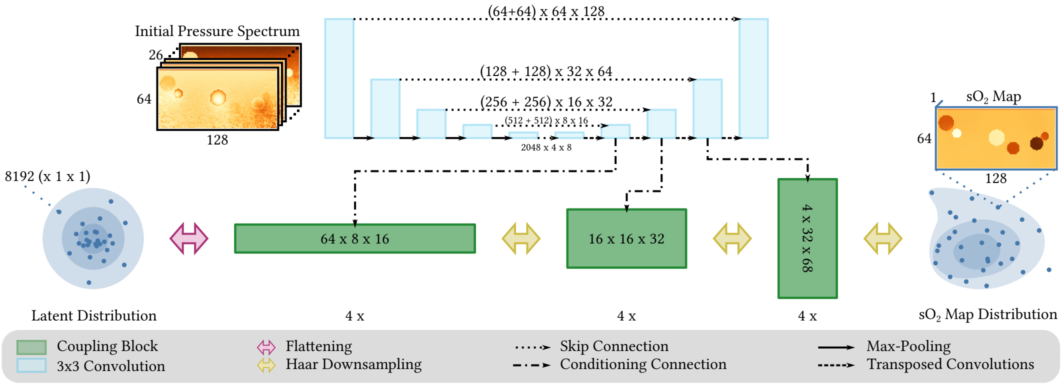


Figure 7.8.: **Architecture of the convolutional conditional invertible neural network.** The conditioning network is a U-Net with 3x3 convolutional layers (turquoise), instance normalization, and leaky ReLU activation. At the correct resolutions, the conditioning input is forwarded to the affine coupling blocks (green). The coupling blocks operate at different spatial resolutions. This is achieved using Haar downsampling layers (yellow), which trade spatial resolution for channel resolution. After the last block before the latent space, the representation is flattened (purple). The figure was adapted form [Nöl21b].

The other cINN architecture (the *global model*) was applied to whole initial pressure images. In this case, a complete sO_2 map for the whole image was generated instead of a posterior for a single pixel, as in the local models. Fully-connected networks are ill-adapted at handling whole images. Hence, we used a convolutional cINN (conv-cINN), where the subnetworks were CNNs made of two blocks consisting of a convolution layer with 128 hidden channels, followed by 2D batch normalization, and ReLU activation, and a final convolution layer to the correct output channel dimension. The kernel size of the convolutions was set to 3×3. We employed Haar downsampling as introduced in section 4.4.3. Thus, coupling blocks on three levels of the spatial resolution were employed before the representation was flattened for the latent space representation. This process is described in figure 7.8.

In contrast to the fc-cINN case, the conditioning input (the initial pressure image) was preprocessed by a conditioning network, which was a standard U-Net [RFB15] adapted to the number of channels of the initial pressure image. The spatial resolutions of the U-Net correspond to the spatial resolutions of the conv-cINN, such that each corresponding resolution was used as conditioning input to the appropriate coupling block. This setup is described in figure 7.8.

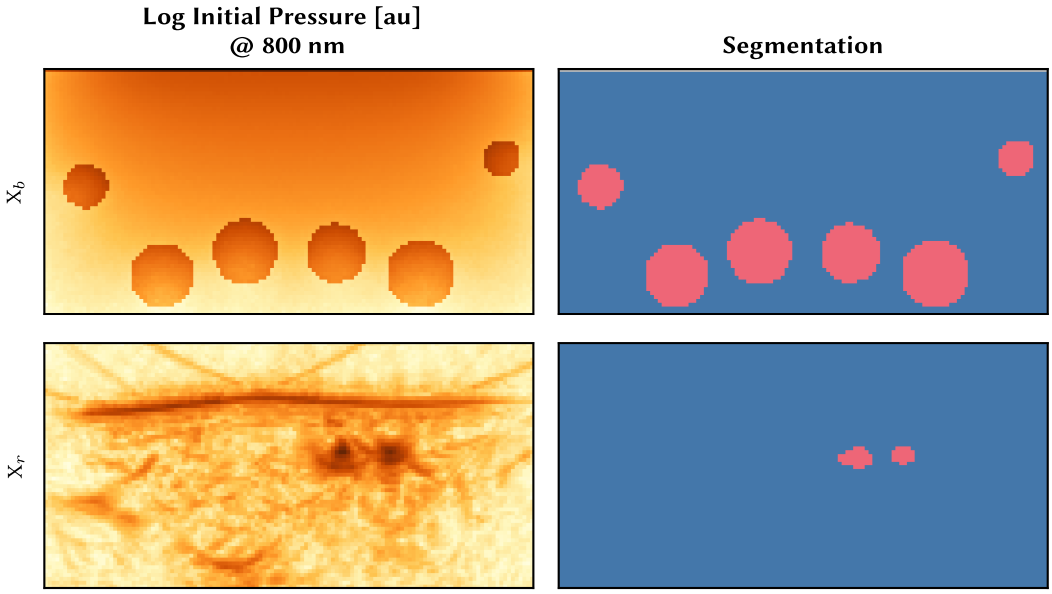


Figure 7.9.: **Comparison of *in silico* to *in vivo* PAI data. Left column:** Logarithm of the initial pressure distribution at 800 nm. The *in vivo* data set X_r exhibits strong noise patterns. **Right column:** Segmentation map with background (blue) and vessels (red). The figure was adapted from [Nöl21b].

The conv-cINN was trained for 600 epochs. The learning rate, optimizer, and L2 weight regularization were the same as for the local models. The learning rate was reduced by a factor of 10 after epochs 200, 300, and 400. The batch size was chosen as 32, the norm of the gradients was clipped to 10, and the individual gradients were clipped to 0.25. We used Gaussian noise augmentation with standard deviation $\sigma = 0.001$ for the normalized sO_2 and $\sigma = 0.12$ for the normalized initial spectra dimensions. Additionally, random horizontal flips were applied with $p = 0.5$.

Multimode Detection To quantify the number of modes in a posterior, we applied the KDE score as introduced in section 6.1.2. We scanned the bandwidth parameter equidistantly in the range of 5 pp and 20 pp at 31 values for each sO_2 posterior.

Mode Localization To compute the absolute error of our predictions, we needed to aggregate the full posterior to a single measure, which should reflect the location of the dominant mode even in the case of a multimodal posterior. As a first rough approximation, we employed the median for this task, as it is efficiently computable and robust to outliers. As alternatives, we located the mode using a KDE and the half-sample method [BF06] (cf. section 6.1.2).

7.2.2. Experiment description

Data Sets We used three data sets for this experiment, two of which were simulated and one *in vivo* data. For the simulation data sets, we modeled volumes with a depth of 22.1 mm and width of 44.2 mm. The third axis had a length of 40 mm, but as the geometry and all other properties were assumed constant along this axis, we will neglect it from further discussion. The voxels were spaced at 0.34 mm. At the top of the volume, the model of the photoacoustic probe was stacked. For more details, please refer to [Dre20].

Each volume comprised three types of tissue: muscle (or background), skin, and vessels. For each tissue type, the chromophore concentrations were drawn from an appropriate distribution, which can be found in table 7.2. The skin was placed at the top of the volume, and the thickness was drawn from a Gaussian distribution with mean 0.22 mm and a standard deviation of 0.1 mm. The vessels were represented by disks[1] with the radius uniformly distributed between 1 mm and 3 mm. The number and placement of the vessels differed for the two simulated data sets as follows.

For the *base data set* X_b, the volume was split into 2 to 7 regions horizontally. The number of regions was drawn uniformly, and all regions had the same width. A single vessel was placed within each region at a uniformly random location. However, each region's outermost 3 mm was excluded to avoid a vessel spilling over in another region.

[1] Or tubes if the third dimension is taken into account.

For the *geometric data set* X_g, the volume was split into three regions horizontally with constant width. First, the radii r_1 and r_2 of two vessels were drawn as described above. Afterward, the distance between the two vessels was drawn uniformly at random between $r_1 + r_2$ and 3 mm. The first vessel was positioned at the center of the region with regard to width and uniformly at random with regard to depth but with the restrictions that both vessels would still fit in the volume if the second vessel were positioned vertically above or below the first at the chosen distance. Then an angle in $[0, 2\pi]$ was chosen uniformly at random to determine the final location of the second vessel.

For both data sets, we obtained 1000 photoacoustic images with 128×64 pixels. For X_b, this led to approximately 530 k vessel pixels and for X_g, to approximately 700 k vessel pixels. 70 % of the data was used as the training set, 15 % for validation, and 15 % was reserved as a test set. The split was performed on the image level, such that all pixels of the same vessels were assigned to the same split. Each simulation was performed at 26 wavelengths equidistantly chosen in the range of 700 nm to 950 nm. The obtained spectra were L2 normalized for each pixel independently. As further preprocessing, the data was z-score normalized (cf. section 4.4.4).

The in vivo *data set* X_r was recorded using the MSOT Acuity Echo system (iThera Medical GmbH, Munich, Germany) on healthy human volunteers. The study was approved by the ethics committee of the medical faculty of Heidelberg University. The reference number is S-451/2020, and on the German Clinical Trials Register the study can be found under DRKS00023205. For each participant, the neck, the forearms, and the calves were recorded. The 26 wavelengths from the simulation were chosen. To obtain the initial pressure spectrum, i. e. to solve the acoustic inverse problem, the DAS beamforming algorithm [GJ82; Kim+16] as mentioned in section 4.2.3 was applied. In a second step, the images were corrected for changes in laser pulse energy at different wavelengths, and five images were averaged to reduce noise. Lastly, the images were resampled to a spacing of 0.31 mm. The vessels in the images were annotated manually, taking a co-registered ultrasound image into account. As preprocessing, the spectra were L2 normalized and z-score normalized using the mean and standard deviation of the corresponding training data set. A comparison of X_r and X_b can be found in figure 7.9.

Posterior Validation The whole evaluation was restricted to vessel pixels. In the case of the local models, the model was only applied to vessel pixels, while for the global model, the prediction was performed on whole images, but none-vessel pixels were discarded afterward. The local models predicted the sO_2 value twice because of the minimum input dimension restriction mentioned above. In [Nöl21b], it was shown that the deviation between the two components is 0.04 pp on average; hence, we simply drop one of the two values.

First, we performed a quantitative evaluation of the simulated data set. We began with the calibration curves and the calibration errors to gauge the reliability of the posterior

Tissue Type	c(H_2O)	c(HbO_2)	c(Hb)	c(Melanin)
Skin	0.68	0	0	$\mathcal{U}(0.002, 0.005)$
Muscle	$\mathcal{U}(0.64, 0.72)$	$\mathcal{U}(0, 1)$	1− c(HbO_2)	0
Vessel	0.68	$\mathcal{U}(0, 1)$	1− c(HbO_2)	0

Table 7.2.: **Chromophore concentrations for the PAI simulation.** c: Chromophore concentration as a fraction between 0 and 1, HbO_2: Oxyhemoglobin, Hb: Hemoglobin (not oxygenated), $\mathcal{U}(a, b)$: Uniform distribution between a and b.

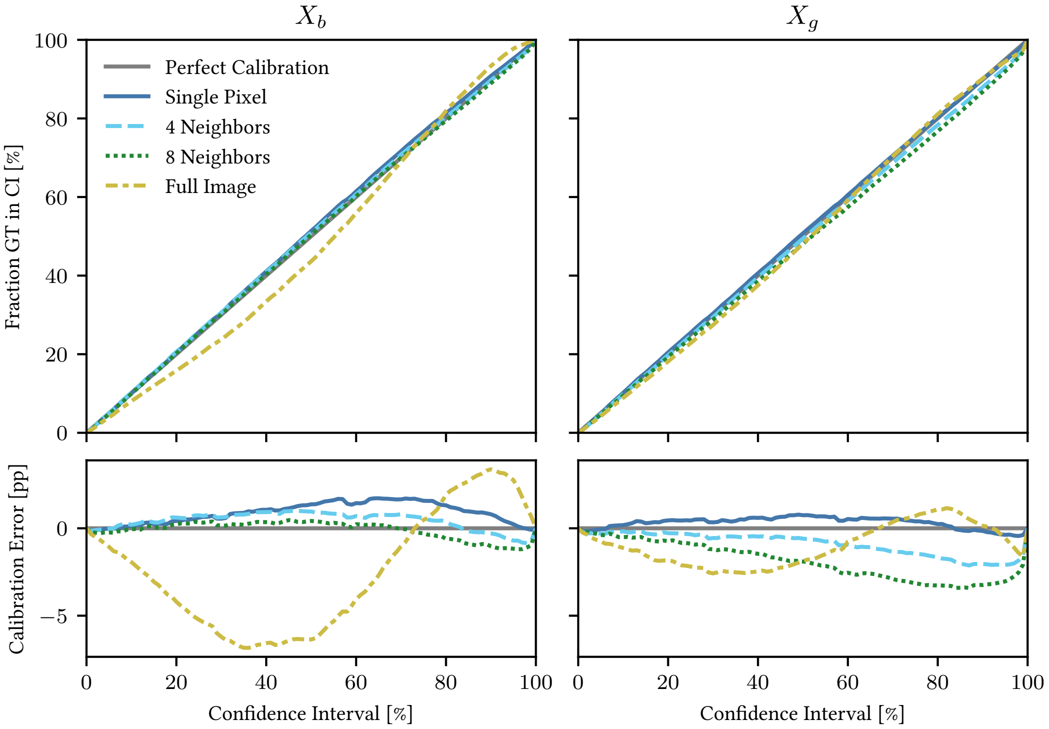


Figure 7.10.: **Independent of spatial context, the models are well calibrated. First row:** For each confidence interval (CI), the fraction of the ground truth sO_2 values falling into the CI of the corresponding posterior is plotted. **Second row:** The Calibration is the difference between the fraction of ground truths (GT) in the CI and the CI. The figure was adapted from [Nöl21b].

width. Then we observed the development of the absolute error with changing spatial context, where we used the mode of the posterior as sO_2 predictor. Lastly, we compared the uncertainty as measured by IQR and ambiguities as measured by the KDE score introduced above.

The data sets exhibit a hierarchical structure: Each volume contains multiple vessels, and each vessel consists of multiple pixels. We respected this hierarchical structure in the evaluation by successively aggregating the metrics, i. e. the absolute error or IQR is computed per pixel, then averaged over all pixels in each vessel, and then this value is averaged for all vessels in the volume. Hence, each data point visible in the jitter plots corresponds to a volume, not a pixel. The only exception to this scheme is the calibration curve/error, where we need access to the full posterior of every pixel.

We perform only a qualitative evaluation for the volunteer study data due to a lack of ground truth data. To this end, we compute posteriors for all vessel pixels and use the median as an sO_2 predictor. Finally, we aggregate all individual vessel predictions per vessel using the mean. Using these values, we check whether the predictions are in physiologically plausible ranges for healthy human veins and arteries.

7.2.3. Results

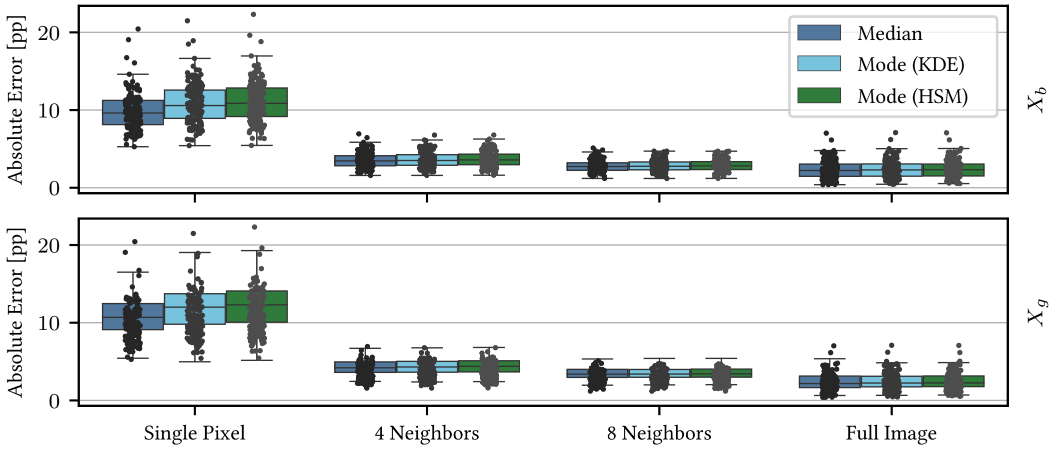


Figure 7.11.: **Increased spatial context reduces error on in-distribution (ID) data.** Absolute blood oxygenation (sO_2) error hierarchically aggregated over vessels and volumes for the base data set X_b (first row) and the geometric data set X_g (second row). The location of the sO_2 posterior is estimated using the median (blue), the mode as detected by kernel density estimation (KDE) (turquoise), and the mode as detected by half-sample mode (HSM) (green). The figure was adapted from [Nöl21b].

Figure 7.10 contains the calibration curves and errors for the two simulation data sets. All models are well calibrated on both data sets, even though they were only trained on X_b. The conv-cINN is slightly over-confident on X_b. Overall, these results imply that the shape of the posteriors is reliable.

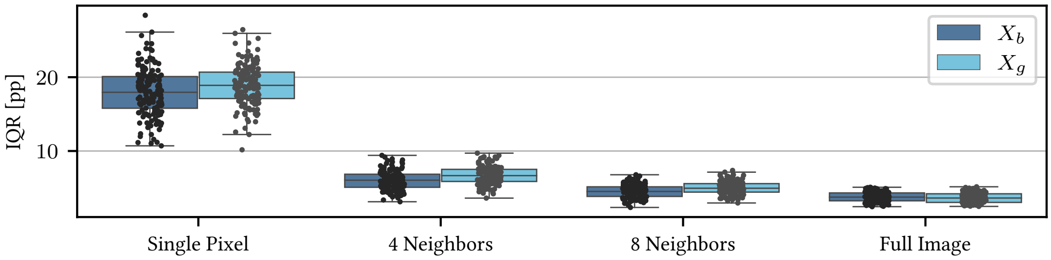


Figure 7.12.: **Spatial context decreases uncertainty.** The IQR of each posterior was first averaged over all vessel pixels and then averaged over all vessels per volume. Results were grouped by the base data set X_b and geometric data set X_g. The figure was adapted from [Nöl21b].

The absolute error drops with additional spatial context for both X_b and X_g as shown in figure 7.11. The largest jump is from single pixel predictions to incorporating the four neighboring pixels. There is no large difference in using the median, the KDE mode, or the HSM mode as a predictor. Only in the single pixel case is there a perceivable difference between the three methods, where the median is slightly better than the other two.

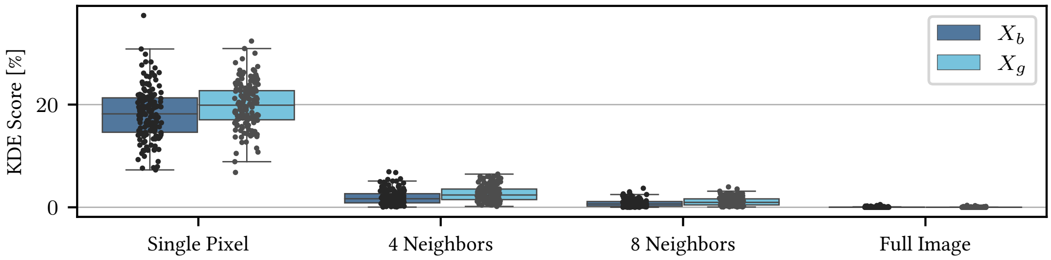


Figure 7.13.: **Spatial context decreases ambiguity.** The KDE score of each posterior was first averaged over all vessel pixels and then averaged over all vessels per volume. Results were grouped by the base data set X_b and geometric data set X_g. The figure was adapted from [Nöl21b].

A similar picture emerges with regard to the IQR of the posteriors, as seen in figure 7.12. More spatial context leads to more confident predictions. For the local models, the IQR on X_g seems elevated compared to the IQR on X_b. For the full image model, this effect is not visible.

