Inaugural – Dissertation

zur

Erlangung der Doktorwürde

der

Gesamtfakultät für Mathematik, Ingenieur- und Naturwissenschaften

der

Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Patrick Godau, Ms. Sc.

aus Heidelberg

Tag der mündlichen Prüfung: _____

Lifelong Machine Learning for Biomedical Image Classification

Supervisor: Prof. Dr. Lena Maier-Hein

For Facts, not Fear.
For Clarity, not Claims.
For Discourse, not Division.
For Progress, not Propaganda.
For Evidence, not Echo Chambers.

For Science, not Silence.

Abstract

Despite rapid advances in the capabilities of Artificial Intelligence (AI), its application in healthcare faces unique challenges: stringent requirements to prove patient benefit, severe data scarcity, and shifting distributions across clinical environments. These barriers span the entire AI lifecycle, requiring solutions at each stage. To address this, we propose a holistic *Lifelong Learning* framework that systematically addresses these challenges through three independent metacognitive loops – continuous improvements of the learning process itself: one to align validation efforts with clinical needs during the *Design* phase, another for effective knowledge transfer between tasks in the *Develop* phase, and a third to adapt models to changing environments in the *Deploy* phase.

Adding these loops to a learning system helps to overcome the challenges outlined above. First, we develop a structured interview process that captures the *problem fingerprint* of biomedical applications and enables automatic determination of appropriate performance measures aligned with clinical objectives. Second, we establish a method for quantifying task similarity and facilitating cross-institutional knowledge transfer while preserving patient data privacy. Our proposed binned Kullback-Leibler Divergence (bKLD) measure underwent extensive evaluation across heterogeneous biomedical imaging tasks, setting new standards for task transferability estimation. Third, we comprehensively analyze prevalence shifts in deployment environments and propose a novel five-step workflow for model adaptation using only unlabeled samples from the deployment environment. Therein we quantify the present class prevalences and post-hoc re-calibrate a model, carefully considering the impact on decision rules and performance measures.

The experiments conducted demonstrate significant advancements in each area. Our recommendation process for aligning model performance metrics with actual clinical utility, reflects the consensus of an international consortium of 73 experts. Our knowledge transfer methodology allows the system to leverage experience from related tasks, exceeding previously proposed estimates of knowledge transferability in the most comprehensive benchmark we are aware of. Our prevalence shift compensation workflow prevents performance degradation across diverse biomedical imaging scenarios, enabling the system to automatically detect and adapt to changing environmental conditions without requiring new annotations.

This work represents the first comprehensive investigation of Lifelong Learning for biomedical image analysis, with tens of thousands of models trained and evaluated. By systematically leveraging metacognitive loops, we lay the groundwork for truly autonomous Lifelong Learning systems in healthcare that can continuously evolve in changing healthcare contexts.

Zusammenfassung

Trotz der rasanten Fortschritte im Bereich der Künstlichen Intelligenz (KI) steht diese bei der Übertragung in medizinische Anwendungen vor einzigartigen Herausforderungen: strenge Anforderungen an den Nachweis des Patientennutzens, unzureichende Datenverfügbarkeit und Differenzen in der statistischen Verteilung zwischen verschiedenen klinischen Umgebungen. Diese Hindernisse erstrecken sich über den gesamten KI-Lebenszyklus und erfordern daher auch Lösungen in jeder Phase. Um diesen Herausforderungen zu begegnen, schlagen wir ein ganzheitliches *Lifelong Learning* System vor, das durch drei unabhängige metakognitive Schleifen ergänzt wird, welche kontinuierlich den Lernprozess selbst verbessern: eine für die Abstimmung von Validierungsmaßen mit klinischen Bedürfnissen in der *Design*-Phase, eine für den effektiven Wissenstransfer zwischen Datensätzen in der *Develop*-Phase und eine dritte für die Anpassung von Modellen an veränderte Umgebungen in der *Deploy*-Phase.

Die Einbindung dieser Schleifen hilft, die oben genannten Herausforderungen zu bewältigen. Erstens entwerfen wir einen strukturierten Interviewprozess, der den *Problem-Fingerabdruck* biomedizinischer Anwendungen erfasst und die automatische Bestimmung geeigneter, auf klinische Ziele ausgerichteter Leistungsmetriken ermöglicht. Zweitens etablieren wir eine Methode zur Quantifizierung der Ähnlichkeit von Datensätzen und erleichtern den institutionenübergreifenden Wissenstransfer bei gleichzeitiger Wahrung des Datenschutzes für Patient*innen. Unser vorgeschlagenes bKLD-Maß wurde umfassend an heterogenen biomedizinischen Bilddatensätzen evaluiert und setzt neue Standards für die Abschätzung der Transferfähigkeit von Wissen. Drittens analysieren wir umfassend die Auswirkungen von Verschiebungen in der Zielklassenverteilung und schlagen einen neuen fünfstufigen Algorithmus zur Anpassung von Modellen vor, der auf nicht kategorisierten Bildern der Einsatzumgebung basiert. Dabei quantifizieren wir die neu entstandene Klassenverteilung und rekalibrieren ein Modell unter sorgfältiger Berücksichtigung der Auswirkungen auf Entscheidungsregeln und Leistungsmaße.