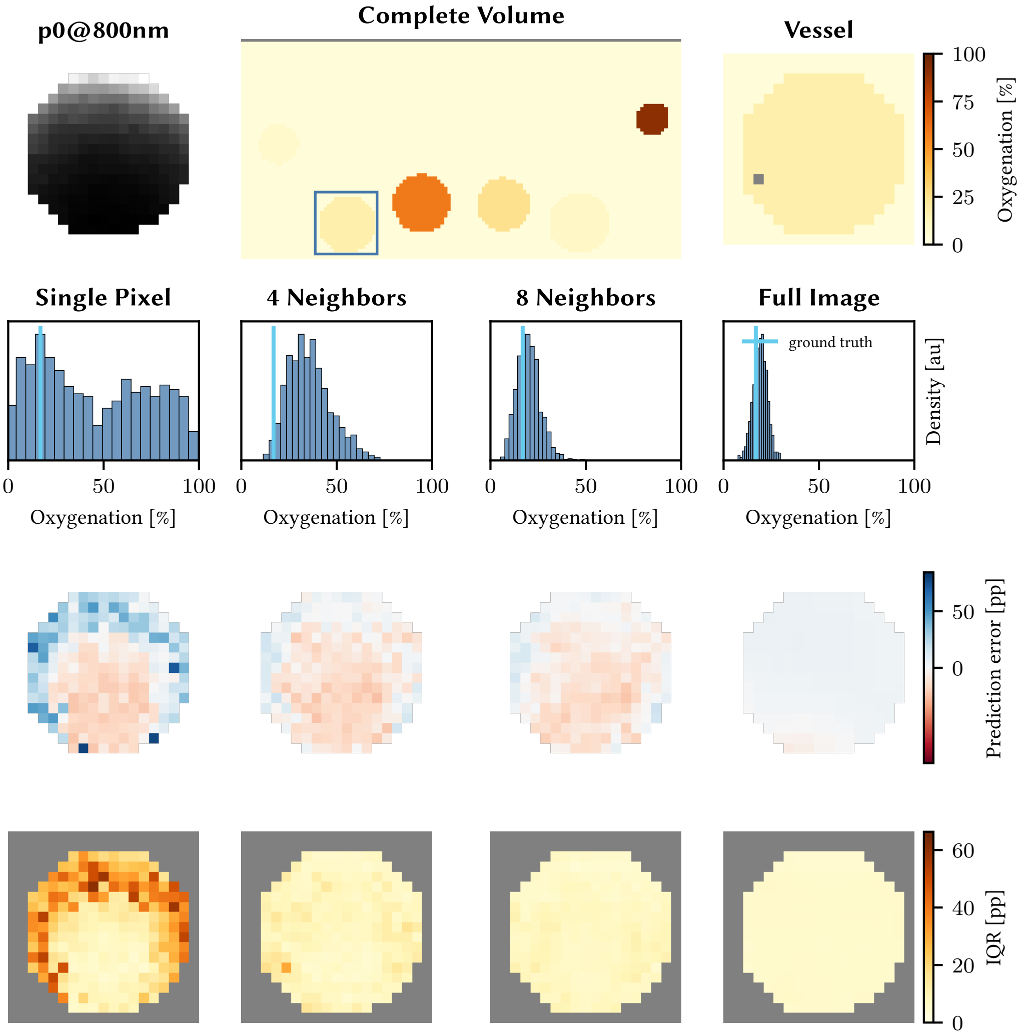


Figure 7.14.: **Example of how spatial context resolves ambiguities. First row:** The oxygenation map of the whole volume and restricted to a single vessel (blue box). The initial pressure (p0) distribution at wavelength 800 nm is also depicted. The gray pixel in the vessel oxygenation map is the pixel for which the sO_2 distribution is shown in the second row. **Second row:** sO_2 posterior of the pixel marked in the first row for the four cINN models. The ground truth is depicted by a turquoise line. **Third row:** Error map of the marked vessel (first row) for the four cINN models. **Fourth row:** IQR map of the marked vessel (first row) for the four cINN models. The figure was adapted from [Nöl21b].

The four neighbors, eight neighbors, and full image model produce almost no multimodal posteriors as measured by the KDE score seen in figure 7.13. While the KDE score for the single pixel model is around 20 %. An example of a multimodal posterior can be found in figure 7.14, second row. As seen there, the posterior for the single pixel model is wide and strongly bimodal, but with added spatial context, the posteriors become unimodal and more narrow. The posteriors are located almost perfectly around the ground truth value for the eight neighbors and full image model.

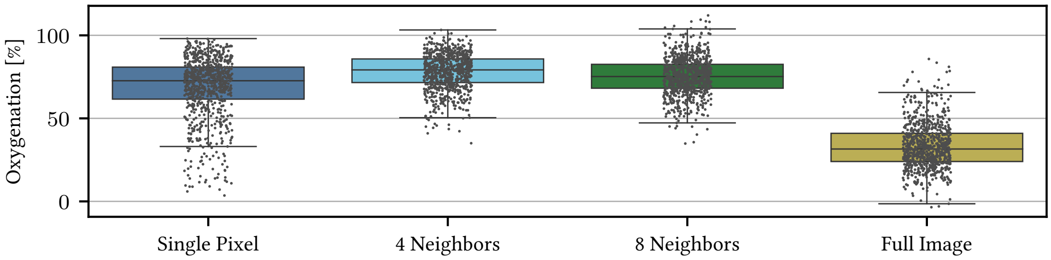


Figure 7.15.: **Predicted blood oxygenation (sO_2) for the *in vivo* data set.** We used the median as the sO_2 predictor on the pixel level and aggregated the results over individual vessels using the mean. The figure was adapted from [Nöl21b].

Figure 7.14, third row, shows the prediction error distribution for a single vessel. For the single pixel model, there is a pattern visible in the error, where a half-moon at the top of the vessel has systematically positive prediction errors[2], while the prediction errors in the center are too low. While less expressive, the same pattern can be observed for the four neighbors and eight neighbors model. The IQR depicted in figure 7.14 fourth row captures this error pattern, i. e. the IQR is elevated in the same regions where the prediction errors are highest.

The *in vivo* results can be found in figure 7.15. The local models predict higher sO_2 values than the global model, whose median sO_2 prediction is below 50 %. However, all prediction distributions are very wide, with the four and eight neighbors model predicting values above 100 % and all three local models predicting values below 60 %. An example can be found in figure 7.16, which shows a volume crop around two vessels. The posteriors of one vessel pixel are shown. We see that all three posteriors are very narrow, implicating a high certainty in the prediction. While the local models predict values above 90 %, the full image model predicts values below 50 % sO_2. The oxygenation map in the first row shows that this behavior is consistent over all vessel pixels.

[2] This means the predicted values are systematically too high.

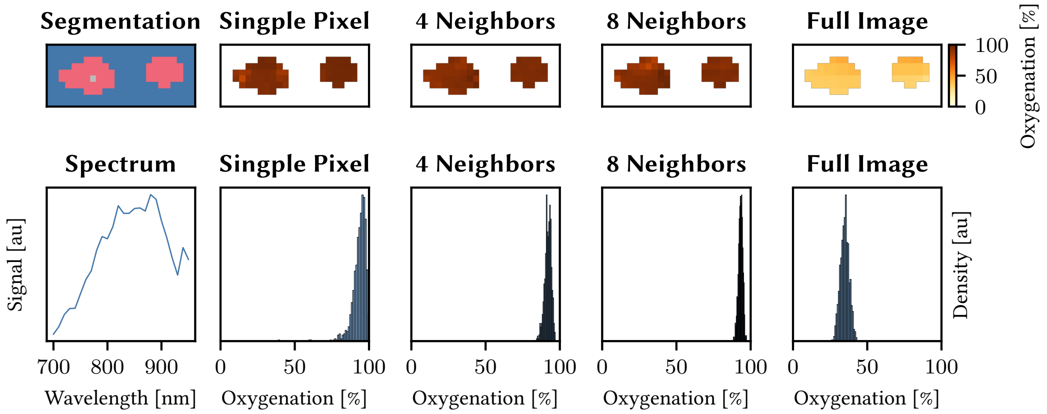


Figure 7.16.: **Example blood oxygenation (sO_2) map and histograms. First row:** Manual segmentation with vessels in red and the pixel, for which the spectrum and histogram are plotted, is shown in grey in the first column. The remaining columns show the predicted sO_2 maps using the median as a predictor. **Second row:** The initial pressure spectrum of the marked pixel and the corresponding sO_2 posteriors. The figure was adapted from [Nöl21b].

7.2.4. Discussion

The findings show that added spatial context increases the prediction performance and the confidence of the cINN prediction. Furthermore, small changes in the domain, as presented in the differences of X_b, on which all models were trained, and X_g were no problem for the models. This is promising as X_g was assumed to be more challenging than X_b as the first does not contain vessels that can shield one another from the probe. However, for larger domain shifts like the shift to *in vivo* data, this is no longer true, as seen on the X_r data set, where many predictions are unphysiological for a healthy human, like sO_2 below 60 % and completely impossible predictions like sO_2 above 100 %. Even worse, the cINN is confident in its (wrong) predictions, as indicated by the narrow posteriors. Hence, the uncertainty score of cINNs trained to solve inverse problems is unsuitable for detecting OoD data.

At first, it might be surprising that the mode predictions for the single pixel model are worse than the median prediction (cf. figure 7.11), while this effect is not observable for any of the other models. However, this observation is cast in a new light under figure 7.13, where we see that the single pixel model is the only model with a substantial proportion of multimodal posteriors. For unimodal posteriors, the difference between the median, KDE mode, and the HSM mode should be very small, but for a multimodal posterior, the difference might be larger, and for the two mode detection methods the distance to the ground truth might be even larger, in cases where the highest mode is the incorrect mode, i. e. the ground truth is closer to a/the smaller mode. As the median's location is less extreme, the error on average is smaller. This can explain the higher absolute error for the KDE mode and HSM mode in the single pixel case.

While not directly visible in the results, a downside of the cINN approach is the necessity of a second input dimension. For the local models, this leads to doubling the sO_2 values, which is undesirable because the constructed distribution is not absolutely continuous relative to the Lebesgue measure on $\mathbb{R}^2$. In theory, this could trip the maximum-likelihood training of the cINNs, but in our case, with the added noise augmentation, the cINN seems powerful enough to overcome this. For the global model, the problem manifests in the form that we cannot apply coupling blocks at full image resolution[3]. Instead, we had to apply a Haar downsampling operation which halves the dimension along each spatial axis. As there are no fine-grained structures in the simulations[4], the reduced spatial resolutions did not seem to impact the performance negatively, but for more realistic data with smaller structures, a negative impact should be expected. There are two avenues to address this issue. First, the coupling block architecture could be adapted. In the conditional case, building an invertible coupling block that can operate on a single input dimension should be possible. Second, we could consider another tissue parameter to increase the input dimension naturally. This is difficult with the current simulation framework as the only other varying quantity is the light fluence Φ (cf. section 4.2), which is not interesting for clinical practitioners in and of itself. In principle, the blood volume fraction (vHb) as in the MSI applications might be a candidate, but since we focus almost exclusively on blood vessels, this quantity should be approximately 1 in all cases. A third alternative might be the speed of sound of the tissue, but this is not sufficiently covered by the simulation at the moment.

Overall, the experiments showed how cINNs could be used to find ambiguities in the optical inverse problem of PAI. Furthermore, we could see how these ambiguities were resolved with higher spatial resolution and how the confidence of the predictions increased. This is especially interesting as we can easily locate ambiguous pixels in the larger volume, which gives us another tool to find patterns that make the inverse problem ill-posed. This might then be the first building block to increase PAI performance by suggesting measurements from additional poses or optimizing the probe shape, as both suggested in our work [Nöl+21]. At the same time, the domain gap to *in vivo* PAI data is too large for the cINNs to generalize to this setting leading to confident but wrong predictions. To overcome this obstacle, future work on both the simulation and the cINN sides is needed.

[3] At least without adding a second channel dimension via doubling of sO_2 or a noise dimension.

[4] The vessels have a minimum radius.

7.3. Gauging the Realism of Synthetic Multispectral Imaging Data

As elaborated in section 4.3.2, OoD detection is inherently error-prone. Especially for high-dimensional data, the curse of dimensionality can lead to prohibitive performance drops. Hence, these three experiments aim to showcase the potential of WAIC computed by an INN ensemble in the setting of MSI.

As a first step, we performed an *in silico* validation. We used simulated spectra and trained on a subset of this data which was chosen based on a threshold of the first principal component. The test set spectra outside this threshold should be detected as OoD. This experiment was performed as a pure sanity check of the method. For the second experiment, we exploited a shortcoming of our simulation pipeline. It assumes that the main chromophores in the tissue are Hb, HbO_2, and if the skin is involved, melanin. However, this assumption is violated for the gallbladder, where bilirubin is another major chromophore. We trained a WAIC ensemble on simulated spectra and evaluated it on multiple porcine organ spectra, including the gallbladder. Our OoD detector should assign higher scores to the gallbladder than to organs closer to the simulation domain. The last experiment analyzed the potential for WAIC to detect changes in lighting during recordings. Most oxygenation quantification methods are sensitive to the chosen light source as it influences the spectrum [Aya+20]. Hence, a method that could automatically notice such lighting changes would be potentially beneficial as it could warn practitioners or disable downstream tasks until the correct conditions are restored. To this end, we recorded a multispectral video of the lip of a healthy human volunteer. During the recording, the light source was switched. The WAIC ensemble was trained on simulated spectra adapted to one of the light sources to detect the second light source as OoD.

Next to validating the general feasibility of WAIC as an OoD score for MSI, the experiments address research question D.2, where we want to gauge the realism of *in silico* MSI data. The light source change experiment can be seen as a very preliminary experiment toward research question D.3, which is concerned with detecting physiological tissue parameter changes (instead of light source changes) via an OoD approach.

Disclosure and Contributions The results of this section were previously published at the MICCAI Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE) 2019 workshop [Adl+19c] as well as a long abstract at Medical Imaging with Deep Learning Conference (MIDL) 2019 [Adl+19a]. Lena Maier-Hein and I developed the experimental setup. Lena Maier-Hein advised me and supervised the whole process. Leonardo Ayala and Anant Vemuri simulated the MSI data. Anant Vemuri and Sebastian Wirkert recorded the porcine organ spectra. Hannes G. Kenngott and Beat P. Müller-Stich performed the surgery on the pigs and operated the laparoscope during measurements. Leonardo Ayala and I recorded the human lips. Lynton Ardizzone, Ullrich Köthe, and

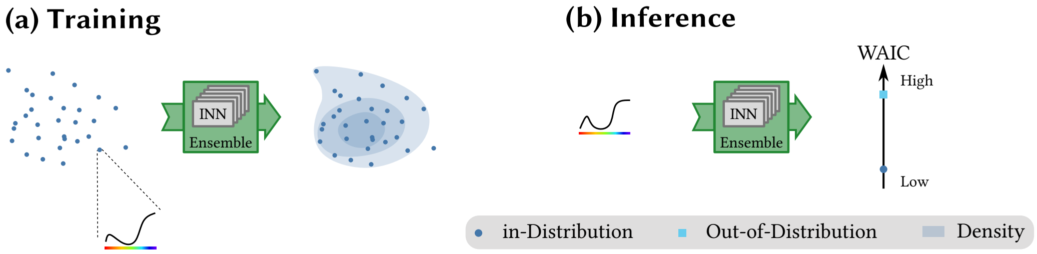


Figure 7.17.: **Training and Inference of the WAIC approach. (a)** The invertible neural network (INN) ensemble is trained on in-distribution (ID) data only to estimate its density. **(b)** During inference, the widely applicable information criterion (WAIC) is computed and used to separate ID from OoD data.

Carsten Rother introduced the INN and gave advice regarding INN implementation details. I implemented all code, performed all experiments, and evaluated all results.

7.3.1. Method details

Training of the WAIC Ensemble For each experiment, we trained an ensemble of INNs consisting of five members. Each member was randomly initialized using the Kaiming initialization scheme [He+15] and trained for 100 epochs using PyTorch and the FrEIA framework [Ard+18a]. Each INN consisted of 20 affine coupling blocks, exponential clamping of 1.0, and global transformation initialized at 0.7. The subnetworks of the coupling blocks were fully-connected shallow NNs with a single hidden layer of width 256 and ReLU activations (see section 4.4 for an introduction of the hyperparameters). For optimization, we used the maximum-likelihood-loss as introduced in equation (4.21) with AdamW as optimizer, an initial learning rate of 10^{-4}, L2 weight regularization parameter 10^{-4}, and AdamW-betas 0.9 and 0.95. If the validation loss did not increase for 15 epochs, we reduced the learning rate by a factor of 10 to a minimum of 10^{-7}. We used a batch size of 2048. The training data was z-score normalized as introduced in section 4.4.4 and Gaussian noise with $\sigma = 0.05$ was applied as augmentation. All other data sets were normalized using the mean and standard deviation computed on the training data. An overview of the setup can be found in figure 7.17.

Evaluation The widely applicable information criterion (WAIC) was computed following equation (6.12). It was computed on an untouched, unseen test set. For far OoD data, WAIC can explode. We computed the natural logarithm of WAIC for easier visualization in these cases.

7.3.2. Experiment description

Data Sets The simulated data set was described in detail in section 7.1. It consists of 450,000 spectra for training, 50,000 spectra for validation, and 50,000 for testing. The *in vivo* data was recorded with two different multispectral cameras: 1.) the Pixelteq (Largo, FL, USA) SpectroCam (8 bands) abbreviated as SpectroCam and 2.) the XIMEA (Münster, Germany) MQ022HG-IM-SM4x4-VIS (16 bands) abbreviated as IMEC. Hence, the simulated spectra were adapted to each camera respectively. As a light source for the adaptation, we used a Wolf (Knittlingen, Germany) IP20 Xenon light source. We abbreviate the data sets $X_{\text{sc}}^{\cdot}$ and $X_{\text{IMEC}}^{\cdot}$, where $\cdot$ indicates the training (tr), validation (va), or test (te) split. For the *in silico* validation, we performed a PCA on $X_{\text{sc}}^{\text{tr}}$ and retained only samples, whose values on the first principal component were below 0.1. This process artificially introduced OoD components in the test data. This reduced the training set size to approximately 337,000 spectra. The validation set was projected to the same coordinates and reduced analogously (Size $\approx 37,500$). These two data sets are denoted by $X_{\text{pca}}^{\text{tr}}$ and $X_{\text{pca}}^{\text{va}}$, respectively. The test set was not restricted such that $X_{\text{pca}}^{\text{te}} := X_{\text{sc}}^{\text{te}}$. A visualization of the validation and test set distribution in the first two principal components can be found in figure 7.18, left column.

The porcine data set was recorded using the SpectroCam and consisted of the organs: abdominal wall, bowel, diaphragm, liver, spleen, and gallbladder. For each organ, there are 200 spectra available. The data set is denoted by X_{pig}.

The lip data was recorded using the IMEC camera and consisted of 100 frames of the lip of a healthy human volunteer. In the beginning, a Wolf IP20 Xenon light source was used for illumination. At approximately frame 80, the light source was changed to a Wolf Endolight LED 2.2 light source. The different stages are depicted in figure 7.20 (c). The lip data set is denoted X_{lip}.

***In Silico* Validation** A WAIC ensemble was trained on $X_{\text{pca}}^{\text{tr}}$ and evaluated on $X_{\text{pca}}^{\text{te}}$ with the aim to test the hypothesis that data points with a low WAIC score should be in the support of the training distribution, while data points with a high WAIC score should be outside the support. To this end, we projected the data using PCA computed on $X_{\text{sc}}^{\text{tr}}$, plotted the density of the validation and training data set and highlighted the location of the test data points with the highest and lowest WAIC score (1 % of $X_{\text{pca}}^{\text{te}}$ respectively). Furthermore, the distribution of WAIC values is compared.

Anomaly Detection in Porcine Organs A WAIC ensemble was trained on $X_{\text{sc}}^{\text{tr}}$ and evaluated on X_{pig}. The distribution of WAIC values for the different organs and the simulated data (using $X_{\text{sc}}^{\text{te}}$) were compared.

Change Detection for Lighting Conditions A WAIC ensemble was trained on $X_{\text{IMEC}}^{\text{tr}}$ and evaluated on X_{lip}. A ROI of 10×10 pixels in a well-illuminated region of the lip was chosen, and the time series of the WAIC values was computed. We expected higher WAIC values after the light source was switched to LED as the simulation was adapted to a Xenon light source.

7.3.3. Results

The results of the *in silico* validation can be found in figure 7.18. The test data points with the highest WAIC value fall into the tail of the test distribution, while the data points with the lowest WAIC value fall mostly into the support of the training and validation distribution. The median WAIC value of the validation set is -1.47 while that of the test data set is -1.36. If the test set is split in the ID and the OoD portion (according to the PCA threshold), the ID median WAIC becomes -1.47 and the OoD median WAIC -0.62.

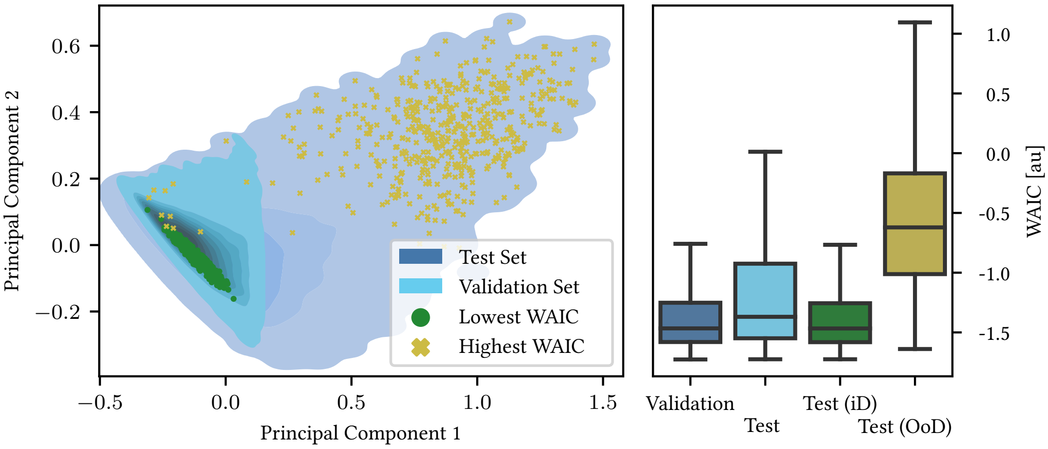


Figure 7.18.: **High WAIC scores are outside the support of the ID distribution. Left:** The 1 % test data points with the highest widely applicable information criterion (WAIC) are in the tail of the test distribution, while the 1 % lowest is at the center of the validation (i. e. the ID) distribution. **Right:** The WAIC distribution of the test set is higher than that of the validation set. If the test set is separated in in-distribution (ID) and out-of-distribution (OoD) data points the ID distribution is in accordance with the validation set, while the OoD portion has higher WAIC values. This figure was adapted from [Adl+19c].

In figure 7.19, we see that all porcine organs have increased WAIC values compared to the simulated data. However, the distance is particularly distinct for the gallbladder with a median WAIC value of 25.9 compared to the simulated test set of -1.6.

An example of the recorded lip data can be found in figure 7.20 (b), where the MSI data was transformed into an RGB image. Figure 7.20 (a) depicts the corresponding WAIC values for the example in (b). The time series in figure 7.20 (c) shows a distinct jump after switching from the Xenon to the LED light source, with peaks at frames 90 and 100. The

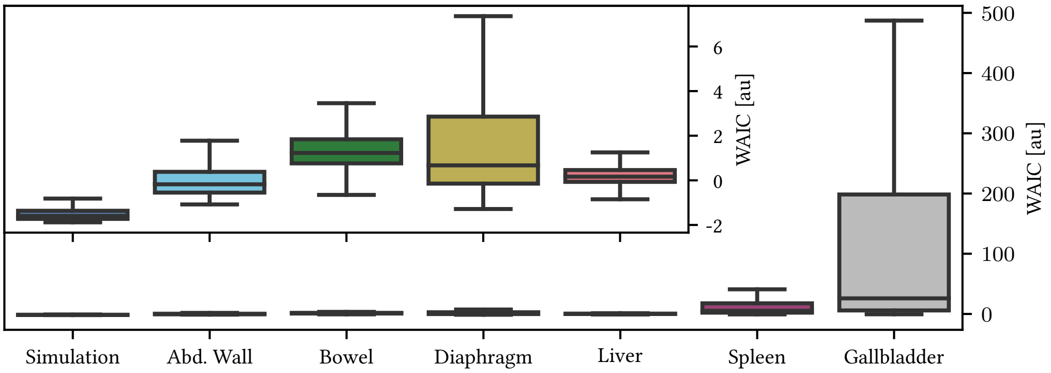


Figure 7.19.: **Distribution of WAIC values separated by porcine organ.** The inset is a zoomed-in version of the box plots for better visibility. The figure is adapted from [Adl+19c]. WAIC: widely applicable information criterion.

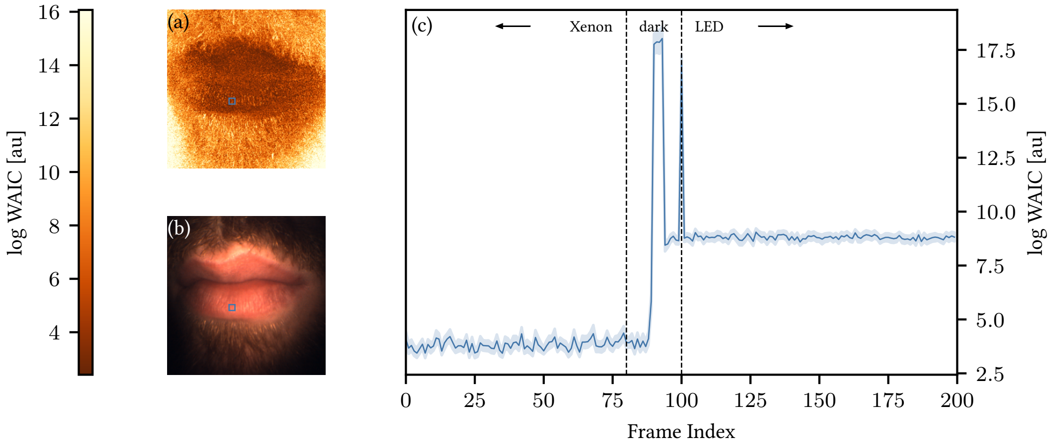


Figure 7.20.: **WAIC time series for the light source change experiment. (a)** Example widely applicable information criterion (WAIC) map with ROI used for (c) inscribed as a blue square. **(b)** Example RGB image corresponding to (a). The RGB image was reconstructed from the multispectral image. **(c)** WAIC time series with values over the ROI aggregated via mean. The confidence band denotes the 95 % confidence interval. Please note the logarithmic values. The figure was adapted from [Adl+19c].

ROI, which was used as a basis for the time series computation, is marked by a blue square in figure 7.20 (a) and (b). Even for the Xenon time range the WAIC values are high with a median of 190 (the median log WAIC is 3.8), but the median WAIC of the LED time range is even higher at $10,200$ (the median log WAIC is 8.8).

7.3.4. Discussion

In all three experiments, we observerd good separation between ID and OoD data using WAIC. This can be seen as evidence that INNs with WAIC are well-suited for OoD detection applications in MSI. For the *in silico* validation the distribution of the ID portion of the test set is almost identical to the validation set, while the OoD portion shows increased WAIC values.

In the porcine organ experiment, the gallbladder is correctly identified as OoD. However, the WAIC distribution of each organ is easily separable from the WAIC distribution of the simulated test set. This indicates that there remains a domain gap between the simulation and the *in vivo* domain. Against our expectation, the WAIC values for the spleen are elevated compared to the other ID organs. As the data set consists of a very small set of spectra, the reason for this is hard to pinpoint. The simulation parameters might not cover the configuration of the spleen sufficiently or analogously to the gallbladder, there might be a missing chromophore. Alternatively, there could have been an issue during measurement, e. g. the camera losing focus or being soiled.

The lip experiment confirms the capability of WAIC to detect changes in the light source, and the LED light source is correctly identified as OoD. The jump during the switch between the Xenon and the LED light source is due to the darkness, which led to a loss of signal in the recorded frames. The peak at frame index 100 is due to a movement of the subject and a subsequent loss of focus of the camera. The WAIC values in the lip experiment are orders of magnitude higher than in the *in silico* and pig experiments, while the values are still small on the test set. This is evidence of a domain gap between the simulation and the measured domain. However, since the porcine data was collected with a different camera, the WAIC values are not comparable. Hence, it is impossible to judge whether the increased values are caused by the change in camera or a worse fit of the simulation to human lip spectra. The higher spectral resolution of the IMEC camera might make it easier for the ensemble to detect OoD data.

While the results are promising, there are caveats connected to WAIC that are worth mentioning. We operated on 8D and 16D data in our setting, which is still rather low-dimensional. As INNs cannot reduce the dimension of the input data, application to truly high-dimensional data will most likely introduce new challenges. Furthermore, WAIC needs an ensemble of INNs. In this offline setting, this was no issue, but for time-critical applications, WAIC might be too slow, or the networks might become too large. To this end, additional experiments to determine a sufficient ensemble size might be insightful to conserve resources.

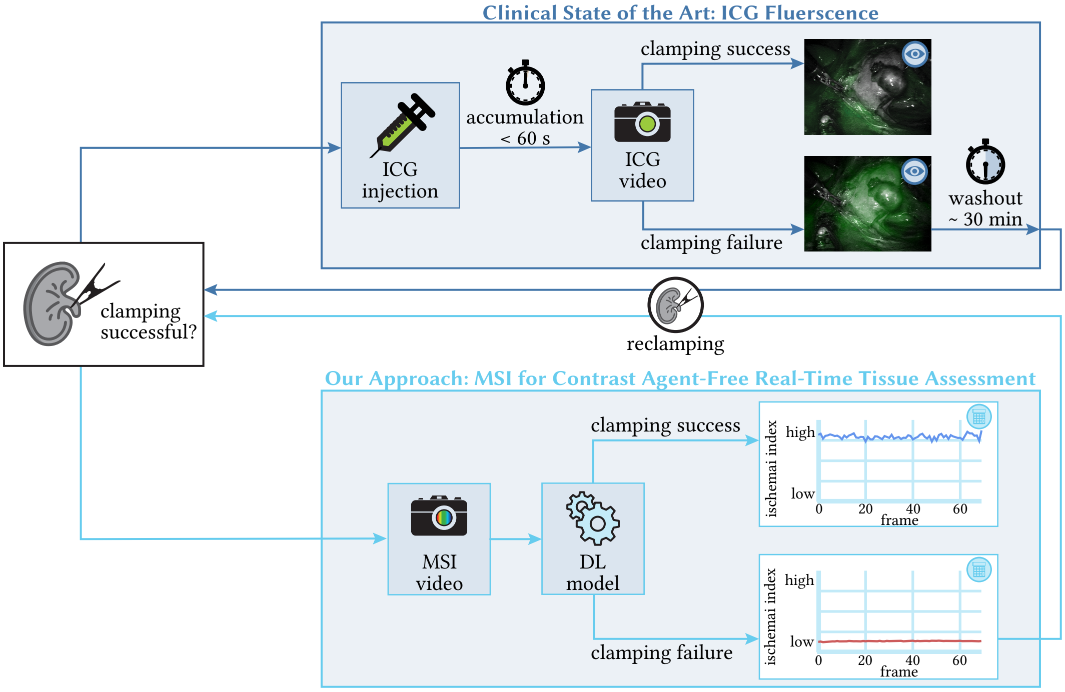


Figure 7.21.: **Comparison of the clinical state of the art to an MSI powered system.** In current clinical practice, ischemia induction via clamping of the renal artery is confirmed using the contrast agent ICG (blue). Successful induction is visible as a lack of fluorescence signal. After a failed induction, a washout period of 30 min is necessary before the technique can be applied again. We propose a new, contrast agent-free, video-rate approach based on MSI and DL (turquoise). The figure was adapted from [Aya+22]. DL: deep learning, ICG: indocyanine green, MSI: multispectral imaging.

Overall, the experiment underlines the power of INNs and WAIC for OoD detection in multispectral imaging.

7.4. Detecting Ischemia in Minimally Invasive Kidney Surgery

Detecting whether an organ is still perfused or ischemic is an important task during many types of surgery [Tho+07; Bor+13; McC+14]. Oftentimes, an organ is artificially removed from the blood supply to avoid bleeding during the actual procedure. This process is called *ischemia induction*. An example of where ischemia induction is beneficial is partial nephrectomy, where a tumor is surgically removed from the kidney. The kidney is removed from the blood supply to avoid bleeding during the tumor resection. The difficulties in ischemia induction arise from the variability of human anatomy. For example, the number of arteries supplying the kidney varies from person to person [LT20]. Furthermore, the arteries are often hardly visible on pre-operative images, so the surgical team has to confirm the actual number during the procedure. In this light, it is nigh impossible to confirm ischemia induction through visual inspection alone, so solutions based on the contrast agent indocyanine green (ICG) are widely spread (cf. figure 7.21). Once the surgical team suspects that all arteries are clamped, ICG is injected. Normally, it accumulates in the kidney on a time scale of seconds. If ICG does not accumulate, the ischemia induction was successful.

However, there are downsides to the use of ICG. The most prominent is the long washout period of the agent (ca. 30 min), which prevents a second application of the method in the same surgery. In other words, if the ischemia induction failed on the first try, there is no possibility of confirming the induction on the second try. Additionally, while ICG is overall safe, there are rare cases of complications (e. g. anaphylactic shocks) [Chu+17; PGM17]. Lastly, ICG requires a specialized camera to make the fluorescence visible. Altogether, there is a case to be made for an alternative ischemia detection method that works without a contrast agent.

In this experiment, which was previously published as [Aya+22], we wanted to demonstrate the benefits of MSI in ischemia detection in a surgical setting. We suspected that the additional spectral information might enable us to detect perfusion changes that are invisible to the human eye. We restricted our attention to minimally invasive partial nephrectomy.

We phrase the task as a binary classification task (perfused/ischemic), but while the final aim is to classify new spectra accordingly, we did not follow the classical supervised learning paradigm but instead focused on OoD detection. The rationale behind this decision is twofold: First, we had to collect all the data ourselves, which limited our data set size. Second, preliminary analyses (shown in figure 7.23) indicated high inter-patient variability in the spectra. Both together highly diminish the generalization capabilities of supervised learning methods to unseen data and patients.

The OoD approach allowed us to train our method on each patient separately, circumventing the inter-patient variability issue. This was possible because we trained on perfused spectra only and detected ischemic spectra as OoD. Nevertheless, this approach comes with its own limitations. Most importantly, we had to confirm that the detected change was truly due to a change in perfusion state and not to some confounding factor (like a change in pose or lighting). We addressed this concern with our experimental setup.

In the following, we will mostly focus on comparing the OoD detection methods as this was my major contribution to [Aya+22]. Naturally, we will introduce all necessary details of the study, but for an in-depth treatise, we refer to the original [Aya+22].

This experiment is aimed completely at answering research question D.3, which was concerned with the possibility of monitoring physiological tissue parameter changes via an OoD approach.

Disclosure and Contributions This section is based on work that was previously published [Aya+22]. Lena Maier-Hein started the work on MSI, supported data interpretation, and suggested the idea of phrasing the ischemia detection task as an OoD task. Lena Maier-Hein, Leonardo Ayala, and I developed the experimental setup and the recording protocol. Leonardo Ayala prepared the checklist for the protocol. Dogu Teber and Lena Maier-Hein initiated the collaboration between Städtisches Klinikum Karlsruhe and German Cancer Research Center (DKFZ). Leonardo Ayala and Alexander Seitel coordinated all legal and ethical aspects of the study from the DKFZ side. Christina Engels, Alexey Aksenov, and Leonardo Ayala coordinated the recording schedule. Leonardo Ayala, Silvia Seidlitz, Jan Sellner, Diana Mindroc, Brittaney Everitt, and I recorded the data. Christina Engels, Dogu Teber, Alexey Aksenov, Matthias Bodenbach, Pia Bader, and Sebastian Baron performed the surgical procedure and operated the laparoscope with the MSI camera. Sebastian Wirkert, Anant Vemuri, and Leonardo Ayala developed the recording software, hardware, and data preprocessing pipeline. Sebastian Pirmann and Leonardo Ayala developed and implemented the tracking algorithm. Leonardo Ayala performed the data preprocessing steps and performed the exploratory analysis on the distribution of the MSI data. I implemented all OoD methods, performed the method comparison, and developed the ischemia index. Manuel Wiesenfarth, Nicholas Schreck, and Annette Kopp-Schneider performed the generalized linear mixed-effects model analysis on the MSI data. Silvia Seidlitz and Jan Sellner created the final data visualizations (for [Aya+22]). Minu Tizabi consulted us on medical considerations and the clarity of our exposition.

7.4.1. Experiment description

Study Design The study was approved by the Landesärztekammer Baden-Württemberg (DE/EKBW01, study reference number: B-F-2019-101), it is in accordance with the Declaration of Helsinki, and it is registered with the German Clinical Trials Register (DRKS-00020996). The data collection took place at Städtisches Klinikum Karlsruhe, Germany. All patients of full age undergoing minimally invasive partial nephrectomy were eligible. Each patient was informed about the study by a doctor and was free to choose whether to consent to it or not.

The recording setup during the procedure can be found in figure 7.22. After the kidney had been prepared by the surgical staff, we inserted the MSI laparoscope in an unused port and recorded a section of cleaned tissue surface for 60 s. This sequence is called $perfused_1$. Afterward, the laparoscope was removed and directly reinserted. The surgeon was asked to locate the exact same pose as previously. A second 60 s sequence of the perfused kidney was recorded, which is called $perfused_2$. The MSI laparoscope was removed, and the surgical team continued with the clamping of the kidney. After the ischemia induction, the MSI laparoscope was reinserted, and a 60 s MSI video of the cleaned kidney surface was recorded. This last sequence is denoted *ischemic*. During the last 15 s of the *ischemic* sequence, the success of the ischemia induction was verified using ICG. The MSI laparoscope was extracted, and the surgery continued as usual.

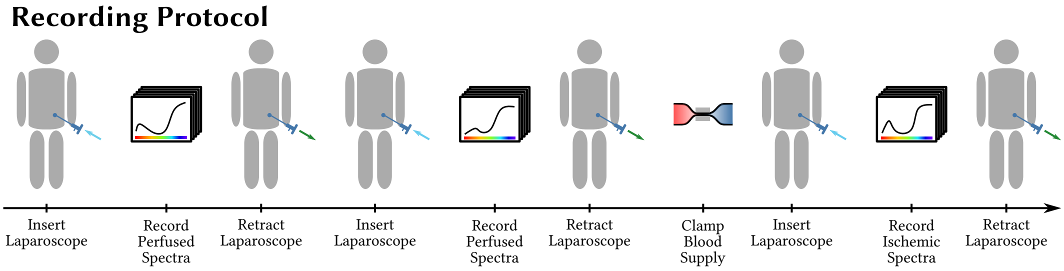


Figure 7.22.: **Experimental setup for the ischemia index validation study.** We recorded three multispectral video sequences with intermittent retraction and reinsertion of the laparoscope. Two sequences of perfused kidney, one of ischemic kidney. The figure was adapted from [Aya+22].

We recorded two sequences of perfused spectra to capture variability in pose and lighting introduced by the surgeon operating a free-hand instrument. This allowed us to confirm whether our OoD approach detects the change in the perfusion state or a confounder.

Based on the recorded images and the ICG signal, we excluded all patients in which the ischemia induction failed and all patients with a bloody kidney surface.

For image acquisition, we used the following hardware:

MSI Camera: MQ022HG-IM-SM4x4-VIS, XIMEA GmbH, Münster, Germany with a spatial resolution of 272×512 pixels and 16 bands

Filter: 335–610 nm band-pass filter (FGB37, Thorlabs Inc., Newton, New Jersey, United States) to restrict the light to the physiologically interesting regions

Laparoscopes: standard 30 ° laparoscope; either 26003BA, KARL STORZ SE & Co. KG, Tuttlingen, Germany or Panoview, Richard Wolf, Knittlingen, Germany

Camera Laparoscope Adapter: C-Mount Adapter (20200043, KARL STORZ SE & Co. KG, Tuttlingen, Germany)

Light source: Xenon (IP20, Richard Wolf GmbH, Knittlingen, Germany)

Recording equipment: Laptop (MSI GE75 Raider 85G, Intel i7, NVIDIA RTX 2080) with custom C++ recording software.

Region of Interest Annotation and Tracking On each kidney, Leonardo Ayala annotated two square ROIs of size 30×30 at the beginning of each recorded sequence. These ROIs were tracked automatically for at least 70 frames using a tracking algorithm based on discriminative correlation filter with channel and spatial reliability (DCF-CSR) [LuN+18] which was adapted to our setting in a master thesis [Pir20]. The ROIs were chosen to be completely located on the kidney surface, in a region without blood, and a region that was well illuminated (at most 5 % of the ROI pixels were allowed to be overexposed). Furthermore, if the tracker was unable to track the ROI for 70 frames, the ROI was discarded, and a new one was chosen instead. Overall this led to six disjoint ROI sequences per patient: Four on perfused kidney and two on ischemic kidney.

Data Normalization To reduce the influence of hardware imperfections, each multispectral image was normalized using a white and a dark image. As the absolute value of the spectrum depends on the light intensity, which fluctuates due to movement, we rescaled all spectra to L2-norm one. Furthermore, the training data set was z-score normalized as introduced in section 4.4.4 on a per-pixel basis. The testing data set was normalized using the mean and standard deviation computed on the training data.

Simulation Data Set for Pre-Training To speed up the training on the patient data, we pre-trained all INNs on simulated data. We utilized the simulated data set introduced in section 7.1 adapted to the IMEC camera used for data collection. We applied the same data normalization as for the patient data, i. e. L2 normalization of all individual spectra and z-score normalization along each band.

7.4.2. Method details

OoD Detection Methods Our proposed method is based on WAIC as introduced in section 6.3, which is computed using an ensemble of five INNs (see section 4.4). As training data, we used the first perfused sequence, i. e. the training data consisted of $2 \cdot 70 = 140$ ROIs. This is too little data to train a INN directly. Instead, we trained the INNs on single pixels, which drastically increased the data set size ($140 \cdot 30 \cdot 30 = 126,000$). However, we had to remove individual pixels which were oversaturated. The training procedure of the INN ensemble can be found in figure 7.17 (a).

At inference time, the log-probability as approximated by the five INNs was aggregated into the WAIC score following equation (6.12). High WAIC indicates an OoD sample, while low WAIC indicates ID (see figure 7.17 (b)).

As a baseline, we used k-nearest neighbor (k-NN) distances. k-NN is based on computing the distance of a new spectrum to the k nearest neighbors in the training data set. The training data set was the same as for the INN ensembles, i. e. the k-NN distances were computed on single spectra. We chose $k = 5$ and used the mean of the 5 distances as an OoD score.

Ischemia Index By construction, we received one OoD score per pixel. In the end, we wanted a single value per frame. This is what we call an *ischemia index*. Let (i, j) denote the spatial coordinates of an ROI then the OoD score (e. g. WAIC or k-NN distance) was aggregated per ROI via

$$\text{score(ROI)} = \text{median}_{(i,j)}[\text{score}(i,\ j)], \tag{7.1}$$

where score denotes any OoD score.

To end up with the final ischemia index, we averaged the two ROIs per frame:

$$\text{ischemia index(frame)} = \frac{1}{2}\left(\text{score}(\text{ROI}_1) + \text{score}(\text{ROI}_2)\right). \tag{7.2}$$

For ease of visualization, the ischemia index was min-max normalized. As this is a strictly monotone transformation[5], the operation has no influence on the area under receiver operating characteristics curve (AUROC) metric.