Die durchgeführten Experimente zeigen signifikante Fortschritte in mehreren Aspekten. Unser Empfehlungsprozess zur Abstimmung der Leistungsmetriken auf den tatsächlichen klinischen Nutzen spiegelt den Konsens eines internationalen Konsortiums von 73 Experten wider. Die Methodik des Wissenstransfers ermöglicht es dem System, Erfahrungen aus ähnlichen Datensätzen zu nutzen und übertrifft dabei bisher vorgeschlagene Alternativen auf unserem umfangreichen Benchmark. Unser Workflow zur Kompensation von Prävalenzverschiebungen verhindert Leistungseinbußen in diversen Bildgebungsverfahren und befähigt das System, Modelle automatisch an veränderte Umgebungsbedingungen anzupassen – ohne dass neue Annotationen erforderlich sind.

Diese Arbeit stellt die erste umfassende Untersuchung des Lifelong Learning in der biomedizinischen Bildanalyse mit zehntausenden trainierten und evaluierten Modellen dar. Durch das Hinzufügen metakognitiver Schleifen schaffen wir die Grundlage für autonome Lifelong Learning Systeme im Gesundheitswesen, die sich in wechselnden medizinischen Kontexten kontinuierlich weiterentwickeln können.

Acknowledgments

This thesis stands as a testament to the collaborative nature of academic research, made possible through the support and expertise of numerous individuals. I would like to extend my sincere appreciation to those who have contributed to this work in various ways.

First I would like to express my deepest gratitude to my supervisor, Prof. Dr. Lena Maier-Hein, for her invaluable guidance, continuous support, and remarkable patience throughout my research journey. Her trust in my skills, the freedom she granted me, the communication on an equal level, putting quality over quantity and the open management of the entire team will always be an inspiration to me. I am grateful for the repeated security in financing, shielding me from a lot of the background politics in research, and the understanding for all kinds of family restrictions. I will always remember your core lesson: '*strong beginning, strong ending*'!

This journey would not have been the same without the support and friendship of my colleagues at the division of '*Intelligent Medical Systems*' (IMSY) at the *German Cancer Research Center* (DKFZ). Two people in particular stand out here: Dr. Annika Reinke, my primary partner in the metrics project, and Piotr Kalinowski my companion in the prevalence shifts project. I would like to thank you for the stimulating discussions, the sleepless nights we worked together before deadlines, and all the fun we have had over the years. You are great scientists and people!

Next, my home team '*Intelligent Systems for Surgery and Endoscopy*' (ISSE) should be mentioned. I thank Dr. Tobias Ross and Dr. Tim Adler, my previous group leads, for their supervision and teaching – your traces are all over this document. Deepest appreciation to the help of my co-lead Dr. Keno März – you do an incredible job! I give you credit for being open to merge our groups and that you can take a step back without me ever having doubts about your willingness to support. I am particularly grateful to my mates Nuong Tran and Amine Yamlahi for their intellectual stimulation, inviting me to some of their projects, and staying with me in times when ISE was the smallest group in IMSY. The collaborative atmosphere of our group has been instrumental to this work. Special thanks to Dominik Michael for our numerous coffee break discussions and SMART hacking sessions, Dr. med. Minu Tizabi for her invaluable feedback on my manuscripts, Dr. Doreen Heckmann-Nötzel for all the support in writing grants and reports, Dr. Evangelia Christodoulou for all statistical advice, and all other ISSE members for always asking the right questions.

But I would like to express my heartfelt appreciation also to my colleagues outside ISSE, who created an environment of intellectual curiosity and mutual support. Our countless discussions, both formal and informal, have significantly shaped this work. Thanks to Matthias Eisenmann for being an awesome onboarding tutor and supervisor during my first research project, Maike Rees for years of re-designing my figures, Marcel Knopp for revising drafts, Dr. Melanie Schellenberg for pushing forward our shared vision of Good Scientific Practice, the IMSY office for being so helpful and approachable with any question and all organizers of retreats, Christmas symposiums, workshops, journal clubs, or other events. Please forgive me for not being able to list everyone, but you can still be sure of my gratitude.

My sincere appreciation goes to the '*Helmholtz Information & Data Science School for Health*' (HIDSS4health) for their support. Special thanks to Prof. Dr. Rainer Stiefelhagen and Dr. med. Hannes Kenngott for their constructive criticism and productive discussions as part of my thesis advisory committee.

I am deeply grateful to my family and friends for their unfailing support and continuous encouragement throughout my years of study. Particularly I would like to thank my parents for my initial scientific education, my brothers who accompanied me on much of my journey, my fellow students, especially Markus Schäfers and Jan Sieber, whose enthusiasm always fueled my passion for mathematics and computer science, and my mother-in-law Susan Godau who took care of my children incredibly often while I was writing these pages.

Finally, I could not have completed this dissertation without the support of my wife, who had to endure a lot of overtime and provided persistent 'encouragement'.