Implementation Details The INNs and the WAIC computation were implemented in PyTorch using the FrEIA framework [Ard+18a]. Each INN consisted of 20 affine coupling blocks. The subnetworks were fully-connected with a single hidden layer of width 256 and ReLU activations. The INNs were trained using the AdamW optimizer [LH17] with the maximum likelihood loss as introduced in equation (4.21) and a learning rate of $1 \cdot 10^{-4}$ and L2 weight regularization with factor $1 \cdot 10^{-4}$. Each INN was randomly initialized using the Kaiming initialization scheme [He+15]. The training data was augmented using additive Gaussian noise with standard deviation $\sigma = 0.05$. The INNs were trained for 100 epochs on the simulated data as pre-training and fine-tuned on the patient data for ten epochs. For the ablation without pre-training, the INNs were trained for 100 epochs on the patient data. The hyperparameters were determined on pre-existing data from a previous study. Only the number of training frames was determined on this data set, and only patients 1-5 were used to determine this number.

For the k-NN model, we used the scikit-learn [Ped+11] implementation to build the nearest neighbor tree. We used $k = 5$ neighbors, the L2 metric as a distance between spectra, and the kd_tree algorithm to build the neighborhood tree.

Generalized Linear Mixed-Model Analysis We used a generalized linear mixed-effects model to determine the proportion of variance in the spectra, which can be explained by the perfusion state (fixed effect) and the change in patients (random effect). This analysis was performed by Manuel Wiesenfarth with support from Leonardo Ayala. For method details please refer to [Sch19; Aya+22].

Generation of RGB Spectra The MSI camera is unable to record RGB images. Hence, we depended on reconstructed RGB images to gauge the performance difference between MSI and RGB. To this end, we trained a linear regression model between the MSI filter responses and RGB filter responses and used this model to transform the higher resolution spectra into RGB spectra. For more details, please see [Aya+22].

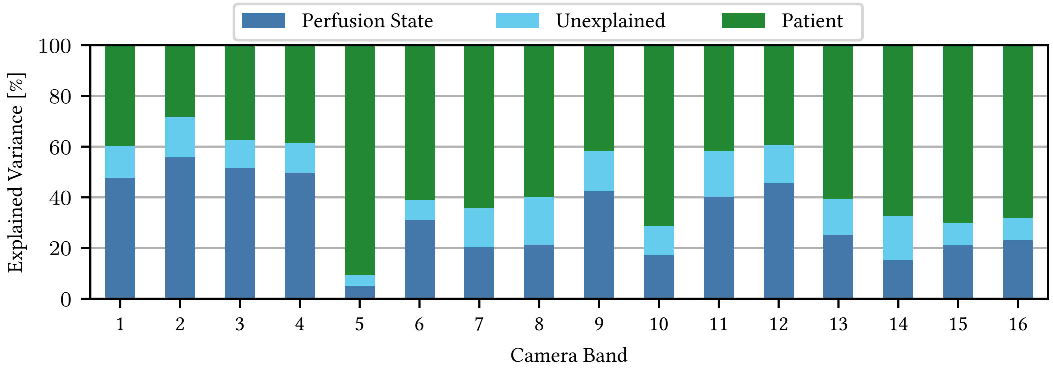


Figure 7.23.: **The perfusion state can explain only a minority of the variance.** The variance was computed using a generalized linear mixed-effects model as introduced in [Sch19]. The figure was adapted from [Aya+22].

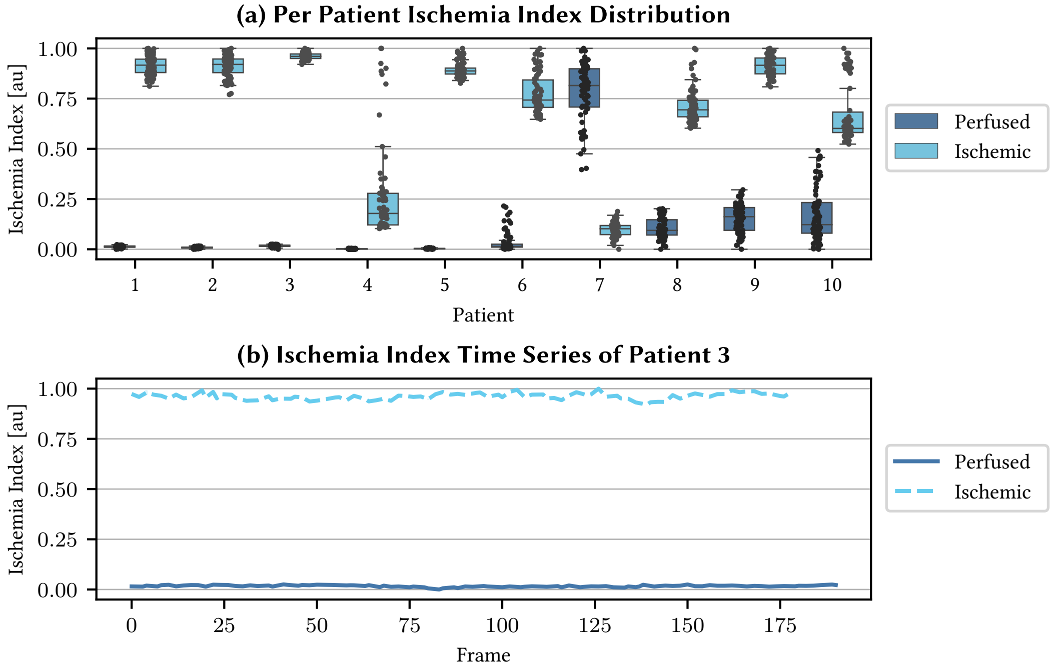


Figure 7.24.: **Ischemia index computed for ten patients. (a)** Ischemia index distribution for the *perfused*$_2$ and *ischemic* sequence of each patient. The distributions are well separated, except for patient 7. For ease of visualization, the index was min-max normalized. **(b)** Example time series of the ischemia index highlights the index's separation over the complete time periods. The figure was adapted from [Aya+22].

7.4.3. Results

Fist application of MSI laparoscope at video-rate in humans To the best of our knowledge, we were the first to bring a multispectral, laparoscopic imaging system to surgery in humans. Central properties of the system are its compact nature (26 × 26 × 31 mm), its small weight (32 g), and its frame rate of 25 Hz. These properties are necessary for easy operation by the surgical team.

Generalized Mixed-Effects Model As can be seen in figure 7.23, for almost all bands, the perfusion state explains less than half of the variance. The only exceptions to this are bands 2 and 3, where roughly half of the variance is explained. On the other hand, the inter-patient variability is dominant in explaining the variance in nine of the 16 bands.

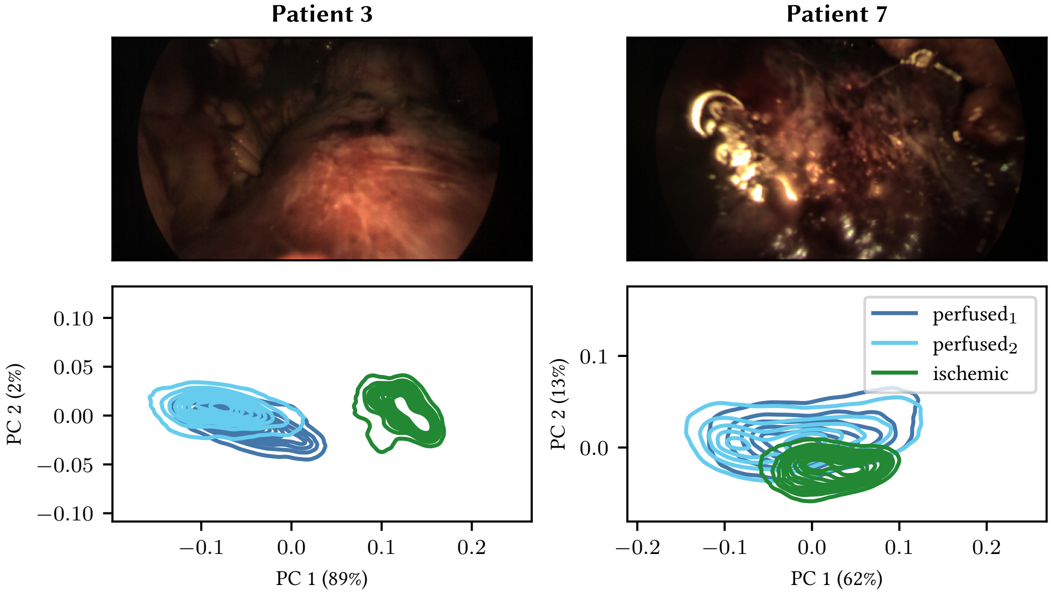


Figure 7.25.: **Spectra distribution comparison between Patient 3 and 7. First row:** Reconstructed RGB images of the two patients taken from sequence *perfused*$_2$. **Second row:** A principal component analysis (PCA) was computed on the spectra of each patient individually. The results were plotted as a density plot using a kernel density estimation (KDE). The explained variance of each principal component (PC) is depicted in brackets at the axis labels.

[5] As long as the score is not constant.

WAIC-based Ischemia Index Performance Figure 7.24 (a) shows the ischemia index distribution (based on WAIC) for all ten patients. Perfused and ischemic frames are well separated, except for patient 7, which is inverted. Figure 7.24 (b) highlights the change of the ischemia index over time in the example of patient 8. The good separation performance is visible in the mean and median AUROC values of 0.9 and 1.0, respectively.

A comparison between a representative patient (patient 3) and the failure patient (patient 7) can be found in figure 7.25. A reconstructed RGB frame is shown for visualization, and the first two principal components of an PCA on the spectral data are plotted. While the spectra from the *ischemic* sequence are clearly separated from the *perfused*$_1$, and *perfused*$_2$ sequences for patient 3, they overlap for patient 7.

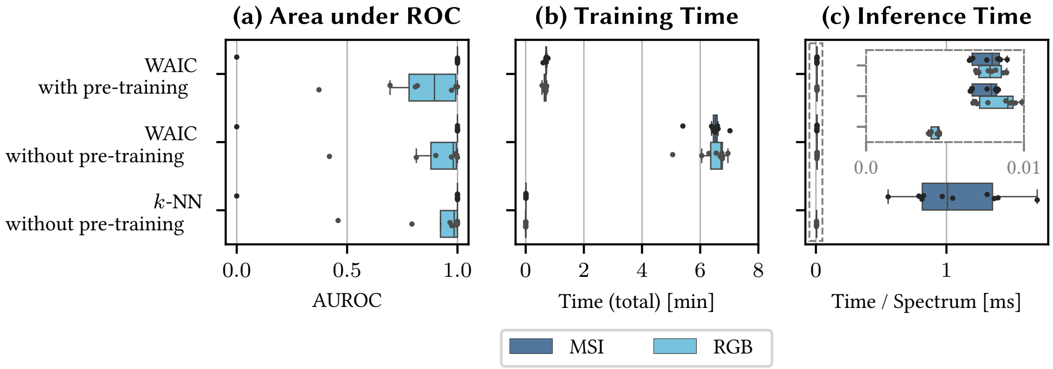


Figure 7.26.: **Different ischemia indices exhibit different trade-offs with regard to accuracy and speed. (a)** Classification accuracy of the methods measured by the area under receiver operating characteristics curve (AUROC) for each patient. **(b)** Total training time for each method and patient. **(c)** Inference time per spectrum for the different patients and methods. The inset magnifies the inference time in the gray dashed box.

OoD Detection Method Comparison Figure 7.26 summarizes the performance of the considered OoD detection methods along the axes of classification performance as measured by AUROC, total training time, and inference time per spectrum. Except for the failure of patient 7, we see that all methods have a perfect AUROC score of 1.0 when using MSI data. For RGB data, the performance drops. k-NN and the non-pre-trained WAIC ensemble are on par and beat the pre-trained WAIC ensemble in terms of median AUROC (0.98 vs. 0.90).

The k-NN training time is almost instantaneous (MSI: 0.4 s, RGB: 0.1 s median training time), followed by the pre-trained WAIC ensemble with 42 s median training time on MSI data and 41 s median training time on RGB data. The WAIC ensemble without pre-training is far off with a per network training time of over 6 min (CPU: AMD Ryzen 9 3900X 12-Core

Processor, RAM: 64GB, GPU: NVIDIA GeForce RTX 3090). The training time is almost equal for the MSI and RGB settings.

For the inference time, which includes only the evaluation of new spectra but not e. g. model loading, k-NN runs distinctly slower on MSI data than all other configurations with approximately 1 ms per spectrum on average. As expected, the inference time for WAIC with and without pre-training are indistinguishable and below 0.01 ms. The k-NN evaluated on RGB data is even faster with an inference time of less than 0.005 ms per spectrum.

7.4.4. Discussion

With a median and mean AUROC of 1.0 and 0.9, the experiment confirms that contrast agent-free ischemia detection is indeed possible. Because of this, multiple applications during the same surgery are possible. As the camera has a recording speed of 25 Hz and the inference speed of the WAIC ensemble per ROI is 140 Hz, our approach is is video rate capable.

The proportion of explained variance underlines the need for our personalized approach, as there is high inter-patient variability. Additionally, a training time of less than 1 min makes it realistic for the models to be trained during surgery.

Figure 7.26 shows that ischemia detection is possible with RGB data too, but with decreased performance. In addition, the RGB data is simulated from MSI data. This fact introduces uncertainties that are hard to quantify. In particular, it is unclear how well our results would transfer to actual RGB images.

While all OoD methods perform almost perfectly with MSI data, there is a trade-off regarding training and inference time. While k-NN trains the fastest, the inference time on MSI data is too slow to be used for live ischemia monitoring. At the same time, training the WAIC ensemble from scratch is potentially too slow during surgery. The sweet spot is achieved with the pre-trained WAIC ensemble, which takes less than 1 min to train and exhibits an adequate inference time.

All MSI methods are successful on nine out of ten patients but fail on patient 7. Counter-intuitively, this is encouraging, as the finding is consistent between different OoD methods indicating that the problem is located in the data and not in the OoD detection method. Indeed, a closer analysis of the data revealed that there is a lot of scarring and fatty tissue at the kidney surface of patient 7. This change in tissue composition, together with the limited penetration depth of the light, seems to overlay the contribution of the perfusion state to the spectra. Interestingly, the methods do not exhibit this strong failure case on the simulated RGB data. One reason for this might lie in the simulated nature of the RGB data. Depending on the exact filter locations, it might be that the parts of the MSI spectrum that differs between perfused and ischemic state is amplified, while the part of the spectrum that makes the ischemic spectra look similar to the perfused spectra might be aggregated and reduced to a single band of the RGB camera. However, it is worth mentioning that the RGB data exhibits outlier patients too. To distinguish between the effect of the RGB reconstruction process and that due to the actual RGB representation, we would need to record MSI and RGB images at the same time. Currently, we do not have the hardware to perform such an experiment, rendering the conclusive identification of the cause of this result impossible.

We would like to point out that patient 4 deviated from our described preprocessing pipeline. This was necessary because the white measurement, used for normalization, was corrupted. Hence, we used the white measurement of patient 3, which used the identical hardware setup. Because of the identical hardware setup, this should have a negligible impact on the results of patient 4. In fact, we examined the influence of using white measurements from other patients, which did not change the performance on patient 4.

Overall, we could show that ischemia detection without a contrast agent is possible, the detection task can be phrased as an OoD detection, and training and inference time is fast enough for use in the operating room (OR). Lastly, there is a clear performance boost of MSI over RGB on average, although MSI fails on one patient. Hence, the experiment highlights the usefulness of INNs in medical OoD detection tasks.

Part IV.

Closing

8. Discussion

This section will discuss the experimental results, methodological peculiarities of the INN architecture, and their impact on the medical domain. To this end, this discussion is separated into two sections. Section 8.1 will discuss technical and methodological aspects of our findings, whereas section 8.2 will discuss the domain aspects.

8.1. Technical Aspects

This section is separated into three parts. Section 8.1.1 addresses all remarks that relate to the (c)INN architecture in general, i. e. to (c)INN properties which are shared by the inverse problem and by the OoD setting. Section 8.1.2 discusses our findings regarding cINNs for biophotonic inverse problems and the open questions regarding the validation of multimodal posteriors. Section 8.1.3 closes this chapter with a treatment of the application of INN ensembles together with WAIC as OoD detectors.

8.1.1. (Conditional) Invertible Neural Networks

The main strength of the INN architecture is the exact $\log p$ computation. This enables maximum-likelihood training and enables INNs to represent arbitrary distributions. This is in contrast to other generative models like VAEs or generative adversarial networks (GANs). VAEs have access to an approximation of $\log p$ via the evidence lower bound, but GANs cannot evaluate the density. This implicit representation is responsible to some extent for the difficulties that GANs exhibit, like mode collapses. INNs do not suffer from this problem because the invertibility (and hence the exact $\log p$) forbids the network from discarding information. At the same time, the generative power of INNs is still not on par with more specialized architectures like GANs. Especially in high-dimensional applications like image generation, the GAN images are sharper and more pleasing to the eye than anything an INN can produce [XYA20]. One reason for this is the curse of dimensionality. GANs and VAEs are not invertible. Hence, they can operate on a lower-dimensional latent space, while the INN latent space has the same dimension as the input dimension. Thus sampling in the INN latent space is inherently less efficient than sampling in the GAN or VAE latent spaces. Since our applications are all in a low- to medium-dimensional domain (1D - 26D)[1], we do not suffer as much from this curse. Still, we need to keep it in mind if

[1] Except for global PAI model with 80,192 dimensions. Still, the evaluation was restricted to 1D marginal distributions.

we want to extend our local, single-pixel models to larger image patches or whole MSI or PAI images. As the PAI experiments in section 7.2 show, we might expect roadblocks, e. g. in the form of a stronger susceptibility to domain shifts.

Theoretical results have shown that INNs can approximate arbitrary diffeomorphisms [Ish+22]. This highlights the architecture's flexibility, and with a large enough training set, we can hope to estimate a large set of data distribution. At the same time, the architecture in the form of the affine coupling blocks and its subnetworks is peculiar, such that it is often difficult to translate new, successful NN tricks to INNs. For example, dropout is a staple operation in most modern NN architectures, and it can be applied to INNs, too, but removing the dropout during evaluation time drastically reduced the performance of the INN in some of our experiments. This is surprising for the INN novice as such behavior seldomly occurs for more widely spread architectures. Similarly, in our experience, INNs and batch normalization require specialized fine-tuning. This all goes to show that while the architecture is very expressive, it is far enough removed from other widely spread NN architectures that a specialized toolkit for "network engineering" and "network optimization" is necessary, such that new users need to perform a cost-benefit analysis whether the unique selling points of INNs outweigh the special needs during architecture development and training.

In this thesis, we have mostly focused on the INN architecture, like the number of coupling blocks or the setup of the subnetworks, but there is another important ingredient when INNs are used as (conditional) density estimators. This is the latent distribution. We fixed it to a standard Gaussian. Any continuous function maps connected components to connected components, and homeomorphisms even conserve the number of "holes" in the connected components [Hat04]. Consequently, an INN with a Gaussian latent space cannot represent a distribution with two or more truly disjoint modes. There is always a small path connecting the two components. For most practical applications, this is of little concern because most real-world distributions do not show completely disjoint modes, but it might still be worthwhile to consider other latent distributions. GMMs can alleviate the mode problem and have been used in previous work [Ard+20]. The trade-off is that a GMM has a fixed number of modes, which we might not be willing or capable of specifying a priori. Another consideration regarding the latent distribution is that the distributions of interest have compact support in our applications. $\mathrm{sO_2}$ and vHb are restricted to the interval $[0, 1]$, but a Gaussian distribution has an infinite support. It might be an interesting inductive bias for the network to work with a distribution with compact support in the latent space. This could potentially simplify training as the INN does not need to concentrate the latent space as much. However, a deep enough network with loose enough exponent clamping should resist this problem, as seen in our experiments.

A last technical consideration is that the current affine coupling block architecture requires at least two input dimensions. For the local PAI applications, this was a problem, as we have only a single quantity of interest ($\mathrm{sO_2}$). We overcame this obstacle by simply doubling the $\mathrm{sO_2}$ dimension and checking that the predictions in both components were consistent. This approach has the problem that a “diagonal distribution” in $\mathbb{R}^2$ is degenerate and has vanishing volume. Hence, the cINN might have trouble representing the distribution, but with suitable noise augmentation, the cINN succeeded. An alternative approach is to add a noise dimension, i. e. a dimension with standard Gaussian noise. This has the advantage that the constructed distribution is absolutely continuous with regard to the Lebesgue measure on $\mathbb{R}^2$ such that it should be easier for the cINN to transform it to the latent distribution. An alternative avenue could be to adapt the coupling blocks. In the conditional setting, half of the channels and the conditioning input are fed to the subnetworks, and then the other half of the channels are (invertibly) transformed. In the 1D case, we could only feed the conditioning input to the subnetworks and transform the 1D input. This operation would still be invertible (as long as the conditioning is provided). At the same time, 1D affine transformations consist of only two parameters (the slope and the bias) which might not be powerful enough for our distributions.

Overall, we see that INNs and cINNs are very versatile (conditional) density estimators. They are naturally well-suited for medium-dimensional data (10s to 100s of dimensions) as present in our biophotonic imaging applications. To tap their full potential, it pays well to be aware of the idiosyncrasies hard-coded invertibility entails.

8.1.2. Representing Solutions to Biophotonic Inverse Problems

The experiments in sections 7.1 and 7.2 show that cINNs are a useful tool to explore uncertainty and ambiguity in biophotonic imaging modalities. When operating on single pixels, i. e. spectra, they lead to well-calibrated posteriors, unless the inverse problem is “too deterministic”, which leads to very narrow posteriors. At some point, the cINN cannot concentrate the latent space enough to achieve this. Especially if other parameters are harder to reconstruct. Thus, we observe under-confident predictions for the multispectral cameras in section 7.1. Still, under-confident posteriors are useful for uncertainty quantification because they give an upper bound on the actual uncertainty. For the MSI cameras the median IQRs on sO_2 were very low with 1.6 pp and 2.3 pp for IMEC and SpectroCam respectively.

In the photoacoustic experiments in section 7.2, we saw that it is possible to apply a cINN to whole images and still yield well-calibrated posteriors, but we also saw that the generalization performance suffered. While the local models (single pixel, four neighbors, and eight neighbors) generalized somewhat from the simulation to the real domain, the global model failed completely. The reason for this is most likely a mixture of a large training set size mismatch and a larger domain gap for whole images than for local models. While a single simulated volume generates 100s of vessel pixels, which make up the data points of the local models, it is still a single volume for the global model. Of course, the images allow for more data augmentation, but this cannot compensate for the two orders of magnitude in data set size difference. At this point, we have not even considered the difference in model size. Additionally, while single simulated spectra seem close to real data, global noise patterns are hard to recreate. Furthermore, for this study, the cINN was trained on the initial pressure spectrum, i. e. we ignored the acoustic inverse problem. However, we cannot record the initial pressure directly, so we needed to apply a reconstruction algorithm to the *in vivo* data, introducing some of the mentioned noise patterns. To alleviate the domain shift, we could either simulate the acoustic forward problem and train on the time series data or reconstruct the initial pressure from the acoustic forward simulation and use that as conditioning input for the cINN. In parallel, improvements to the simulation framework should be pursued like in [Dre20; Grö+22].

In addition to the uncertainty score given by a suitable measure of variability (like the standard deviation or IQR), which are available through simpler architectures [KG17; Grö+18], the posteriors allowed us to examine the well-posedness of the inverse problem by evaluating the modes. This property sets cINNs apart from classical ML approaches. In the MSI experiments, we found higher uncertainty with a lower spectral resolution but few multimodal posteriors for sO_2. For vHb, we found high uncertainty with few multimodal posteriors for all spectral resolutions. In the PAI experiment, we found that there are some ambiguous posteriors for the single pixel model. Furthermore, the ambiguities virtually vanish with more spatial context. This analysis was only possible because of the full posterior distribution.

While the posteriors are the greatest strength of the cINNs, there are still open questions relating to their validation. We are in the unfortunate position that there is only a single reference sO_2 and vHb value per spectrum. In terms of our posterior validation framework, this means that our reference consists of a discrete set of mode locations[2] and the number of reference modes might be incomplete. This strongly restricts the number of available metrics. In fact, next to the number of TP and FP, which would require a suitable localization threshold, we are left with only centroid-based distance metrics between the reference and the posterior modes. These are exactly the metrics we chose for our experiments. So we see that while our reference properties limit the number of available metrics, our posterior validation framework was still applicable to the setting.

So, with the current reference, we can validate if at least one posterior mode is located correctly, but verifying the location of the remaining mode(s) is hard. A remedy could be re-simulation, i. e. simulating new spectra using the reconstructed tissue parameters. However, the cINNs do not reconstruct all the parameters necessary for re-simulation. In the MSI setting, we could include all parameters for the three-layer model, but as was shown in [Ard+18b], most of them cannot be recovered with any certainty. The problem is even more fundamental for the local PAI models. Because of the locality, the model is ignorant of the absolute position of the pixel in the volume. Hence, the computed posterior is implicitly averaged over all pixel locations. For re-simulation, this implies that we would have to try out all pixel locations (including all possible tissue parameter configurations for the remaining pixels of the volume) to test if one of them generates the correct spectrum under

[2] In this case, the set consists of a single element.

re-simulation. This leads to a combinatoric explosion that is computationally prohibitively expensive. In this regard, improving the global models and the simulation pipeline is even more important to enable re-simulation. However, this would require mode detection, i. e. clustering, of a very high-dimensional posterior (80,192D in the PAI case), which introduces its own problems. As we were in the fortunate position that our inverse problems were rather well-posed, we did not have to validate multimodal posteriors, and the unimodal posteriors matched the references well in the case of sO_2 while the vHb posteriors for MSI were very wide at all considered levels of spectral resolution, indicating that reconstruction of vHb might be impossible at the current spectral range.

Our experiments show the value of cINNs as an analysis tool for inverse problems, but at the moment, it remains a tool for theoretical and methodological study. For practitioners, like surgical or medical staff, multimodal posteriors are impractical to interpret and act on. There is a need for further automatic post-processing of the posteriors and extracting valuable insights for the user. This is especially important in the medical domain, where robust uncertainty quantification is key to the safe usage of ML methods. Hence, there are many interesting open questions in the intersection of cINNs and medical imaging, highlighting the potential for meaningful impact of this research area.

8.1.3. Out-of-Distribution Detection for Biophotonoic Imaging

The experiments in sections 7.3 and 7.4 show that INNs in conjunction with WAIC can successfully detect OoD data in an *in silico* as well as in an *in vivo* setting. We observed a gap between the simulation domain and the obtained spectra from porcine organs. This domain gap was especially high for the gallbladder, which contains a chromophore not covered in our simulation (bilirubin). The lighting change experiment on human lips showed that WAIC has potential on human organs as well and that the change could be detected. At the same time, the experiment revealed an even higher domain gap between the simulation and the lip spectra. This is likely due to the lip not being an inner organ and hence containing some amount of melanin, which is also mostly ignored in the simulation pipeline. In addition, we used a different camera for the lip experiment with a higher spectral resolution which might also contribute to the larger domain gap. As a proof of concept, the experiment was still a success.

Our INN ensemble consisted of five members, which seem to be sufficient. We performed preliminary experiments, where we increased the member size up to 20 for the porcine experiments. The WAIC values for the ID organs stabilized for values lower than ten, and the separation effect of ID and OoD data was good for all ensemble sizes. The WAIC values for the gallbladder increased throughout. Hence, higher ensemble sizes might increase the separation effect of WAIC but at the cost of significantly higher computation power and time. This trade-off has to be evaluated for each individual application.

One downside of WAIC concerns its arbitrary scale. WAIC can become negative, and there is no clear lower bound like zero for a (mathematical) metric. Furthermore, WAIC values between different ensembles and/or use cases are not comparable, as the scale depends on the complexity of the ID data and the degree of convergence of the ensemble members. Nevertheless, WAIC introduces an order on the data points of a single task. Even if the exact numerical values are hard to interpret, we can resort to the induced ranks and work with ranking-based methods, like discarding the n % data points with the highest WAIC values. Such an approach can already increase the prediction performance of downstream tasks and harden them against OoD samples. One approach to make WAIC more interpretable might lie in the calibration of the WAIC values on an ID validation data set. A simple approach could be to z-score normalize the WAIC values on that validation set. However, this process might require some care if the WAIC distribution is highly skewed. In addition, we could take the final training or validation losses of the networks into account to normalize the scale.

Motivated by the first encouraging results, we used WAIC's change detection capabilities to detect the perfusion state of human kidneys during minimally invasive surgery. We addressed the domain gap by replacing the simulation completely and training on a single patient's perfused spectra. For nine out of ten patients, we could successfully detect the change in the perfusion state. As discussed in the experiment's discussion section, the failure patient, patient 7, showed a lot of fatty tissue and scarring on the kidney surface. This most likely confused our model. The WAIC computed on reconstructed RGB images did not suffer as much from this outlier, but overall performance was not on par with WAIC built on MSI data. One reason might be that WAIC suffers from the curse of dimensionality on the 16D MSI data, while the 3D RGB data is low dimensional enough to avoid it. Indeed, further analyses using PCA (cf. figure 7.25) and the variance of the spectra indicated that the support of the ischemic distribution was contained in the support of the perfused distribution, leading to the inversion of the WAIC scores.

This leads us to one of the main open questions regarding WAIC. While we found empirical evidence for its usefulness and others found the same [CJA18], there are so far no theoretical guarantees for its effectiveness. In fact, patient 7 can be seen as a failure mode of WAIC. With a lack of theoretical guarantees, empirical validation and analysis of the failure modes[3] become even more important to make sure that ML models can be safely applied in the clinics.

Another observation was that WAIC did not outperform k-NN on MSI data, but they showed similar classification performance. However, the evaluation time of WAIC was orders of magnitude faster. This indicates that the spectral information in the MSI data already leads to a strong perfusion signal such that a rather simple approach like k-NN can pick up on it. However, the evaluation time for k-NN on higher-dimensional data will only grow longer, as they do not scale well, such that the application to cameras with higher spectral resolution or working with a larger spatial context is out of the question. An INN ensemble could be applied to both settings with a negligible drop in speed, as suggested by

[3] As we performed for patient 7.

the minimal inference time change between RGB and MSI data.

While we observed a domain gap between the simulation and *in vivo* measurements, the simulated MSI data still seemed to provide a useful inductive bias, such that pre-training on the simulation and fine-tuning on the patient data led to the same accuracy as complete training on the patient data. This circumstance enabled an acceptable training time during surgery.

There is one more big open question in general when it comes to phrasing ischemia detection as OoD detection. This problem is attribution. By design, OoD detection methods like WAIC only detect whether new data differs from the training data, but they cannot tell us why or how. So, during the design of such a system, we need to be careful to ensure that the change we detect is actually due to the "change of interest". In the very controlled setting of our experiments, we can guarantee this, but for translation in the generally more volatile clinical practice, we have to address that we actually measure a change in perfusion state and not a change in e. g. lighting or pose. This will most likely require suitable representations of the surgical scene that are invariant to the confounding factors.

We have seen the power that exact $\log p$ estimation can bring to the field of biophotonic imaging. We can validate our simulation data, detect changes in the measurement environment, and detect changes in physiological tissue parameters. While there remain open questions before the OoD methods are ready for translation to the clinic, these experiments revealed the utility that these methods can bring to constructing robust ML applications in the medical domain.

8.2. Domain Aspects

This section will discuss the domain impact of our findings. To this end, it is subdivided into two parts. Section 8.2.1 will treat our contributions to uncertainty aware ML methods for biophotonic imaging. This subsumes the treatment of ill-posed inverse problems and the treatment of OoD data points. Section 8.2.2 will discuss our personalized approach to perfusion monitoring.

8.2.1. Uncertainty-Aware Biophotonic Imaging

The representation of ambiguities opens new doors for uncertainty-aware biophotonic imaging. In the previous discussions of our MSI experiments, we have taken the cameras as a given and discussed their influence on the well-posedness of the inverse problem and how the cINNs are capable of representing the ambiguities. However, from a medical perspective, it is more interesting to think about it the other way around and ask what the optimal camera for a certain application is. Generally, cameras with lower spectral resolutions have higher spatial resolutions and/or faster acquisition speeds, which might be necessary if e. g. strong organ movement is expected during the intervention. In such a scenario, the "best" camera would be the camera with the lowest possible spectral resolution, which still solves the inverse problem adequately well. Our framework based on cINNs enables the computation of previously inaccessible metrics to answer this question. With classic ML models, we could only compare the prediction to the reference and evaluate error metrics. cINNs allow the computation of the same error metrics but give us access to additional metrics based on the spread of the posterior (e. g. the standard deviation or IQR) or based on the number and location of detected modes. With the additional metrics, we can generate a more complete picture of the advantages and disadvantages of a certain camera. An open challenge in this regard is the development of robust, automated mode detection algorithms. We have proposed clustering algorithms, but many introduce additional hyperparameters or are restricted to low-dimensional or 1D use cases. A unified framework would be desirable.

One peculiarity of our analysis is that it only used *in silico* data. This is, at the same time, an advantage and a disadvantage. It is advantageous in the sense that we can easily apply it to any recording device for which the filter response functions are available without the necessity of purchasing the camera in question and performing cost-intensive and bureaucratic *in vivo* measurements. This can speed up the turn-around time and might even inform the decision to purchase a certain camera. At the same time, we have to cope with the possibility that the *in silico* data is OoD with regard to the *in vivo* data, i. e. if the simulation pipeline is not realistic enough, there is some uncertainty regarding the transferability of the results to the *in vivo* domain. However, the great strength of our approach is its flexibility so that we can easily adapt our analysis to evolving simulation frameworks, closing the domain gap step by step.

Further evidence for our framework's flexibility can be seen in the ease with which we could transfer it to PAI. While we compared cameras in the MSI setting, we worked with varying amounts of spatial context in the PAI case. Such an analysis can be useful to gauge the trade-off between incorporating more context for a more stable prediction and the prediction speed of the method. For example, we mostly focus on vessel pixels, so it would be desirable to restrict the sO_2 estimation model to these pixels instead of using pixels that are discarded afterward. With cINNs, we can quantify the context amount necessary such that the inverse problem turns well-posed.

Another interesting property of our ambiguity detection approach is its spatial resolution. Instead of predicting whether the inverse problem is ill-posed for a whole volume, we can create "ambiguity maps" highlighting pixels for which the (marginalized) posterior is multimodal. This enables us to use the predictions in regions with high certainty while warning users about regions where the predictions might not be trustworthy. Such an ambiguity map could also be the basis of a method to resolve ambiguities, e. g. by suggesting additional recording poses to avoid shielding by other structures. Furthermore, visualizing the ambiguities and uncertainties can be seen as one way to make ML methods more interpretable which is an important ingredient in increasing trust and acceptance of the models in clinical practice.

While we have focused on biophotonic imaging applications in this thesis, our approach to ambiguity detection can easily be extended to other medical imaging applications, like image registration [Tro+20]. The value proposition in all cases is the detection of potentially ambiguous situations that might spoil the results of classical ML models using point estimates.

Next to ambiguity detection, we have introduced an OoD detection method based on INN ensembles and WAIC. Our experiments show that our method can detect domain shifts between the *in silico* and *in vivo* domain but also changes in the interventional environment, like changes in lighting. These findings highlight the potential of our methodology to be used as a building block for robust medical ML systems.

One core property of our methodology is that it is task agnostic. At training time, we only need to know how ID data looks like without further specification of what the downstream ML system does. At inference time, the OoD detector could then be used to filter data before it can lead to spurious results in the downstream task. This would require no change to the main ML model. If a global change in the recording condition, like a change in lighting, were detected, the OoD detector could go one step further and warn the practitioners about a continued input data mismatch. Instead of a silently failing system, the surgical team would get the chance to troubleshoot. To stay with our lighting example, this could lead to the surgical staff identifying an erroneously turned-on lamp that would otherwise have invalidated the network predictions.

Overall, we see that INNs have great potential as enabling factor for robust ML in medicine.

8.2.2. Personalized Perfusion Monitoring

The proposed OoD detection approach to perfusion monitoring presents a new paradigm in MSI. We avoid physics-based models, like e. g. the Lambert-Beer law, that are not sufficiently realistic while simultaneously being able to train our model in a very low data regime, i. e. on a single patient. This allows us to overcome the observed high inter-patient variability and lets us avoid confounders due to different patients [ZAP20; Die+21]. Interestingly, our findings of high inter-patient variability are in contrast to observations in porcine organs [Stu+21], where the inter-specimen variability in the MSI spectra is lower. Since the patient cohort is rather limited, it is difficult to find the cause for the changes in spectra. Indeed, the cancer type with the comorbidities is sufficient information to identify a patient in our cohort. At this point, we have not even considered confounders due to the complex surgical environment, like the free-hand operation of the laparoscope, which introduces variation in the poses. We can use data normalization to reduce the effect of some confounders, but overall, the extent of the possible analysis stays limited. In this regard, a larger patient cohort would be desirable, but the scientific benefit has to be weighed against bureaucratic and, more importantly, ethical considerations when using prototype equipment in the OR.

However, independent of the cause of the high inter-patient variability, its mere existence underlines our personalized approach's advantage. In particular, in the light of recent findings suggesting that ML performance is often overestimated due to biased test data selection [Rob+21; Sha+21]. One disadvantage of the personalized approach consists in the necessity to train the INN ensemble on each patient. However, with suitable pre-training, the training time can be reduced to less than 1 min per network (five networks, training can be parallelized). This training time is short enough to avoid delays in the surgical workflow. The training could, for example, be performed after the kidney's preparation during the localization of the renal artery. After training, the proposed method can perform inference at video rate with an ROI inference rate of 140 Hz.

Currently, the success of ischemia induction in partial nephrectomy is confirmed using ICG fluorescence. Our approach has the decided advantages that it does not require a contrast agent, i. e. it is non-invasive, it is video rate capable, and it can be performed multiple times during the same procedure. A drawback of our approach is that it currently requires a clean kidney to work reliably. In fact, blood, scars, or fatty tissue on the kidney's surface can lead to our approach's failure. We hope to overcome this limitation by using multispectral cameras that are sensitive in the near-infrared region of the spectrum, as this range offers deeper tissue penetration. However, such recordings are out of scope for our current equipment, both with regard to the camera and a suitable light source.

We want to stress that while there is previous work using tissue parameter estimation-based approaches to detecting ischemia [Wir+17; Aya+19], we resorted to the OoD detection-based approach because these models failed in preliminary experiments on the (diseased) human kidney spectra. As mentioned previously, we require MSI data with annotated tissue parameters to train such estimation models. As a gold standard method for annotation of *in vivo* data is missing, we are restricted to *in silico* training data. Further analysis showed that the *in vivo* spectra were OoD with regard to the *in silico* training data.

The proposed approach worked on nine out of ten patients. However, it failed on patient 7. We conducted further experiments and found that the spectra for this patient did not markedly differ between the perfused and ischemic stages. In fact, a KDE on the first two principal components showed a large intersection of the two stages (cf. figure 7.25). A second peculiarity was that the perfused spectra showed a higher variance than the ischemic spectra. Technically, these findings suggest that the support of the ischemic spectra lay within the support of the perfused spectra, which might explain the misclassification. As for the reasons why the distributions overlapped so much, we suspect a suboptimal kidney surface. The patient exhibited a lot of scarring and burned fatty tissue on the kidney surface. This tissue composition seems to have overridden the signal due to the change in the perfusion state. It is further worth mentioning that patient 7 was the only smoker in the cohort, but with the limited cohort size, we cannot gauge whether this caused the failure. Lastly, we would like to point out that our clinical collaborators described the video of the kidney as "unusually looking", which might be taken as soft evidence that patient 7 is an outlier in our cohort.

Overall, our approach performed exceptionally well considering the overly complex surgical setup. In particular, we needed to retract our laparoscope in between the acquisition of the perfused and ischemic spectra. The retraction was necessary because the port for the MSI laparoscope was needed during regular parts of the surgery, and we wanted to keep the changes to the intervention as non-invasive as possible. However, our evaluation became more difficult because we could not guarantee to identify the same ROIs for the different recording stages. In fact, the free-hand operation of the laparoscope led to changes in the pose at each stage, but also artificial relocations of the kidney due to the clamping of the artery led to a change in view. With a lack of landmarks because of the small field of view, it was even hard for a human to identify the same regions for some of the patients. To make sure that the detected change was actually due to a change in perfusion state and not due to a change in pose, we acquired two perfused sequences, $\textit{perfused}_1$, and $\textit{perfused}_2$, with re-insertion of the laparoscope in between. We trained on $\textit{perfused}_1$ and evaluated on $\textit{perfused}_2$. If the detected change had been due to a change in pose, our method should have failed on the $\textit{perfused}_2$ sequence. However, as mentioned before, there was a clear separation between $\textit{perfused}_2$ and the *ischemic* sequence for nine out of the ten patients. Overall, the results might be seen as a lower bound for the actual performance of a clinically certified system which would simplify the tracking between stages as the re-insertion could be avoided.

9. Conclusion

This thesis presented a framework for uncertainty quantification for inverse problems which builds upon INNs as core components. We used this framework to address important domain-related research questions. We successfully analyzed the well-posedness of inverse problems in a data-driven way with the help of cINNs. Furthermore, we introduced a new method to examine the realism of our simulation framework based on OoD detection, which used INN ensembles and WAIC as OoD detectors as central ingredients. Lastly, we introduced a personalized approach to tissue parameter monitoring, circumventing the challenges due to inter-patient variability. This approach introduced a new paradigm for medical ML formulating change monitoring as another OoD detection problem, with the current tissue parameter configuration interpreted as ID. In this way, we were the first to apply a laparoscopic MSI system in human surgery, where we applied the proposed tissue parameter monitoring approach to the task of perfusion monitoring during partial nephrectomy.

In the following, we will discuss this thesis's contribution in more detail. In section 9.1, we will circle back to our research questions T.1 – T.3 and D.1 – D.3 and note our progress toward answering them. Please recall that solutions to T.1 – T.3 are methodological contributions that are, in principle, broadly applicable across modalities and domains. In contrast, D.1 – D.3 are specific to the field of medical biophotonic imaging with potentially high domain impact. In section 9.2, we will continue with an outlook of open challenges and opportunities.