Contents

Lists

List of Acronyms

AC Accuracy 9
AI Artificial Intelligence . 229
AP Average Precision . . . 58
AUC Area under the curve . 58
AUROC Area under the Receiver Operating Characteristic Curve 58
BA Balanced Accuracy . . . 43
BIAS Biomedical Image Analysis Challenges 96
bKLD binned Kullback-Leibler Divergence 230
BPMN Business Process Model and Notation 116
BS Brier Score 69
BSS Brier Skill Score 69
BVM German Conference on Medical Image Computing 238
CE Calibration Error 64
CK Cohen's Kappa 46
CNN Convolutional Neural Network 73
CT Computed Tomography 17
CWCE Class-wise Calibration Error 66
DL Deep Learning 5
DNN Deep Neural Network . 17
EC Expected Cost 231
ECE Expected Calibration Error 65
ECE^{KDE} Expected Calibration Error Kernel Density Estimate 67
EHR Electronic Health Record 196
EMD Earth Mover's Distance 163
EQUATOR Enhancing the QUAlity and Transparency Of health Research 97
ER Error Rate 42
F1 F1-Score 53
FED Fisher Embedding Distance 166
FID Fréchet Inception Distance 165
FM Foundation Model . . . 233
FN False Negative 39
FNR False Negative Rate . . 41
FP False Positive 39
FDR False Discovery Rate . . 49
FOR False Omission Rate . . 48
FPR False Positive Rate . . . 41

GAN Generative Adversarial Network 86
GPT Generative Pre-trained Transformer 5
GPU Graphics Processing Unit 5
ImLC Image-level Classification 233
INN Invertible Neural Network 161
InS Instance Segmentation . 34
IOU Intersection over Union 53
IR Imbalance Ratio 22
J Youden's Index 48
JAC Jaccard Index 53
KCE Kernel Calibration Error 67
KLD Kullback-Leibler Divergence 69
LLM Large Language Model . 238
LR+ Positive Likelihood Ratio 47
LR- Negative Likelihood Ratio 47
MAML Model-Agnostic Meta Learning 80
MCC Matthews Correlation Coefficient 54
MCE Marginal Calibration Error 64
MICCAI Medical Image Computing and Computer Assisted Interventions 237
MIDL Medical Imaging with Deep Learning 95
MK Markedness 51
ML Machine Learning . . . 20
MLP Multilayer Perceptron . 72
MMD Maximum Mean Discrepancy 161
MONAI Medical Open Network for Artificial Intelligence . . 237
MRI Magnetic Resonance Imaging 17
NB Net Benefit 59
NEC Normalized Expected Cost 45
NN Neural Network 4
NeurIPS International Conference on Neural Information Processing Systems . . 237
NKLD Normalized Kullback-Leibler Divergence
NLL Negative Log Likelihood 68
NLP Natural Language Processing 5
NPV Negative Predictive Value 49
ObD Object Detection 34
PACS Picture Archiving and Communication System 19
PEFT Parameter-Efficient Fine-Tuning 233
PPV Positive Predictive Value 48
PR Precision-Recall 58
PSR Proper Scoring Rule . . 118
RBS Root Brier Score 70
RKHS Reproducing Kernel Hilbert Space 66
RNN Recurrent Neural Network 4
ROC Receiver Operating Characteristic 57
SDS Surgical Data Science . 223
SGD Stochastic Gradient Descent 71

SemS	Semantic Segmentation	238	**TNR**	True Negative Rate	41
TCE	Top-label Calibration Error	65	**TP**	True Positive	39
			TPR	True Positive Rate	41
TN	True Negative	39	**WCK**	Weighted Cohen's Kappa	46

List of Tables

List of Figures

Part I

Inception

Introduction 1

> *Any consistent formal system F within which a certain amount of elementary arithmetic can be carried out is incomplete; i.e., there are statements of the language of F which can neither be proved nor disproved in F.*
>
> — Kurt Friedrich Gödel

This chapter first motivates the research on *Lifelong Machine Learning* embedded in a historical overview of computer science and Artificial Intelligence (AI) in Sec. 1.1. Next, Sec. 1.2 derives and presents the three research questions at the core of this thesis. Finally, an outline of the structure of the thesis is provided in Sec. 1.3 to ease navigation for interested readers.

1.1 Motivation

The history of AI is essentially a story of increasingly sophisticated loops – a trajectory that traces back to Gödel's transformative insight into the capacity of formal systems for self-reference. Just as his 'First Incompleteness Theorem' [140] demonstrated that any sufficiently powerful formal system must necessarily contain statements that refer to themselves, modern AI systems achieve unprecedented skills precisely through their ability to represent and reason about their own operations. Hofstadter [170] links the self-referential capabilities of systems directly to the emergence of cognition in his Pulitzer Prize awarded book *Gödel, Escher, Bach: an Eternal Golden Band*: "It is an inherent property of intelligence that it can jump out of the task which it is performing, and survey what it has done; it is always looking for, and often finding, patterns.". The evolution of AI reflects multiple integrations of such self-analyzing loops, often marking a significant evolutionary step in system capability.