9.1. Summary of Contributions

This section's structure mirrors chapter 2, where the research questions were introduced. Here, we will collect the progress toward answering the research questions and reference corresponding publications.

T.1. How can we encode inherent ambiguities in inverse problems?

We were the first to analyze biophotonic inverse problems using posterior distributions over the tissue parameter space generated with cINNs. We successfully examined the influence of the spectral resolution on the well-posedness of the inverse problem at the example of MSI (section 7.1) and the influence of spatial resolution on the well-posedness of the inverse problem at the example of PAI (section 7.2). Our experiments highlight that cINNs are powerful enough to learn biophotonic inverse problems and that they are capable of encoding uncertainty in general and ambiguities in particular in their posteriors. Overall, this underlines the potential of our proposed framework and INNs for the analysis of ambiguities in inverse problems, not only in biophotonic imaging but beyond the frontiers of the medical domain.

The results on analyzing ambiguities in MSI using INNs led to a publication [Adl+19b] at IPCAI and an associated journal publication at the International Journal of Computer Assisted Radiology and Surgery (IJCARS). My short presentation received the audience vote for one of the interventional imaging session's two best presentations. My work on tissue parameter estimation in MSI also contributed to a patent application [Mai+18a]. My initial work on ambiguities in biophotonics was extended to PAI by Jan-Hinrich Nölke in his master thesis [Nöl21b] under Lena Maier-Hein's and my supervision. Parts of the results were published at BVM [Nöl+21]. Jan's latest manuscript, with me as second author, on well-posedness in PAI, is currently under review at IEEE Transactions in Medical Imaging. Next to biophotonic imaging, we have extended our cINN approach to C-arm pose estimation. This work was led by Darya Trofimova and supported by me and culminated in a MIC meets Conference on Neural Information Processing Systems (NeurIPS) workshop contribution (preprint of the work presented at the workshop [Tro+20]). A follow-up to this work led to a master thesis by Sebastian Gruber [Gru22] under Lena Maier-Hein's and Jan-Hinrich Nölke's primary and my secondary supervision.

T.2. How can we validate posteriors?

We were the first to propose a posterior validation framework taking key properties of the inverse problem and the available reference data into account to end up with the best set of metrics. To this end, we discovered and expanded an analogy between the task of object detection in computer vision and the process of validating posterior modes. We successfully transferred and extended object detection metrics to the posterior validation setting. Overall, we found that the extent to which a posterior could be validated strongly depended on the quality of the reference data. In particular, it is very hard to validate multimodal posteriors if the reference data is ignorant of ambiguous solutions, which is often the case if the reference data is generated using a simulation of the forward problem. Nevertheless, our proposed framework covers these cases and will provide a potentially smaller but suitable set of metrics.

This thesis is the first time that we revealed this framework to the public. We are currently working on a manuscript to introduce a wider audience to our results.

T.3. How can we detect out-of-distribution data in biophotonic imaging?

We were the first to apply INN ensembles in conjunction with WAIC as an OoD detector to SI. We successfully validated the OoD capabilities in an *in silico* setting and applied the detector to porcine and human data to examine the realism of our simulation framework. Lastly, we displayed the strength of our proposed OoD methodology by using it to detect physiological parameter changes in an unsupervised manner during minimally invasive kidney surgery. To this end, we addressed challenges due to small data sets which required training on single pixels instead of whole multispectral images. However, this solution increased the noise in our predictions, which we tackled with suitable post-processing steps. Overall, we observed a strong separation between ID and OoD data using (post-processed) WAIC as the score. We compared our methodology to a baseline method based on k-NN and found similar classification performance. However, k-NN does not scale well to higher dimensions, such that the inference time was orders of magnitude slower than that of the INN ensemble with WAIC. At the same time, with suitable offline pre-training, the INNs exhibited an adequate training time for live training during interventions. All in all, our results highlight the fit of OoD detection based on WAIC for SI.

Our results analyzing the realism of synthetic MSI data using our OoD approach were published at the MICCAI UNSURE workshop [Adl+19c], where I received one of three oral presentation spots. Our overarching framework joining OoD detection with ambiguity detection was published as a long abstract at MIDL [Adl+19a]. Furthermore, I supported David Zimmerer during the MOOD 2020 challenge, which aimed at introducing a common benchmark for OoD detection in medical imaging[1]. The challenge results were published at IEEE Transactions in Medical Imaging [Zim+22]. Our work on ischemia detection in minimally invasive kidney surgery led to a manuscript that is currently under a second round of revision at Science Advances. Leonardo Ayala and I are shared first authors. A preprint of the manuscript can be found on medRxiv [Aya+22]. In addition, our results were accepted as a long abstract at IPCAI 2022. Our OoD-based framework to change detection of physiological parameters led to a patent application [Mai+18b].

D.1. Are inverse problems in biophotonic imaging well-posed?

We were the first to analyze the well-posedness of biophotonic inverse problems over varying spectral context (for MSI) and spatial context (for PAI) using our data-driven framework based on cINNs. We focused on the recovery of sO_2 for MSI and PAI and of vHb for MSI and adapted the cINN architecture to operate on such low-dimensional spaces. We found that sO_2 estimation is well-posed in most cases. Only with minimal spatial context (i. e. a single pixel) did we find a perceivable amount of multimodal posteriors in PAI. Besides that, we found the expected increase of certainty (measured by the posterior IQR) in the prediction with increasing spectral and/or spatial context. The problem of vHb estimation in MSI was found to be ill-posed at all provided levels of spectral context. However, the ill-posedness did not lead to multimodal posteriors but to very wide posteriors, indicating that the cINNs could not narrow down the solution space at all. All the results

[1] With a focus on CT and MRI.

were produced on *in silico* data because of the lack of a gold standard reference method for *in vivo* data. This limits the expressiveness of our results. Nevertheless, the very flexible nature of our framework allows for its easy application to new data sets, such that the analysis can keep up with evolving simulation frameworks. If a gold standard reference method became available, we could easily apply our framework to that setting, too, with no need for modifications.

The dissemination of our contributions toward research question D.1 overlaps with the dissemination of our contributions toward research question T.1 and can be found in that paragraph.

D.2. Are synthetic spectral images sufficiently realistic?

We were the first to propose an OoD detection approach based on INN ensembles and WAIC to gauge the realism of simulated SI data. We did this with the explicit goal in mind that an unsupervised OoD approach would allow us to use it as a filtering stage in a robust ML pipeline filtering data points that could lead to spurious predictions on downstream tasks. Overall, we found a perceivable domain gap between our *in silico* data and *in vivo* data for both porcine and human spectra. These findings indicate that further improvements to our simulation framework are necessary to bridge the domain gap. At the moment, we should expect non-negligible performance drops when ML models that were trained on the *in silico* data are applied to *in vivo* data.

The dissemination of our contributions toward research question D.2 overlaps with the dissemination of our contributions toward research question T.3 and can be found in that paragraph.

D.3. Can physiological tissue changes be detected via an out-of-distribution approach?

We were the first to apply a personalized ML approach to the problem of monitoring physiological parameter changes using MSI data. To this end, we phrased the problem as an OoD detection task and applied our OoD detection methodology based on INN ensembles and WAIC. Furthermore, to our knowledge, we were the first to develop a laparoscopic MSI system that was successfully applied in human surgery. To achieve our goal, we overcame obstacles due to data sparsity and noise introduced due to the volatile setting of a surgical intervention. Overall, we found that our OoD approach is capable of detecting tissue parameter changes. Besides the good classification performance, we found that with suitable pre-training, the INN ensemble could be adapted to the current patient in less than 1 min, implying an adequate training time for in-surgery training. While the classification performance of our approach was on par with the results of a baseline method based on k-NN, the inference time of the proposed methodology was orders of magnitude faster than the baseline leading to video rate capabilities of our approach. Lastly, we found a clear classification performance drop when using (reconstructed) RGB data to monitor the perfusion state as compared to MSI data. Our proposed personalized approach to ML models introduces a new paradigm to medical AI, which, together with MSI, can open the door to new non-invasive functional imaging applications and increased patient benefit.

The dissemination of our contributions toward research question D.3 overlaps with the dissemination of our contributions toward research question T.3 and can be found in that paragraph.

Miscellaneous Contributions

My ML expertise, together with my core project contributions, led to a number of fruitful collaborations with me in a supporting role. In this paragraph, I wanted to mention some of the highlights. I supported Thuy Nuong Tran and Amine Yamlahi in their work on the Computer Vision in Endoscopy (EndoCV) 2022 challenge [Tra+22; Yam+22]. Our efforts culminated in first place in the polyp detection challenge and third place in the polyp segmentation challenge [AG22]. Furthermore, I participated in the Hackathon on Energy Efficient AI (AI-HERO) 2021[2] organized by the Helmholtz Association, where my team won second place on the medical classification task [Deb+21].

The second half of my Ph. D. studies was under the influence of the coronavirus disease 2019 (COVID-19) pandemic. At the height of the first wave, my supervisor Lena Maier-Hein initiated a collaboration with the department of tropical medicine and infectious diseases at Heidelberg university hospital, which is led by Claudia Denkinger. Other members of our department and I supported the department of tropical medicine in analyzing COVID-19 symptom screening data and developing methods for optimized testing strategies. Our contribution to this work led to a joint manuscript which is currently under review at eClinicalMedicine (part of The Lancet publishing group). The corresponding study protocol can be found under [Dec+21].

9.2. Outlook

This thesis has presented a framework for uncertainty handling in medical imaging validated with the example of biophotonic imaging based on INNs. This work lays the foundation for exciting new opportunities with regard to robust medical applications based on ML. Our contribution to analyzing the well-posedness of inverse problems can directly be applied for camera[3] selection. Furthermore, with little adaptation, we can use it to analyze recording poses or geometries and their influence on the inverse processes. Our work on OoD detection can be directly incorporated as quality control in simulation frameworks. Additionally, it can be used as a task-agnostic filtering step for downstream ML methods to

[2] Postponed to early 2022.

[3] Or more generally, recording device

increase their prediction robustness. Lastly, our approach to personalized ML models is not restricted to tissue parameter monitoring in settings with high inter-patient variability. On the contrary, it can be applied to general (binary) classification tasks, where variability due to confounders covers the class signal. This variability might be due to e. g. a change in the recording device or in the recording environment (like different hospitals). Next to these immediate applications, our framework has the potential for a wider impact which we will describe below.

Detecting ambiguities is a valuable first step, but optimally, we do not stop there. Instead, it would be desirable for the medical device to suggest ways of resolving the ambiguity. For example, using a handheld PAI device, the model could propose new recording poses that could reduce the uncertainty of some tissue parameter estimation. Such a device is only thinkable with a robust method for ambiguity detection, as the proposed framework promises. Nevertheless, additional work on representing the device's environment in the ML model and advanced data fusion strategies are necessary to achieve this goal.

Our OoD detection approach could become a stepping stone for more robust ML methods in medicine. Next to its video rate capabilities, it is completely task-agnostic. Hence, it can be used as a filter in front of arbitrary downstream ML models. This approach can potentially prevent spurious predictions on OoD data, leading to more reliable results, which is a central requirement for medical applications. In addition, our OoD detection methodology could become an enabling factor in continual learning. For example, if the domain shifts over time, maybe due to the aging of the recording device, we might need to adapt models to this new reality. OoD detectors could trigger such retraining of the models once a sufficient domain shift is detected. Bringing such an approach to the clinic would require further standardization and possibly a common platform for ML models so that the different components, like the OoD detector and the downstream model, can communicate using well-defined interfaces.

The proposed personalized approach to physiological tissue parameter monitoring introduces a completely new paradigm to ML applications in medicine. While we are still at the early stages of the methodology, the first results are very promising. The biggest open challenge concerns the hardening of the OoD detection to confounders, like changes in pose or lighting. This will most likely require suitable (learned) representations of the surgical scene that are confounder invariant. However, once this obstacle is overcome, the medical applications are nigh countless. Whenever there is a stable status quo at the beginning of an intervention and if deviations from this status quo might have medical implications, our personalized approach has the potential to be adapted to the task. The most obvious example, and the example examined in this thesis, is perfusion monitoring, but there are other examples that fit this general setting, like monitoring the concentration of an administered agent or monitoring a patient for the onset of sepsis.

Next to the thesis's medical potential, the proposed validation framework for posteriors could initiate a community-wide discussion that might lead to standardized benchmarks for data-driven examination of (medical) inverse problems. Such an exchange could be very valuable for developing robust ML methods in medicine.

Overall, our results regarding uncertainty quantification are a first step toward the spectacular opportunities for medical ML waiting on the horizon.

With this, we have reached the end of my thesis, and there is only one thing left to say:

So long, and thanks for all the fish.[4]

[4] from Adam Douglas's "The Hitchhiker's Guide to the Galaxy"

Bibliography

[Adl+19a] Tim J Adler, Lynton Ardizzone, Leonardo Ayala, Janek Gröhl, Anant Vemuri, Sebastian J Wirkert, Beat P Müller-Stich, Carsten Rother, Ullrich Köthe, and Lena Maier-Hein. "Uncertainty handling in intra-operative multispectral imaging with invertible neural networks". In: *International Conference on Medical Imaging with Deep Learning–Extended Abstract Track.* 2019 (cit. on pp. 120, 153).

[Adl+19b] Tim J Adler, Lynton Ardizzone, Anant Vemuri, Leonardo Ayala, Janek Gröhl, Thomas Kirchner, Sebastian Wirkert, Jakob Kruse, Carsten Rother, Ullrich Köthe, and Lena Maier-Hein. "Uncertainty-aware performance assessment of optical imaging modalities with invertible neural networks". In: *International journal of computer assisted radiology and surgery* 14.6 (2019), pp. 997–1007 (cit. on pp. 100, 102–104, 107, 152).

[Adl+19c] Tim J Adler, Leonardo Ayala, Lynton Ardizzone, Hannes G Kenngott, Anant Vemuri, Beat P Müller-Stich, Carsten Rother, Ullrich Köthe, and Lena Maier-Hein. "Out of distribution detection for intra-operative functional imaging". In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures.* Springer, 2019, pp. 75–82 (cit. on pp. 120, 123, 124, 153).

[AG22] Sharib Ali and Noha Ghatwary, eds. *Proceedings of the 4th International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2022)* (Kolkata, India, Mar. 28, 2022). CEUR Workshop Proceedings. Aachen, 2022. URL: `http://ceur-ws.org/Vol-3148` (cit. on p. 155).

[Ard+18a] Lynton Ardizzone, Till Bungert, Felix Draxler, Ullrich Köthe, Jakob Kruse, Robert Schmier, and Peter Sorrenson. *Framework for Easily Invertible Architectures (FrEIA).* 2018. URL: `https://github.com/VLL-HD/FrEIA` (cit. on pp. 101, 109, 121, 131).

[Ard+18b] Lynton Ardizzone, Jakob Kruse, Carsten Rother, and Ullrich Köthe. "Analyzing Inverse Problems with Invertible Neural Networks". In: *International Conference on Learning Representations.* 2018 (cit. on pp. 5, 46, 50, 58, 61, 142).

[Ard+19] Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. "Guided image generation with conditional invertible neural networks". In: *arXiv preprint arXiv:1907.02392* (2019) (cit. on pp. 5, 51, 52, 61, 62).

[Ard+20] Lynton Ardizzone, Radek Mackowiak, Carsten Rother, and Ullrich Köthe. "Training normalizing flows with the information bottleneck for competitive generative classification". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 7828–7840 (cit. on pp. 51, 140).

[Arr99] Simon R Arridge. "Optical tomography in medical imaging". In: *Inverse problems* 15.2 (1999), R41 (cit. on p. 12).

[Aya+19] Leonardo A Ayala, Sebastian J Wirkert, Janek Gröhl, Mildred A Herrera, Adrian Hernandez-Aguilera, Anant Vemuri, Edgar Santos, and Lena Maier-Hein. "Live monitoring of haemodynamic changes with multispectral image analysis". In: *OR 2.0 context-aware operating theaters and machine learning in clinical Neuroimaging*. Springer, 2019, pp. 38–46 (cit. on pp. 5, 148).

[Aya+20] Leonardo Ayala, Silvia Seidlitz, Anant Vemuri, Sebastian J Wirkert, Thomas Kirchner, Tim J Adler, Christina Engels, Dogu Teber, and Lena Maier-Hein. "Light source calibration for multispectral imaging in surgery". In: *International Journal of Computer Assisted Radiology and Surgery* 15.7 (2020), pp. 1117–1125 (cit. on p. 120).

[Aya+22] Leonardo Ayala, Tim J Adler, Silvia Seidlitz, Sebastian Wirkert, Christina Engels, Alexander Seitel, Jan Sellner, Alexey Aksenov, Matthias Bodenbach, Pia Bader, Sebastian Baron, Anant Vemuri, Manuel Wiesenfarth, Nicholas Schreck, Diana Mindroc, Minu Tizabi, Sebastian Pirmann, Brittaney Everitt, Annette Kopp-Schneider, Dogu Teber, and Lena Maier-Hein. "Spectral imaging enables contrast agent-free real-time ischemia monitoring in laparoscopic surgery". In: *medRxiv* (2022) (cit. on pp. 29, 97, 102, 126–129, 131, 132, 153).

[Bak+14] Wesley B Baker, Ashwin B Parthasarathy, David R Busch, Rickson C Mesquita, Joel H Greenberg, and AG Yodh. "Modified Beer-Lambert law for blood flow". In: *Biomedical optics express* 5.11 (2014), pp. 4053–4075 (cit. on p. 27).

[Bay63] Thomas Bayes. "LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S". In: *Philosophical transactions of the Royal Society of London* 53 (1763), pp. 370–418 (cit. on pp. 23, 44).

[BB81] Arthur W Burks and Alice R Burks. "First general-purpose electronic computer". In: *IEEE Annals of the History of Computing* 3.04 (1981), pp. 310–389 (cit. on p. 32).

[BBK19] Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. "The need for uncertainty quantification in machine-assisted medical decision making". In: *Nature Machine Intelligence* 1.1 (2019), pp. 20–23 (cit. on pp. 57, 60).

[Bee52] August Beer. "Bestimmung der absorption des rothen lichts in farbigen flussigkeiten". In: *Ann. Physik* 162 (1852), pp. 78–88 (cit. on p. 27).

[Bel14] Stefan Bellini. *Fingerpulseoxymeter.* 2014. URL: `https://de.wikipedia.org/wiki/Pulsoxymetrie#/media/Datei:Pulox_Pulse_Oximeter.JPG` (visited on 09/29/2022) (cit. on p. 6).

[BF06] David R Bickel and Rudolf Frühwirth. "On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications". In: *Computational Statistics & Data Analysis* 50.12 (2006), pp. 3500–3530 (cit. on pp. 72, 111).

[Bie+21] Sebastian Bieringer, Anja Butter, Theo Heimel, Stefan Höche, Ullrich Köthe, Tilman Plehn, and Stefan T Radev. "Measuring QCD splittings with invertible networks". In: *SciPost Physics* 10.6 (2021), p. 126 (cit. on p. 61).

[Big+13] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. "Evasion attacks against machine learning at test time". In: *Joint European conference on machine learning and knowledge discovery in databases.* Springer. 2013, pp. 387–402 (cit. on pp. 7, 11).

[BN06] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning.* Vol. 4. 4. Springer, 2006 (cit. on p. 40).

[Bor+13] Michael S Borofsky, Inderbir S Gill, Ashok K Hemal, Tracy P Marien, Isuru Jayaratna, Louis S Krane, and Michael D Stifelman. "Near-infrared fluorescence imaging to facilitate super-selective arterial clamping during zero-ischaemia robotic partial nephrectomy". In: *BJU international* 111.4 (2013), pp. 604–610 (cit. on p. 126).

[Bou29] Pierre Bouguer. *Essai d'optique, sur la gradation de la lumiere.* Claude Jombert, 1729 (cit. on p. 27).

[Bro+17] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. "Adversarial patch". In: *arXiv preprint arXiv:1712.09665* (2017) (cit. on p. 7).

[Bro+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901 (cit. on pp. 3, 34, 42).

[Bro12] L Egbertus J Brouwer. "Beweis der Invarianz des n-dimensionalen Gebiets." In: *Mathematische Annalen* 71 (1912), pp. 305–313. URL: `http://eudml.org/doc/158534` (cit. on p. 21).

[Bry61] Arthur E Bryson. “A gradient method for optimizing multi-stage allocation processes”. In: *Proc. Harvard Univ. Symposium on digital computers and their applications.* Vol. 72. 1961, p. 22 (cit. on p. 41).

[But+21] Anja Butter, Theo Heimel, Sander Hummerich, Tobias Krebs, Tilman Plehn, Armand Rousselot, and Sophia Vent. “Generative networks for precision enthusiasts”. In: *arXiv preprint arXiv:2110.13632* (2021) (cit. on p. 61).

[Chá22] José A Chávez. “Generative Flows as a General Purpose Solution for Inverse Problems”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2022, pp. 1490–1498 (cit. on p. 62).

[Che+20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. “A simple framework for contrastive learning of visual representations”. In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607 (cit. on p. 62).

[Cho+14] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. “On the properties of neural machine translation: Encoder–decoder approaches”. In: *8th Workshop on Syntax, Semantics and Structure in Statistical Translation, SSST 2014.* Association for Computational Linguistics (ACL). 2014, pp. 103–111 (cit. on p. 42).