Universal machines The first mechanical devices that performed *computations* were designed for very narrow applications, such as the ancient Greek 'Antikythera mechanism' [121] to predict astronomical positions, the 'Pascaline' [287] to perform arithmetic calculations, or tide-prediction machines [385] from the late 19th century. Until then,

only simple loops in the form of interacting gears were possible. A groundbreaking milestone was set by Alan Turing in 1936 [389]. With mathematical precision, he described a theoretical computer and deduced the existence of 'universal machines' that could be programmed to perform *any* possible computation. While his results demonstrated the limits of computability[1], on the other hand he had described a fundamental conceptual loop: Using the same strategy as Gödel, he enabled machines to refer to themselves.[2] The first technical realizations of universal computers soon followed.[3] A key component, the 'von Neumann architecture' [406], which is still common in modern computers, implements the self-reference loop of machines more practically: Shared memory for data and instructions allows a machine to interact with its own code. Although it must be made clear that none of the theoretical or physical machines invented up to this point would be considered 'intelligent', in part because they lacked the ability to 'learn'.

Learning machines Based on the artificial neuron [246], the *Perceptron* [325] emerged in the 1950s as the first (partially) trainable multilayer Neural Network (NN). This machine attempted to perform a learning loop, as it back-propagated errors [324], allowing feedback from observed information on 'weights' of the model itself. However, its capabilities were severely limited because the outputs of its neurons were discrete levels, and its (single-layer) variant was shown to be incapable of learning even the XOR function [252]. This limitation may have contributed to the first AI winter and a shift to other approaches.[4] Thus, it took several decades to develop [229, 417] and popularize [330] the now common gradient estimation method of *backpropagation*. For the first time, a system could systematically modify its own internal representations based on experience and the 'backpropagation loop'. Dealing with a different kind of loop, the class of Recurrent Neural Network (RNN), became popular around the turn of the millennium [169]. While a feedforward network by design does not store any presented data[5], RNNs process data over multiple time steps, while each recurrent unit maintains a hidden state. Because of this memory, the insights from previous time steps can be used to process the current time step. However, limited computational resources and the lack of large datasets have

[1] In fact, the set of computable numbers in the interval $[0, 1]$ has measure zero, i. e., it is tiny compared to the uncountable set of potential problems.

[2] This is closely related to the idea of *self-replicating machines* [407]. Indeed, an early result by Kleene, known as his 'Second Recursion Theorem' [202], proves the existence of programs that produce their code as output for any Turing-complete programming language. The name *quine* for such programs was coined by Hofstadter [170].

[3] Noteworthy, the first conceptual universal machine was designed a whole century earlier, but has never been fully built [46].

[4] The field of AI has experienced several (major and minor) 'hype cycles'. Apparently, the topic tends to create exaggerated expectations in the public, followed by disappointment and criticism, which ultimately leads to funding cuts.

[5] It is important to distinguish the *learning phase* from the *inference phase* of such 'classical models'. Indeed, through backpropagation during the learning phase, information is stored through the weight updates – but when these are fixed in an inference setting, no information is kept 'alive' in the system.

prevented deeper architectures from reaching their full potential. Shallow NN are limited in their expressiveness and require feature engineering: The transformation of raw data into a more effective set of inputs – a time-consuming preprocessing step that requires human expertise.

Deep Learning In the late 2000s and into the 2010s, the 'advent of Deep Learning (DL)', enabled by Graphics Processing Units (GPUs), fueled by the large ImageNet dataset [93], and ushered in by AlexNet [210], finally overcomes the feature engineering bottleneck. Modern NNs do not just process pre-engineered features – they discover their own representations through multiple layers of transformation, effectively learning how to represent the raw data in a more powerful way. The following years are characterized by increasing model depth, advances in available computing hardware, and architectural optimizations. In parallel, the backpropagation loop is improved upon: machines no longer necessarily rely on external supervision signals. One way to mitigate the need for paired training data in *Supervised Learning* is given by the more general perspective of *Reinforcement Learning* [189]. There, the setting is vaguely characterized by an 'agent' performing 'actions' in an 'environment'. The environment is repeatedly interpreted, and a state representation as well as a reward is fed back to the agent. The goal of the agent is to learn a policy that maximizes the (expected) reward. This learning loop of the agent is much more interactive and allowed *Alpha Zero* [354] to master the games of chess, shogi and go solely through 'self-play'. (Pre-)training via *self-supervision* is another approach that facilitated the successful series of Generative Pre-trained Transformer (GPT) models [48] in Natural Language Processing (NLP). This 'task-independent' learning technique produced the groundbreaking potential of 'in-context learning' [48] – the ability to learn during *inference* – and drove the materialization of a new class of models in the late 2010s: Foundation Model (FM) [37]. The remarkable success of Large Language Model (LLM)s involves a different looping technique compared to 'classical models': context, queries and previously generated predictions are continuously looped through the model to generate the next prediction. Thus, the model predictions also depend on previously generated responses from the very same model. Apparently, exploiting these advanced loops (along with scaling and other improvements) lead to the emergence of new capabilities of AI. Self-referential loops – the ability of a system to represent, reason about, and interact with itself – appear to be more than just another 'feature' of intelligent systems. Rather, they may be the fundamental mechanism by which true intelligence emerges from simpler computational processes. This idea bridges Gödel's mathematical discoveries, Hofstadter's philosophical observations about strange loops [170], and the empirical success of modern AI architectures. When a system can leverage or improve a loop of self-reference – whether through mathematical self-proof, conscious self-reflection, or algorithmic self-modification – it transcends its original limits.

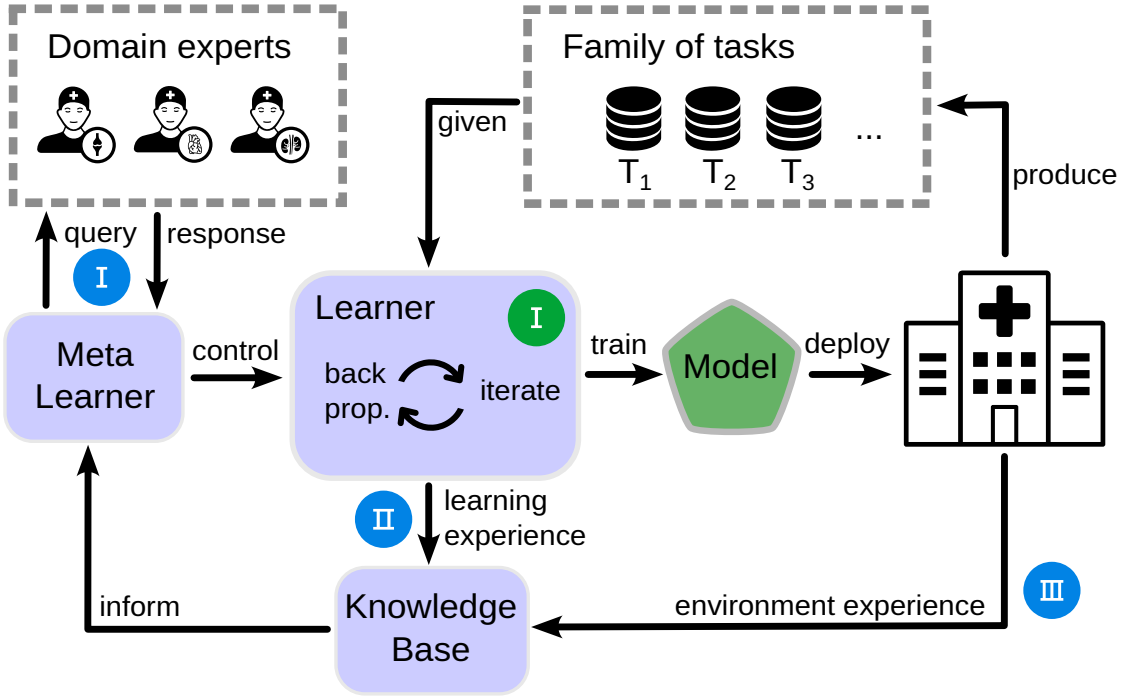


Figure 1.1: Components of (and loops within) a Lifelong Learning system. Given a task, the *Learner* trains a model using backpropagation (inner loop, green I). The *Meta Learner* controls the behavior of the *Learner* and may interact with domain experts, e. g., to design the reward function for the *Learner* (outer loop, blue I). Accumulated learning experience, e. g., successful training pipelines, are persistently stored in the *Knowledge Base*, which is accessed by the *Meta Learner* (outer loop, blue II). Deployed models, that interact with the environment can generate additional experience that can be used by the system (outer loop, blue III). Inspired by Fig. 1.2 from Chen and Liu [60].

Lifelong Learning In 1987, Schmidhuber [340] was the first to introduce methods that *nest* such loops – perceiving the very own weights of a NN as part of the inputs and predicting the updates thereof. The concept of 'learning to learn' or alternatively *Meta Learning* was born. The idea of *Transfer Learning* [278] as a fundamental part for such a *Lifelong Learning* system roots back to the 1990s.[6] Compared to the models available at the time, humans were incredibly efficient at learning new tasks from very few examples, given that they already had experience in a *related* task. For example, learning to play an instrument, given that they can already play another one. Without studying and eventually enabling such behavior in machines, the goal of general AI

[6] According to Pan et al. [278] a 1995 workshop at the International Conference on Neural Information Processing Systems (NeurIPS) [53] played an important role. In the political and social sciences 'Lifelong Learning' refers to the continuous, voluntary, and self-motivated pursuit of learning [114]. We will separate and formally define the different terms Lifelong Learning, Meta Learning and Transfer Learning in our sense in Sec. 2.8.

seemed unattainable. In an early definition of 'learning to learn' by Thrun et al. [386] it is informally characterized as follows. Given

1. a family of tasks,
2. training experience for each of these tasks, and
3. a family of performance measures (e. g., one for each task),

an algorithm is said to be 'learning to learn' if its performance on each task improves with experience *and* with the number of tasks. The key difference from to regular learning (which is restricted to one task, hence also called *isolated learning*), is that the algorithm must improve not only with experience (e. g., training epochs), but also with the number of tasks available to it. Somehow the experience must be *transferred* between tasks and benefit the individual learning processes. A conceptual representation of a Lifelong Learning system in given in Fig. 1.1.