[Chu+17] William Chu, Avinash Chennamsetty, Robert Toroussian, and Clayton Lau. “Anaphylactic shock after intravenous administration of indocyanine green during robotic partial nephrectomy”. In: *Urology case reports* 12 (2017), pp. 37–38 (cit. on p. 127).

[CJA18] Hyunsun Choi, Eric Jang, and Alexander A Alemi. “Waic, but why? generative ensembles for robust anomaly detection”. In: *arXiv preprint arXiv:1810.01392* (2018) (cit. on pp. 39, 94–96, 144).

[Cla+20] Neil T Clancy, Geoffrey Jones, Lena Maier-Hein, Daniel S Elson, and Danail Stoyanov. “Surgical spectral imaging”. In: *Medical image analysis* 63 (2020), p. 101699 (cit. on p. 28).

[Cox+04] Benjamin T Cox, Jan G Laufer, Kornel P Kostli, and Paul C Beard. “Experimental validation of photoacoustic k-space propagation models”. In: *Photons Plus Ultrasound: Imaging and Sensing*. Vol. 5320. SPIE. 2004, pp. 238–248 (cit. on p. 31).

[Cur44] Haskell B Curry. “The method of steepest descent for non-linear minimization problems”. In: *Quarterly of Applied Mathematics* 2.3 (1944), pp. 258–261 (cit. on p. 36).

[CV95] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20.3 (1995), pp. 273–297 (cit. on p. 60).

[CW17] Nicholas Carlini and David Wagner. "Towards evaluating the robustness of neural networks". In: *2017 ieee symposium on security and privacy (sp)*. Ieee. 2017, pp. 39–57 (cit. on pp. 7, 11).

[Cyb89] George Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314 (cit. on p. 41).

[Deb+21] Charlotte Debus, Darya Trofimova, Fabian Isensee, Ines Reinartz, Markus Götz, and Wolfgang Suess, eds. (Feb. 1, 2022). Helmholtz Association of German Research Centres. 2021. URL: `https://events.hifis.net/event/109/` (cit. on p. 155).

[Dec+21] Andreas Deckert, Simon Anders, Manuela De Allegri, Hoa Thi Nguyen, Aurélia Souares, Shannon McMahon, Matthias Meurer, Robin Burk, Matthias Sand, Lisa Koeppel, Lena Maier-Hein, Tobias Roß, Tim J Adler, Tobias Siems, Lucia Brugnara, Stephan Brenner, Kondrad Herbst, Daniel Kirrmaier, Yuanqiang Duan, Svetlana Ovchinnikova, Kathleen Boerner, Michael Marx, Hans-Georg Kräusslich, Michael Knop, Bärnighausen Till, and Claudia Denkinger. "Effectiveness and cost-effectiveness of four different strategies for SARS-CoV-2 surveillance in the general population (CoV-Surv Study): Study protocol for a two-factorial randomized controlled multi-arm trial with cluster sampling". In: *Trials* 22.1 (2021), pp. 1–9 (cit. on p. 155).

[Dev+18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018) (cit. on pp. 3, 34, 42).

[Die+21] Maximilian Dietrich, Silvia Seidlitz, Nicholas Schreck, Manuel Wiesenfarth, Patrick Godau, Minu Tizabi, Jan Sellner, Sebastian Marx, Samuel Knödler, Michael M Allers, Leonardo Ayala, Karsten Schmidt, Thorsten Brenner, Alexander Studier-Fischer, Felix Nickel, Beat P Müller-Stich, Annette Kopp-Schneider, Markus A Weigand, and Lena Maier-Hein. "Machine learning-based analysis of hyperspectral images for automated sepsis diagnosis". In: *arXiv preprint arXiv:2106.08445* (2021) (cit. on pp. 5, 147).

[DKB14] Laurent Dinh, David Krueger, and Yoshua Bengio. "Nice: Non-linear independent components estimation". In: *arXiv preprint arXiv:1410.8516* (2014) (cit. on p. 46).

[Dos+21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event,* *Austria, May 3-7, 2021.* OpenReview.net, 2021. URL: `https://openreview.net/forum?id=YicbFdNTTy` (cit. on p. 42).

[Dre20] Kris K Dreher. *Efficient Simulation of Realistic Photoacoustic Images with Unsupervised Domain Adaptation.* Master thesis. 2020 (cit. on pp. 108, 111, 142).

[DSB16] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using real nvp". In: *arXiv preprint arXiv:1605.08803* (2016) (cit. on p. 46).

[FB09] Qianqian Fang and David A Boas. "Monte Carlo simulation of photon migration in 3D turbid media accelerated by graphics processing units". In: *Optics express* 17.22 (2009), pp. 20178–20190 (cit. on pp. 29, 31).

[Fis14] Gerd Fischer. *Lineare Algebra.* Springer Spektrum Wiesbaden, 2014 (cit. on p. 20).

[FK18] Christopher Fadden and Sri-Rajasekhar Kothapalli. "A single simulation platform for hybrid photoacoustic and RF-acoustic computed tomography". In: *Applied sciences (Basel, Switzerland)* 8.9 (2018), p. 1568 (cit. on p. 108).

[FM82] Kunihiko Fukushima and Sei Miyake. "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition". In: *Competition and cooperation in neural nets.* Springer, 1982, pp. 267–285 (cit. on p. 42).

[GAR19] Xiaoyi Gu, Leman Akoglu, and Alessandro Rinaldo. "Statistical analysis of nearest neighbor methods for anomaly detection". In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 62).

[Gel22] Andrew Gelman. *Epistemic and aleatoric uncertainty: The role of context.* 2022. URL: `https://statmodeling.stat.columbia.edu/2022/02/03/epistemic-and-aleatoric-uncertainty/` (visited on 06/23/2022) (cit. on p. 43).

[GG16] Yarin Gal and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning.* PMLR. 2016, pp. 1050–1059 (cit. on p. 58).

[Gin36] Corrado Gini. "On the measure of concentration with special reference to income and statistics". In: *Colorado College Publication, General Series* 208.1 (1936), pp. 73–79 (cit. on p. 60).

[GJ82] Lloyd Griffiths and C W Jim. "An alternative approach to linearly constrained adaptive beamforming". In: *IEEE Transactions on antennas and propagation* 30.1 (1982), pp. 27–34 (cit. on pp. 31, 112).

[Goo+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014) (cit. on pp. 3, 34, 42).

[Gre+12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. “A Kernel Two-Sample Test”. In: *Journal of Machine Learning Research* 13.25 (2012), pp. 723–773. URL: `http://jmlr.org/papers/v13/gretton12a.html` (cit. on pp. 50, 86).

[Grö+18] Janek Gröhl, Thomas Kirchner, Tim J Adler, and Lena Maier-Hein. “Confidence estimation for machine learning-based quantitative photoacoustics”. In: *Journal of Imaging* 4.12 (2018), p. 147 (cit. on pp. 12, 60, 142).

[Grö+21] Janek Gröhl, Melanie Schellenberg, Kris Dreher, and Lena Maier-Hein. “Deep learning for biomedical photoacoustic imaging: A review”. In: *Photoacoustics* 22 (2021), p. 100241 (cit. on p. 60).

[Grö+22] Janek Gröhl, Kris K Dreher, Melanie Schellenberg, Tom Rix, Niklas Holzwarth, Patricia Vieten, Leonardo Ayala, Sarah E Bohndiek, Alexander Seitel, and Lena Maier-Hein. “SIMPA: an open-source toolkit for simulation and image processing for photonics and acoustics”. In: *Journal of biomedical optics* 27.8 (2022), p. 083010 (cit. on pp. 31, 142).

[Gru22] Sebastian Gruber. *Ambiguous Image Registration using Conditional Invertible Neural Networks*. Master thesis. 2022 (cit. on p. 152).

[GSS14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *arXiv preprint arXiv:1412.6572* (2014) (cit. on pp. 7, 11).

[Haa09] Alfred Haar. *Zur theorie der orthogonalen funktionensysteme*. Georg-August-Universitat, Gottingen., 1909 (cit. on p. 52).

[Had02] Jacques Hadamard. “Sur les problèmes aux dérivées partielles et leur signification physique”. In: *Princeton university bulletin* (1902), pp. 49–52 (cit. on p. 20).

[Had23] Jacques Hadamard. *Lectures on Cauchy’s Problem in Linear Partial Differential Equations*. Yale University Press, 1923 (cit. on p. 20).

[Hal+22] Jonas Haldemann, Victor Ksoll, Daniel Walter, Yann Alibert, Ralf S Klessen, Willy Benz, Ullrich Koethe, Lynton Ardizzone, and Carsten Rother. “Exoplanet Characterization using Conditional Invertible Neural Networks”. In: *arXiv preprint arXiv:2202.00027* (2022) (cit. on p. 61).

[Has+09] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009 (cit. on p. 33).

[Hat04] Allen Hatcher. *Algebraic Topology*. 2004 (cit. on p. 140).

[He+15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision.* 2015, pp. 1026–1034 (cit. on pp. 101, 121, 131).

[He+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016, pp. 770–778 (cit. on p. 41).

[Hel09] Ernst Hellinger. "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen." In: *Journal für die reine und angewandte Mathematik* 1909.136 (1909), pp. 210–271 (cit. on p. 61).

[Her81] Alexander Hertle. "On the problem of well-posedness for the Radon transform". In: *Mathematical Aspects of Computerized Tomography.* Springer, 1981, pp. 36–44 (cit. on p. 21).

[Heu+17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 62).

[HH85] John A Hartigan and Pamela M Hartigan. "The dip test of unimodality". In: *The annals of Statistics* (1985), pp. 70–84 (cit. on p. 72).

[HHS21] Paul Hagemann, Johannes Hertrich, and Gabriele Steidl. "Stochastic normalizing flows for inverse problems: a Markov Chains viewpoint". In: *arXiv preprint arXiv:2109.11375* (2021) (cit. on p. 61).

[Hoc+19] Roman Hochuli, Lu An, Paul C Beard, and Benjamin T Cox. "Estimating blood oxygenation from photoacoustic images: can a simple linear spectroscopic inversion ever work?" In: *Journal of Biomedical Optics* 24.12 (2019), p. 121914 (cit. on p. 31).

[Hop82] John J Hopfield. "Neural networks and physical systems with emergent collective computational abilities." In: *Proceedings of the national academy of sciences* 79.8 (1982), pp. 2554–2558 (cit. on p. 42).

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780 (cit. on p. 42).

[HSW89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5 (1989), pp. 359–366 (cit. on p. 41).

[HW21] Eyke Hüllermeier and Willem Waegeman. "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods". In: *Machine Learning* 110.3 (2021), pp. 457–506 (cit. on p. 43).

[IM17] Abhaya Indrayan and Rajeev Kumar Malhotra. *Medical biostatistics.* Chapman and Hall/CRC, 2017 (cit. on p. 43).

[Ish+22] Isao Ishikawa, Takeshi Teshima, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. “Universal approximation property of invertible neural networks”. In: *arXiv preprint arXiv:2204.07415* (2022) (cit. on pp. 49, 140).

[Jam+13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning.* Vol. 112. Springer, 2013 (cit. on p. 34).

[JPZ97] Ana Justel, Daniel Peña, and Rubén Zamar. “A multivariate Kolmogorov-Smirnov test of goodness of fit”. In: *Statistics & probability letters* 35.3 (1997), pp. 251–259 (cit. on p. 85).

[Jum+21] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Celemns Meyer, Simon A A Kohl, Andrwe J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (2021), pp. 583–589 (cit. on pp. 3, 34, 42).

[Kan+14] Kazuhiro Kaneko, Hiroshi Yamaguchi, Takaaki Saito, Tomonori Yano, Yasuhiro Oono, Hiroaki Ikematsu, Shogo Nomura, Akihiro Sato, Motohiro Kojima, Hiroyasu Esumi, and Atsushi Ochiai. “Hypoxia imaging endoscopy equipped with laser light source from preclinical live animal study to first-in-human subject research”. In: *PLoS One* 9.6 (2014), e99055 (cit. on p. 101).

[Kan+22] Da Eun Kang, Eric W Pellegrini, Lynton Ardizzone, Ralf S Klessen, Ullrich Koethe, Simon CO Glover, and Victor F Ksoll. “Emission-line diagnostics of H ii regions using conditional invertible neural networks”. In: *Monthly Notices of the Royal Astronomical Society* 512.1 (2022), pp. 617–647 (cit. on p. 61).

[Kan60] Leonid V Kantorovich. “Mathematical methods of organizing and planning production”. In: *Management science* 6.4 (1960), pp. 366–422 (cit. on p. 85).

[Kar+15] Subhajit Karmakar, Eno Hysi, Michael C Kolios, and Ratan K Saha. “Realistic photoacoustic image simulations of collections of solid spheres using linear array transducer”. In: *Photons Plus Ultrasound: Imaging and Sensing 2015.* Vol. 9323. SPIE. 2015, pp. 493–500 (cit. on p. 108).

[KD18] Durk P Kingma and Prafulla Dhariwal. “Glow: Generative flow with invertible 1x1 convolutions”. In: *Advances in neural information processing systems* 31 (2018) (cit. on pp. 46, 49).

[Kel60] Henry J Kelley. "Gradient theory of optimal flight paths". In: *Ars Journal* 30.10 (1960), pp. 947–954 (cit. on p. 41).

[KG17] Alex Kendall and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?" In: *Advances in neural information processing systems* 30 (2017) (cit. on pp. 5, 58, 142).

[Kim+16] Jeesu Kim, Sara Park, Yuhan Jung, Sunyeob Chang, Jinyong Park, Yumiao Zhang, Jonathan F Lovell, and Chulhong Kim. "Programmable real-time clinical photoacoustic and ultrasound imaging system". In: *Scientific reports* 6.1 (2016), pp. 1–11 (cit. on pp. 31, 112).

[Kir+18] Thomas Kirchner, Franz Sattler, Janek Gröhl, and Lena Maier-Hein. "Signed real-time delay multiply and sum beamforming for multispectral photoacoustic imaging". In: *Journal of Imaging* 4.10 (2018), p. 121 (cit. on p. 31).

[Kir+19] Thomas Kirchner, Janek Gröhl, Mildred A Herrera, Tim J Adler, Adrián Hernández-Aguilera, Edgar Santos, and Lena Maier-Hein. "Photoacoustics can image spreading depolarization deep in gyrencephalic brain". In: *Scientific reports* 9.1 (2019), pp. 1–9 (cit. on pp. 5, 6).

[KIW20] Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. "Why normalizing flows fail to detect out-of-distribution data". In: *Advances in neural information processing systems* 33 (2020), pp. 20578–20589 (cit. on p. 39).

[KL51] Solomon Kullback and Richard A Leibler. "On information and sufficiency". In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86 (cit. on p. 61).

[Kle13] Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013 (cit. on pp. 23, 44).

[KM22] Johnson Kuan and Jonas Mueller. "Back to the Basics: Revisiting Out-of-Distribution Detection Baselines". In: *arXiv preprint arXiv:2207.03061* (2022) (cit. on p. 62).

[Koh+18] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. "A probabilistic u-net for segmentation of ambiguous images". In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 60).

[Kol33] Andrey Kolmogorov. "Sulla determinazione empirica di una legge di distribuzione". In: *Inst. Ital. Attuari, Giorn.* 4 (1933), pp. 83–91 (cit. on p. 85).

[Kru+21] Jakob Kruse, Lynton Ardizzone, Carsten Rother, and Ullrich Köthe. "Benchmarking invertible architectures on inverse problems". In: *arXiv preprint arXiv:2101.10763* (2021) (cit. on p. 49).

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012) (cit. on p. 42).

[Kso+20] Victor F Ksoll, Lynton Ardizzone, Ralf Klessen, Ullrich Koethe, Elena Sabbi, Massimo Robberto, Dimitrios Gouliermis, Carsten Rother, Peter Zeidler, and Mario Gennaro. “Stellar parameter determination from photometry using invertible neural networks”. In: *Monthly Notices of the Royal Astronomical Society* 499.4 (2020), pp. 5447–5485 (cit. on p. 61).

[Kuh55] Harold W Kuhn. “The Hungarian method for the assignment problem”. In: *Naval research logistics quarterly* 2.1-2 (1955), pp. 83–97 (cit. on p. 88).

[Lam92] Johann Heinrich Lambert. *Lamberts Photometrie:(Photometria, sive De mensura et gradibus luminis, colorum et umbrae)(1760).* 31-33. W Engelmann, 1892 (cit. on p. 27).

[LeC+89] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. “Backpropagation applied to handwritten zip code recognition”. In: *Neural computation* 1.4 (1989), pp. 541–551 (cit. on p. 42).

[LeC+98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324 (cit. on p. 42).

[LH17] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017) (cit. on pp. 101, 109, 131).

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and scalable predictive uncertainty estimation using deep ensembles”. In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 59).

[LT20] Mishan Listmann and R Shane Tubbs. “The Abdominal Aorta”. In: *Surgical Anatomy of the Lateral Transpsoas Approach to the Lumbar Spine.* Elsevier, 2020, pp. 185–188 (cit. on p. 127).

[LuN+18] A LuNežič, Tomáš Vojíř, L Čehovin Zajc, Jiří Matas, and Matej Kristan. “Discriminative correlation filter TracNer with channel and spatial reliability”. In: *International Journal of Computer Vision* 126.7 (2018), pp. 671–688 (cit. on p. 129).

[Mac+21] Radek Mackowiak, Lynton Ardizzone, Ullrich Köthe, and Carsten Rother. “Generative classifiers as a basis for trustworthy image classification”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2021, pp. 2971–2981 (cit. on p. 49).

[Mad+17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. “Towards deep learning models resistant to adversarial attacks”. In: *arXiv preprint arXiv:1706.06083* (2017) (cit. on pp. 7, 11).

[Mah36] Prasanta Chandra Mahalanobis. “On the generalised distance in statistics”. In: *Proceedings of the National Institute of Sciences of India.* Vol. 2. 1. 1936, pp. 49–55 (cit. on p. 87).

[Mai+18a] Lena Maier-Hein, Sebastian J Wirkert, Anant S Vemuri, Leonardo Antonio Ayala Menjivar, Silvia Seidlitz, Thomas Kirchner, and Tim J Adler. “Method and system for augmented imaging in open treatment using multispectral information”. patent (pending) WO2020025684A1. 2018 (cit. on p. 152).

[Mai+18b] Lena Maier-Hein, Sebastian J Wirkert, Anant S Vemuri, Leonardo Antonio Ayala Menjivar, Silvia Seidlitz, Thomas Kirchner, and Tim J Adler. “Method and system for augmented imaging using multispectral information”. patent (pending) EP3830790A1. 2018 (cit. on p. 153).

[Mai+22a] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, Hirenkumar Nakawala, Adrian Park, Carla M Pugh, Danail Stoyanov, S Swaroop Vedula, Beat P Müller-Stich, Kevin Cleary, Gabor Fichtinger, Germain Forestier, Bernard Gibaud, Teodor P Grantcharov, Makoto Hashizume, Hannes Kenngott, Ron Kikinis, Lars Mündermann, Nassir Navab, Sinan Onogur, Raphael Sznitman, Russell H Taylor, Minu D Tizabi, Martin Wagner, Gregory D Hager, Thomas Neumuth, Nicolas Padoy, Pierre Jannin, and Stefanie Speidel. “Surgical data science–from concepts toward clinical translation”. In: *Medical image analysis* 76 (2022), p. 102306 (cit. on p. 43).

[Mai+22b] Lena Maier-Hein, Annika Reinke, Evangelia Christodoulou, Ben Glocker, Patrick Godau, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A Riegler, Manuel Wiesenfarth, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, AEmre Kavur, Tim Rädsch, Minu D Tizabi, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Peter Bankhead, Arriel Benis, M Jorge Cardoso, Veronika Cheplygina, Beth Cimini, Gary S Collins, Keyvan Farahani, Bram van Ginneken, Daniel A Hashimoto, Michael M Hoffman, Merel Huisman, Pierre Jannin, Charles E Kahn, Alexandros Karargyris, Alan Karthikesalingam, Hannes Kenngott, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L Martel, Peter Mattson, Erik Meijering, Bjoern Menze, David Moher, Karel G M Moons, et al. “Metrics reloaded: Pitfalls and recommendations for image analysis validation”. In: *arXiv preprint arXiv:2206.01653* (2022) (cit. on pp. 36, 62, 73, 78, 93).

[Mar+22] Laura Martínez-Ferrer, Álvaro Moreno-Martínez, Manuel Campos-Taberner, Francisco Javier García-Haro, Jordi Muñoz-Marí, Steven W Running, John Kimball, Nicholas Clinton, and Gustau Camps-Valls. “Quantifying uncertainty in high resolution biophysical variable retrieval with machine learning”. In: *Remote Sensing of Environment* 280 (2022), p. 113199 (cit. on p. 61).

[Mat+14] Giulia Matrone, Alessandro Stuart Savoia, Giosuè Caliano, and Giovanni Magenes. “The delay multiply and sum beamforming algorithm in ultrasound B-mode medical imaging”. In: *IEEE transactions on medical imaging* 34.4 (2014), pp. 940–949 (cit. on p. 31).

[McC+14] Tyler R McClintock, Marc A Bjurlin, James S Wysock, Michael S Borofsky, Tracy P Marien, Chinonyerem Okoro, and Michael D Stifelman. “Can selective arterial clamping with fluorescence imaging preserve kidney function during robotic partial nephrectomy?” In: *Urology* 84.2 (2014), pp. 327–334 (cit. on p. 126).

[MF17] Justin Matejka and George Fitzmaurice. “Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing”. In: *Proceedings of the 2017 CHI conference on human factors in computing systems*. 2017, pp. 1290–1294 (cit. on pp. 3, 5).

[Moc+18] Sara Moccia, Sebastian J Wirkert, Hannes Kenngott, Anant S Vemuri, Martin Apitz, Benjamin Mayer, Elena De Momi, Leonardo S Mattos, and Lena Maier-Hein. “Uncertainty-aware organ classification for surgical data science applications in laparoscopy”. In: *IEEE Transactions on Biomedical Engineering* 65.11 (2018), pp. 2649–2659 (cit. on p. 60).

[Mok+21] Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, and Evgeny Burnaev. “Large-scale wasserstein gradient flows”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 15243–15256 (cit. on p. 61).

[Mon+20] Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. “Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 12756–12767 (cit. on p. 60).

[NAB22] Pablo Noever-Castelos, Lynton Ardizzone, and Claudio Balzani. “Model updating of wind turbine blade cross sections with invertible neural networks”. In: *Wind Energy* 25.3 (2022), pp. 573–599 (cit. on p. 61).

[Nal+19a] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. *Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality.* 2019. DOI: `10.48550/ARXIV.1906.02994`. URL: `https://arxiv.org/abs/1906.02994` (cit. on pp. 39, 63, 94).

[Nal+19b] Eric T Nalisnick, Akihiro Matsukawa, Yee W Teh, Dilan Görür, and Balaji Lakshminarayanan. "Do Deep Generative Models Know What They Don't Know?" In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.* OpenReview.net, 2019. URL: `https://openreview.net/forum?id=H1xwNhCcYm` (cit. on p. 39).

[Nes03] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course.* Vol. 87. Springer Science & Business Media, 2003 (cit. on p. 41).

[Nöl+21] Jan-Hinrich Nölke, Tim J Adler, Janek Gröhl, Thomas Kirchner, Lynton Ardizzone, Carsten Rother, Ullrich Köthe, and Lena Maier-Hein. "Invertible neural networks for uncertainty quantification in photoacoustic imaging". In: *Bildverarbeitung für die Medizin 2021.* Springer, 2021, pp. 330–335 (cit. on pp. 31, 108, 119, 152).

[Nöl21a] Jan-Hinrich Nölke. Personal Communication. 2021 (cit. on p. 72).

[Nöl21b] Jan-Hinrich Nölke. *Uncertainty Quantification for Photoacoustic Imaging using Invertible Neural Networks.* Master thesis. 2021 (cit. on pp. 108–110, 112–118, 152).

[Oha04] Tony O'hara. *Dicing with the unknown.* 2004. URL: `http://www.stat.columbia.edu/~gelman/stuff_for_blog/ohagan.pdf` (cit. on p. 43).

[Ped+11] F Pedregosa, G Varoquaux, A Gramfort, V. Michel, B Thirion, O Grisel, M Blondel, P. Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 131).

[PGM17] Andrea Papadia, Maria Luisa Gasparri, and Michael Mueller. "Are allergic reactions to indocyanine green really that uncommon? A single institution experience". In: *Obstetrics and gynecology reports* 1.2 (2017), pp. 1–2 (cit. on p. 127).

[Pin99] Allan Pinkus. "Approximation theory of the MLP model in neural networks". In: *Acta numerica* 8 (1999), pp. 143–195 (cit. on p. 41).

[Pir20] Sebastian Pirmann. *Tracking of Regions of Interest in Multispectral Laparoscopy.* Master thesis. 2020 (cit. on p. 129).

[Pra98] Scott Prahl. *Optical Absorption of Hemoglobin.* 1998. URL: `https://omlc.org/spectra/hemoglobin/index.html` (cit. on p. 28).

[Rad+20] Stefan T Radev, Ulf K Mertens, Andreas Voss, Lynton Ardizzone, and Ullrich Köthe. "BayesFlow: Learning complex stochastic models with invertible neural networks". In: *IEEE transactions on neural networks and learning systems* (2020) (cit. on p. 61).

[Ram+22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* (2022) (cit. on pp. 3, 34, 42).

[Ren+19] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A DePristo, Joshua V Dillon, and Balaji Lakshminarayanan. *Likelihood Ratios for Out-of-Distribution Detection.* 2019. doi: 10.48550/ARXIV.1906.02845. url: https://arxiv.org/abs/1906.02845 (cit. on p. 39).

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention.* Springer. 2015, pp. 234–241 (cit. on pp. 3, 34, 42, 60, 110).

[RHW85] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. *Learning internal representations by error propagation.* Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science, 1985 (cit. on p. 41).

[RHW86] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536 (cit. on p. 41).

[RN20] Stuart J Russell and Peter Norvig. *Artificial Intelligence: a modern approach.* 4th ed. Pearson, 2020 (cit. on pp. 32, 34, 40).

[RNR13] Amir Rosenthal, Vasilis Ntziachristos, and Daniel Razansky. "Acoustic inversion in optoacoustic tomography: A review". In: *Current Medical Imaging* 9.4 (2013), pp. 318–336 (cit. on p. 12).

[Rob+21] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, Alessandro Ruggiero, Anna Korhonen, Emily Jefferson, Emmanuel Ako, Georg Langs, Ghassem Gozaliasl, Guang Yang, Helmut Prosch, Jacobus Preller, Jan Stanczuk, Jing Tang, Johannes Hofmanninger, Judith Babar, Lorena Escudero Sánchez, Muhunthan Thillai, Paula Martin Gonzalez, Philip Teare, Xiaoxiang Zhu, Mishal Patel, Conor Cafolla, Hojjat Azadbakht, Joseph Jacob, Josh Lowe, Kang Zhang, Kyle Bradley, Marcel Wassin, Markus Holzer, Kangyu Ji, Maria Delgado Ortet, Tao Ai, Nicholas Walton, Pietro Lio, Samuel Stranks, Tolou Shadbahr, Weizhe Lin, Yunfei Zha, Zhangming Niu, et al. "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans". In: *Nature Machine Intelligence 2021 3:3* 3 (3 Mar. 2021), pp. 199–217. ISSN: 2522-5839. DOI: `10.1038/s42256-021-00307-0`. URL: `https://www.nature.com/articles/s42256-021-00307-0` (cit. on p. 148).

[RT19] Kui Ren and Faouzi Triki. "A Global stability estimate for the photo-acoustic inverse problem in layered media". In: *European Journal of Applied Mathematics* 30.3 (2019), pp. 505–528 (cit. on p. 12).

[Sah+20] Teemu Sahlström, Aki Pulkkinen, Jenni Tick, Jarkko Leskinen, and Tanja Tarvainen. "Modeling of errors due to uncertainties in ultrasound sensor locations in photoacoustic tomography". In: *IEEE Transactions on Medical Imaging* 39.6 (2020), pp. 2140–2150 (cit. on p. 61).

[Sch19] Nicholas Schreck. "Empirical decomposition of the explained variation in the variance components form of the mixed model". In: *bioRxiv* (2019) (cit. on pp. 131, 132).

[Sel+20] Raghavendra Selvan, Frederik Faye, Jon Middleton, and Akshay Pai. "Uncertainty quantification in medical image segmentation with normalizing flows". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2020, pp. 80–90 (cit. on p. 60).

[Sen+20] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, Hugo Penedones, Stig Perersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T Jonse, David Silver, Koray Kavukcuoglu, and Demis Hassabis. "Improved protein structure prediction using potentials from deep learning". In: *Nature* 577.7792 (2020), pp. 706–710 (cit. on pp. 3, 34, 42).

[SG18] Lewis Smith and Yarin Gal. "Understanding measures of uncertainty for adversarial example detection". In: *arXiv preprint arXiv:1803.08533* (2018) (cit. on p. 59).

[Sha+21] Rohan Shad, John P Cunningham, Euan A Ashley, Curtis P Langlotz, and William Hiesinger. "Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging". In: *Nature Machine Intelligence 2021 3:11* 3 (11 Nov. 2021), pp. 929–935. ISSN: 2522-5839. DOI: `10.1038/s42256-021-00399-8`. URL: `https://www.nature.com/articles/s42256-021-00399-8` (cit. on p. 148).

[Sha48] Claude E Shannon. "A mathematical theory of communication". In: *The Bell system technical journal* 27.3 (1948), pp. 379–423 (cit. on p. 94).

[Sil+16] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587 (2016), pp. 484–489 (cit. on pp. 3, 42).

[Sil+17] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. "Mastering the game of go without human knowledge". In: *nature* 550.7676 (2017), pp. 354–359 (cit. on pp. 3, 42).

[Sil+18] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362.6419 (2018), pp. 1140–1144 (cit. on pp. 3, 42).

[Sil18] Bernard W Silverman. *Density estimation for statistics and data analysis.* Routledge, 2018 (cit. on p. 71).

[Smi48] Nickolay Smirnov. "Table for estimating the goodness of fit of empirical distributions". In: *The annals of mathematical statistics* 19.2 (1948), pp. 279–281 (cit. on p. 85).

[Soh+21] Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister. "Learning and Evaluating Representations for Deep One-Class Classification". In: *International Conference on Learning Representations.* 2021. URL: `https://openreview.net/forum?id=HCSgyPUfeDj` (cit. on p. 62).

[Son+21] Jingwei Song, Shaobo Xia, Jun Wang, Mitesh Patel, and Dong Chen. "Uncertainty quantification of hyperspectral image denoising frameworks based on sliding-window low-rank matrix approximation". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), pp. 1–12 (cit. on p. 60).

[Sri+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958 (cit. on p. 58).

[SSS19] Jacob J Senecal, John W Sheppard, and Joseph A Shaw. "Efficient convolutional neural networks for multi-spectral image classification". In: *2019 International Joint Conference on Neural Networks (IJCNN).* IEEE. 2019, pp. 1–8 (cit. on p. 60).

[Stu+21] Alexander Studier-Fischer, Silvia Seidlitz, Jan Sellner, Manuel Wiesenfarth, Leonardo Ayala, Berkin Özdemir, Jan Odenthal, Samuel Knödler, Karl-Friedrich Kowalewski, Caelan M Haney, Isabella Camplisson, Maximilian Dietrich, Karsten Schmidt, Gabriel A Salg, Hannes G Kenngott, Tim J Adler, Nicholas Schreck, Annette Kopp-Schneider, Klaus Maier-Hein, Lena Maier-Hein, Beat P Müller-Stich, and Felix Nickel. "Spectral organ fingerprints for intraoperative tissue classification with hyperspectral imaging". In: *bioRxiv* (2021) (cit. on p. 147).

[Stu10] Andrew M Stuart. "Inverse problems: a Bayesian perspective". In: *Acta numerica* 19 (2010), pp. 451–559 (cit. on p. 21).

[Suz17] Kenji Suzuki. "Overview of deep learning in medical imaging". In: *Radiological physics and technology* 10.3 (2017), pp. 257–273 (cit. on p. 42).

[Taj+21] Fahim Tajwar, Ananya Kumar, Sang Michael Xie, and Percy Liang. "No True State-of-the-Art? OOD Detection Methods are Inconsistent across Datasets". In: *arXiv preprint arXiv:2109.05554* (2021) (cit. on p. 64).

[Tar+13] Tanja Tarvainen, Aki Pulkkinen, Ben T Cox, Jari P Kaipio, and Simon R Arridge. "Bayesian image reconstruction in quantitative photoacoustic tomography". In: *IEEE transactions on medical imaging* 32.12 (2013), pp. 2287–2298 (cit. on p. 61).

[Tho+07] R Houston Thompson, Igor Frank, Christine M Lohse, Ismail R Saad, Amr Fergany, Horst Zincke, Bradley C Leibovich, Michael L Blute, and Andrew C Novick. "The impact of ischemia time during open nephron sparing surgery on solitary kidneys: a multi-institutional study". In: *The Journal of urology* 177.2 (2007), pp. 471–476 (cit. on p. 126).

[Tia+20] Yu Tian, Gabriel Maicas, Leonardo Zorron Cheng Tao Pu, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. "Few-shot anomaly detection for polyp frames from colonoscopy". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer. 2020, pp. 274–284 (cit. on p. 62).

[Tia+21] Yu Tian, Guansong Pang, Fengbei Liu, Yuanhong Chen, Seon Ho Shin, Johan W Verjans, Rajvinder Singh, and Gustavo Carneiro. "Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer. 2021, pp. 128–140 (cit. on p. 63).

[Tie21] Tamara Tiefenauer. *KSB Blog*. 2021. URL: `https://blog.ksb.ch/wissen/hautkrebs-erkennen/` (visited on 09/29/2022) (cit. on p. 6).

[Tol+17] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. “Wasserstein auto-encoders”. In: *arXiv preprint arXiv:1711.01558* (2017) (cit. on p. 86).

[TPT16] Jenni Tick, Aki Pulkkinen, and Tanja Tarvainen. “Image reconstruction with uncertainty quantification in photoacoustic tomography”. In: *The Journal of the Acoustical Society of America* 139.4 (2016), pp. 1951–1961 (cit. on p. 61).

[TPT19] Jenni Tick, Aki Pulkkinen, and Tanja Tarvainen. “Modelling of errors due to speed of sound variations in photoacoustic tomography using a Bayesian framework”. In: *Biomedical physics & engineering express* 6.1 (2019), p. 015003 (cit. on p. 61).

[Tra+22] Thuy Nuong Tran, Fabian Isensee, Lars Krämer, Amine Yamlahi, Tim J Adler, Patrick Godau, Minu Tizabi, and Lena Maier-Hein. “Heterogeneous model ensemble for automatic polyp segmentation in endoscopic video sequences”. In: *Proceedings of the 4th International Workshop and Challenge on Computer Vision in Endoscopy* (2022) (cit. on pp. 59, 155).

[Tro+20] Darya Trofimova, Tim J Adler, Lisa Kausch, Lynton Ardizzone, Klaus Maier-Hein, Ulrich Köthe, Carsten Rother, and Lena Maier-Hein. “Representing ambiguity in registration problems with conditional invertible neural networks”. In: *arXiv preprint arXiv:2012.08195* (2020) (cit. on pp. 147, 152).

[TW16] Jakub M Tomczak and Max Welling. “Improving variational auto-encoders using householder flow”. In: *arXiv preprint arXiv:1611.09630* (2016) (cit. on p. 48).

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 42).

[Vin+19] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P Agapiou, Max Jaderberg, Alexander S Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L Paine, Caglar Gulcehre, Wang Ziyu, Tobias Pfaff, Yuhuai Wu, roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. “Grandmaster level in StarCraft II using multi-agent reinforcement learning”. In: *Nature* 575.7782 (2019), pp. 350–354 (cit. on p. 42).

[Wai19] Waidmann8x57. *Gehirn eines Rehbocks ca. zwei tunden nach Erlegung*. 2019. URL: `https://de.wikipedia.org/wiki/Gehirn#/media/Datei:Gehirn_eines_Rehbocks_-_brain_of_a_roebuck.jpg` (visited on 09/29/2022) (cit. on p. 6).

[Wan+04] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612 (cit. on p. 62).

[Wat13] Sumio Watanabe. "WAIC and WBIC are information criteria for singular statistical model evaluation". In: *Proceedings of the Workshop on Information Theoretic Methods in Science and Engineering*. 2013, pp. 90–94 (cit. on pp. 39, 95, 96).

[Wer20] Wolfram Wermke. *Ultraschall mit Kontrastmittel erkennt Leberkrebs zuverlässig*. 2020. URL: `https://healthcare-in-europe.com/de/news/ultraschall-mit-kontrastmittel-erkennt-leberkrebs-zuverlaessig.html` (visited on 09/29/2022) (cit. on p. 6).

[Wil13] Richard David Wilkinson. "Approximate Bayesian computation (ABC) gives exact results under the assumption of model error". In: *Statistical applications in genetics and molecular biology* 12.2 (2013), pp. 129–141 (cit. on p. 59).

[Win+20] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, A Taylan Cemgil, S M Ali Eslami, and Olaf Ronneberger. "Contrastive training for improved out-of-distribution detection". In: *arXiv preprint arXiv:2007.05566* (2020) (cit. on p. 63).

[Wir+16] Sebastian J Wirkert, Hannes Kenngott, Benjamin Mayer, Patrick Mietkowski, Martin Wagner, Peter Sauer, Neil T Clancy, Daniel S Elson, and Lena Maier-Hein. "Robust near real-time estimation of physiological parameters from megapixel multispectral images with inverse Monte Carlo and random forest regression". In: *International journal of computer assisted radiology and surgery* 11.6 (2016), pp. 909–917 (cit. on p. 29).

[Wir+17] Sebastian J Wirkert, Anant S Vemuri, Hannes G Kenngott, Sara Moccia, Michael Götz, Benjamin FB Mayer, Klaus H Maier-Hein, Daniel S Elson, and Lena Maier-Hein. "Physiological parameter estimation from multispectral images unleashed". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 134–141 (cit. on pp. 6, 102, 148).

[WJ92] Lihong Wang and Steven L Jacques. "Monte Carlo modeling of light transport in multi-layered tissues in standard C". In: *The University of Texas, MD Anderson Cancer Center, Houston* 4.11 (1992) (cit. on pp. 29, 102).

[WW12] Lihong V Wang and Hsin-i Wu. *Biomedical optics: principles and imaging*. John Wiley & Sons, 2012 (cit. on p. 26).

[WWG06] Brady T West, Kathleen B Welch, and Andrzej T Galecki. *Linear mixed models: a practical guide using statistical software*. Chapman and Hall/CRC, 2006 (cit. on p. 93).

[XW05] Minghua Xu and Lihong V Wang. "Universal back-projection algorithm for photoacoustic computed tomography". In: *Phys. Rev. E* 71 (2005), p. 016706. DOI: `10.1103/PhysRevE.71.016706`. URL: `https://link.aps.org/doi/10.1103/PhysRevE.71.016706` (cit. on p. 31).

[XYA20] Zhisheng Xiao, Qing Yan, and Yali Amit. "Improving Sample Quality by Training and Sampling from Latent Energy". In: *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models* (2020) (cit. on p. 139).

[XYW14] Jun Xia, Junjie Yao, and Lihong V Wang. "Photoacoustic tomography: principles and advances". In: *Electromagnetic waves (Cambridge, Mass.)* 147 (2014), p. 1 (cit. on pp. 26, 30).

[Yam+22] Amine Yamlahi, Patrick Godau, Thuy Nuong Tran, Lucas-Raphael Müller, Tim J Adler, Minu Dietlinde Tizabi, Michael Baumgartner, Paul Jäger, and Lena Maier-Hein. "Heterogeneous model ensemble for polyp detection and tracking in colonoscopy". In: *Proceedings of the 4th International Workshop and Challenge on Computer Vision in Endoscopy* (2022) (cit. on pp. 59, 155).

[Yan+21] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. "Generalized out-of-distribution detection: A survey". In: *arXiv preprint arXiv:2110.11334* (2021) (cit. on p. 38).

[ZAP20] Qingyu Zhao, Ehsan Adeli, and Kilian M. Pohl. "Training confounder-free deep learning models for medical applications". In: *Nature Communications 2020 11:1* 11 (1 Nov. 2020), pp. 1–9. ISSN: 2041-1723. DOI: `10.1038/s41467-020-19784-9`. URL: `https://www.nature.com/articles/s41467-020-19784-9` (cit. on p. 147).

[Zho+21] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises". In: *Proceedings of the IEEE* 109.5 (2021), pp. 820–838 (cit. on pp. 42, 58).

[Zim+19] David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein. "Unsupervised anomaly localization using variational auto-encoders". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 289–297 (cit. on p. 62).

[Zim+22] David Zimmerer, Peter M Full, Fabian Isensee, Paul Jäger, Tim J Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Annika Reinke, Antanas Kascenas, Bjorn S Jensen, Alison Q O'Neil, Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, Bernhard Kainz, Nina Shvetsova, Irina Fedulova, Dmitry V Dylov, Baolun Yu, Jianyang Zhai, Jingtao Hu, Runxuan Si, Sihang Zhou, Siqi Wang, Xinyang Li, Xuerun Chen, Yang Zhao, Sergio N Marimont, Giacomo Tarroni, Victor Saase, Lena Maier-Hein, and Klaus Maier-Hein. "MOOD 2020: A public Benchmark for Out-of-Distribution Detection and Localization on medical Images". In: *IEEE Transactions on Medical Imaging* (2022) (cit. on p. 153).

[ZZ09] Ethan Zhang and Yi Zhang. "Average Precision". In: *Encyclopedia of Database Systems*. Ed. by LING LIU and M TAMER ÖZSU. Boston, MA: Springer US, 2009, pp. 192–193. ISBN: 978-0-387-39940-9. DOI: `10.1007/978-0-387-39940-9_482`. URL: `https://doi.org/10.1007/978-0-387-39940-9_482` (cit. on p. 91).

Uncertainty Quantification in Biophotonic Imaging using Invertible Neural Networks
Ph. D. Thesis

Supervised by Prof. Dr. Lena Maier-Hein

This work has been set using LaTeX and KOMA-Script.

Main Font:	Linux Libertine
Sans Font:	Linux Biolinum
Color Scheme:	`https://personal.sron.nl/~pault/`