Queries A revealing perspective on (self-referential) loops may be provided by **queries**. A 'self-aware' system, may ask queries of itself, e. g., a universal machine, querying parts of its 'source code' to perform analysis and afterwards modify it with custom optimizations. More generally, on a broader view of a Lifelong Learning system, queries can also be directed to other entities – allowing for interactive engagement of the system with its environment. In Fig. 1.1 the system can interact with domain experts (blue I), a central *Knowledge Base* (blue II), and the deployment environment (blue III). In principle, such a system must decide **what** information it is interested in, **how** it can query the information, and finally, considering a volatile environment and potential costs of queries, the system must know **when** it is beneficial to ask its questions and process the response. Necessarily, a Lifelong Learning system, in contrast to isolated learning, is in constant exchange with the *environment* of the system, because of the 'stream' of incoming tasks.[7] It can observe the consequences of its output (which is also a continuous stream), interact with other systems ('agents') and be confronted with new instructions ('tasks'). According to the 'embodiment hypothesis' [360], enabling a system to take actions within its environment and to sense the results is a driver for the emergence of intelligence.

Life cycle Another contrast to isolated learning – including RNNs and Reinforcement Learning – is that the classic 'AI lifecycle' is looped over repeatedly: New tasks arrive, models are generated, and put into action. According to De Silva et al. [89], the AI lifecycle can be divided into three phases, which we summarize as follows:

(i) **Design**: Identification and formulation of the problem. Data preparation, exploration, and further acquisition.

[7] Compared to the informal definition by Thrun et al. [386] given above, in practice there is a temporal aspect to the appearance of tasks.

(ii) **Develop**: Building and executing a model training pipeline. Benchmarking multiple models.

(iii) **Deploy**: Risk assessment and post-deployment review. Monitoring and performance evaluation.

Except for the execution of the model training pipeline, these steps are primarily driven by human decisions in isolated learning. An autonomous and holistic Lifelong Learning system must perform most of these steps by itself and should therefore make informed decisions in all three phases.

Outlook Looking toward the horizon of AI development, we see multiple threads converging toward the grand challenge of 'Lifelong Machine Learning' [60]. Specifically Meta Learning [172] – the ability to learn how to learn – could emerge as a foundational self-referential loop in modern AI. Current systems already demonstrate some Meta Learning capabilities as in-context learning [48], but several critical loops remain to be resolved to achieve true Lifelong Learning capabilities.

I. A **reward-learning loop** that deals with the ability to refine learning objectives, e. g., choose appropriate validation metrics.

II. A **pipeline-learning loop** that allows the system to evolve its own learning strategies, e. g., a selection mechanism for neural architectures.

III. An **environment-learning loop** that feeds back experience from environment interaction and equip the system to adapt to changes in the surrounding, e. g., when the system is transferred.

Such 'metacognitive' loops would likely empower the system to reason about its learning strategies and dynamically adjust them based on context and experience, while actively and repeatedly querying the environment. This vision of a system that can truly learn throughout its operational lifetime, constantly improving and adapting, marks a leap up the evolutionary ladder of AI.

1.2 Research questions

We want to shift the focus to the biomedical imaging domain and put our vision of a Lifelong Learning system in the context of the current challenges in this application. Despite the rapid pace of AI research and the recent breakthroughs in DL, there are only few success stories that translate these advances into patient care in the clinic [193, 239, 280, 348]. The outstanding challenges to achieving transformational benefits are many:

(i) Alignment of performance metrics with the clinical purpose [193, 348],

(ii) a lack of robustness and generalization primarily due to data sparsity [239, 372],

(iii) on-site infrastructure for data management [239, 280],

(iv) a cultural shift that includes continuing education for clinicians [116, 239],

(v) interpretability/explainability of predictions [193, 239],

(vi) and many more.

In the remainder of this section, we will outline the specific research questions of this thesis and show how the metacognitive loops we motivated in Sec. 1.1 might be suited to address open challenges (i) and (ii) in particular. We will focus our theoretical and experimental analysis on a single category of image analysis problems: **Image-level Classification (ImLC)**. ImLC is highly relevant in the diagnostic decision process on medical images with a variety of applications [231]. This simplification has several advantages: First, ImLC is a fundamental problem type in image analysis beyond medical images that has been extensively studied. This allows us to compare ourselves to a large corpus of previous research, while providing us with a rich toolbox of existing open source software to use. Second, the comparatively low effort required to annotate images for this type of problem ensures sufficient availability and variability of data and use cases to conduct experiments that test the generalizability of our methods. Third, because pretraining on ImageNet [93] is particularly useful for this type of problem [176, 243], it generally allows for faster convergence of models [310], making our experiments feasible on a larger scale.

Research Question 1 (RQ1)

Neglecting the individual context of a medical application by choosing a performance measure based on popularity can lead to useless models [242]. Consider, for example, the most common metric, Accuracy (AC), which simply counts the proportion of correctly mapped instances of a model. Relying on AC in a medical screening scenario with very few positive cases, say one out of 100, may lead to an overestimation of model applicability, since consistently predicting the negative class already achieves 99% AC. Furthermore, it would neglect the potentially different consequences of type I and type II prediction errors for individual patients. The general need for better evaluation practices in medical image analysis has also been highlighted in the literature. Maier-Hein et al. [240] conducted a survey on evaluation issues in international biomedical competitions, which identified the selection of performance metrics as the most pressing issue. The need for a systematic understanding of model evaluation was well summarized in the call by Kelly et al. [193], who emphasized that performance measures should "capture real clinical applicability and be understandable to intended users".

Given a particular biomedical image classification application, which family of performance measures (see Sec. 1.1) best reflects the driving medical need? This question

is critical in the **Design** phase of AI. It must be answered before any model can be examined in isolation or any set of models can be compared during the **Develop** phase. Otherwise, no conclusions can be drawn about progress toward the application goals. From a metacognitive perspective, this corresponds to the sketched **reward-learning loop**: Can a *Meta Learner* improve its evaluation mechanisms to produce models of higher quality from an application perspective? We strive to integrate the metric selection as a *systematic* procedure that can be performed by the *Meta Learner* interactively with the knowledge of the application domain experts. For example, we need to answer how intrinsic properties of the data affect specific performance measures. Or, how interests from the medical domain can be mapped to (or reflected in) the performance evaluation. In summary, we ask the following question:

Research Question 1

How can clinical objectives be systematically translated into appropriate AI model validation metrics?

Research Question 2 (RQ2)

Generalizability is a non-negotiable requirement for model applicability in healthcare, but in contrast, the strict regulations on personal health data and the high cost of data annotation only allow model training with sparse data [239]. Incorporating experience from related datasets is an intuitive approach to cope with the very limited samples provided by each individual task in such a scenario. Based on the informal definition of 'learning to learn' (see Sec. 1.1), this idea is at the heart of Lifelong Learning. Recall that knowledge must somehow be *transferred* between individual tasks. The simplified case of knowledge flowing from a 'source task' to a single 'target task' is called *Transfer Learning*. Pan et al. [278] postulate the following three main research issues in Transfer Learning:[8]

1. **What to transfer?** – what part of the knowledge can be transferred across tasks.
2. **How to transfer?** – asks what algorithms need to be developed to transfer the knowledge.
3. **When to transfer?** – in which situations is transfer beneficial, and in which ones may it be harmful ('negative transfer').

A concrete example of Transfer Learning is *fine-tuning*[9]. For an intuitive explanation

[8] Note the close relationship between the posing of *queries* and *transfer of knowledge*: Every cycle of questions (including some answers) constitutes knowledge transfer.

[9] Fine-tuning is indeed the most common variant of Transfer Learning, and often the term 'Transfer Learning' itself is used as an (imprecise) synonym for fine-tuning or, slightly more generally, the transfer of learned parameters [158]. A more precise definition of fine-tuning is given in Def. 2.87.

of this technique, let $\mathcal{S}$ be the source task and $\mathcal{T}$ be the target task. Given the data distribution of $\mathcal{T}$ is 'sufficiently' covered by $\mathcal{S}$ (**when**), the learning machine first learns only on $\mathcal{S}$, where the resulting model parameters (**what**) are then used as initialization for the learning on $\mathcal{T}$ (**how**).

In the Lifelong Learning scenario with multiple potential source tasks, we are primarily interested in a quantifiable measure of 'relatedness' of tasks to answer the **when** question in advance. Furthermore, we are interested in which other parts of the learning pipeline can be optimized based on experience from other tasks (**what**)? And we will explore the additional constraint of actual data availability during transfer, motivated by the limitations of data sharing in the medical domain (**how**). Is it possible to share knowledge in a collaborative environment where network participants are not allowed to share sensitive data of a task?

The second research question follows the line of thought of the **pipeline-learning loop**, which improves the individual learning of tasks by adaptation based on other pipelines and tasks. It is at the core of the **Develop** phase of any model and determines the strategies for training pipeline construction. Overall we formulate our second research question as follows:

Research Question 2

How to enable effective knowledge transfer across biomedical image analysis tasks?

Research Question 3 (RQ3)

Lastly, we strive to improve the methodology once a model enters the final phase of **Deployment**. Panch et al. [280] noted "the inconvenient truth" about AI in healthcare: algorithms that excel in research are not executable in clinical practice. They attribute this in part to the fact that healthcare organizations lack the data infrastructure needed to (i) adapt algorithms to local populations and practice patterns and (ii) validate them for biases, especially when patient cohorts may have been inadequately represented during model training. Any model trained and validated on a particular data distribution will inherently pick up certain biases of that distribution and will face problems if that underlying distribution shifts in an application setting [372]. This scenario is commonly known as *dataset shift* [261]. In the sense of our envisioned **environment-learning loop**, a system that detects and corrects such shifts would be of great value. Unfortunately, these shifts can be of very different types and causes, with some dataset shifts doomed to be resolved only with tedious intervention [54]. *Prevalence shifts* are a specific, frequently encountered type of dataset shifts that may severely impact a model. By restricting ourselves to this type of dataset shift we can solve the problem elegantly and in line with the observed limitations of healthcare organizations' data infrastructure, as mentioned by Panch et al. [280]. Finally, we formulate our third research question as follows:

Research Question 3

What mechanisms enable biomedical imaging models to detect and compensate for prevalence shifts in deployment?

1.3 Outline

This dissertation consists of eight chapters. After the thematic introduction in Chap. 1, Chap. 2 provides the necessary background information on the medical side, specifically the translational obstacles the field faces and the imaging datasets we use. It continues with basic concepts their respective notions for this thesis. This is followed by a detailed presentation of a variety of existing metrics for validating classification models. Next, an introduction to selected machine learning techniques used for deep neural networks and formal definitions of the various learning paradigms we mentioned earlier are given.

In Chap. 3, we present the relevant state of the art in the context of this thesis and conclude with a discussion of prevailing limitations and open challenges. We summarize related work for validation practices of predictive models in medical image classification, compare existing approaches for the training of deep neural networks in sparse data settings, and provide an overview of deployment considerations under dataset and prevalence shifts.

The following three chapters are each dedicated to one of our core research questions and are identically structured into methodological exploration, experimental results, and a discussion. Our framework for application-oriented validation of image classification algorithms is presented in Chap. 4, answering **(RQ1)**. Chap. 5 focuses on our solution to overcome the learning boundaries of individual datasets and to share learning experience between collaborators in a potentially data-sensitive environment, answering **(RQ2)**. Next Chap. 6 elaborates on our insights from deploying classification models under prevalence shifts, answering **(RQ3)**.

In Chap. 7, we bring the different threads back together to an overall discussion of the results of this thesis, their limitations and implications in a general context. Finally, Chap. 8 concludes this work by summarizing our contributions to the research questions and providing an outlook on potential future directions of Lifelong Learning in healthcare.

Tab. 1.1 shows which chapters of the thesis address which of the three research questions.

Table 1.1: Outline overview table. Relation between our three research questions, the corresponding thesis chapters, AI lifecycle phase and meta-loop. We phrase each of our concepts as a knowledge transfer, answering **what** knowledge can be transferred, **how** a transfer is enabled and **when** it is beneficial. For illustration, we formulate sample queries from the *Meta Learner* for each loop.

	(RQ1)	**(RQ2)**	**(RQ3)**
Chap.	Chap. 4	Chap. 5	Chap. 6
Phase	Design	Develop	Deploy
Loop	I. reward	II. pipeline	III. environment
		Knowledge transfer	
what?	formalized application goals & task properties	previous training hyperparameters & additional data	model predictions from deployment
how?	interview domain experts & derive proper metrics	integrate previously successful settings into training	quantify class prevalences & post-hoc adjustments
when?	must be done for each new task	only if tasks are related	discrepancy from training prevalence
		Query	
example	'Are some class confusions more severe?'	'What pipelines worked well for similar tasks?'	'Have class prevalences shifted from training?'

Disclosure of Contributions

The research presented in this thesis is the product of interdisciplinary work with contributions from various team members and collaborators. While this thesis was written independently by me, it uses the 'we' form rather than the 'I' form to reflect the collective efforts of all involved. For the sake of transparency, App. A provides an overview of my personal contributions to the research questions and the corresponding publications.

Part II

Environment

Fundamentals 2

This chapter presents the prerequisites of biomedical imaging and the algorithms used to analyze it. We start with a primarily medical perspective in Sec. 2.1, which assesses the main open challenges in biomedical imaging, and a collection of concrete medical applications for this work in Sec. 2.2. We then switch to a predominantly theoretical perspective, formalizing the problem in a rigorous manner to describe concepts with mathematical precision (Sec. 2.3). The following sections describe in detail a multitude of performance measures that play an important role in this thesis. Specifically, we introduce counting metrics (Sec. 2.4), multi-threshold metrics (Sec. 2.5), and notions of model calibration (Sec. 2.6). Next, we introduce the main concepts of Deep Neural Network (DNN) (Sec. 2.7), today's dominant class of models used to solve image processing tasks. Finally, Sec. 2.8 zooms out of the isolated training for single tasks and defines a variety of paradigms around the idea of knowledge transfer and 'learning to learn'.

2.1 Translational obstacles in biomedical imaging

We begin this section with a brief summary of the history of biomedical imaging and then discuss the prevailing translational obstacles for the success of AI in this field.

2.1.1 History of biomedical imaging

The foundations of modern optical imaging were laid in the late 19th century by Ernst Abbe, who revolutionized microscope design through precision lens manufacturing and theoretical optics [2], enabling unprecedented clarity and magnification. The history of biomedical imaging then unfolded with Wilhelm Röntgen's groundbreaking discovery of X-rays in 1895 [322], which for the first time allowed physicians to visualize internal body structures without invasive procedures.

The mid-20th century saw transformative advances. In the 1940s, ultrasound was first used to image the human body, but it was not until the 1960s that ultrasound became widely available for medical use. In 1972, the first commercially viable Computed Tomography (CT) scanner was invented by Godfrey Hounsfield [174]. At the same time, the conceptual foundations of Magnetic Resonance Imaging (MRI) were laid by Paul C.

Lauterbur, who developed a mechanism for encoding spatial information into a nuclear magnetic resonance signal using magnetic field gradients [216].

According to Litjens et al. [231], medical image analysis evolved in three phases. The field began in the 1970 with rudimentary image processing techniques like edge detection and shape fitting, which were used to construct rule-based systems for specific analytical tasks. By the late 1990s, medical image analysis transitioned to supervised learning methodologies, where systems were no longer entirely human-designed but instead learned from training data. During this period, computers analyzed feature vectors from sample data to determine optimal classification boundaries. The third major shift occurred after the groundbreaking DL success by Krizhevsky et al. [210] in 2012, which triggered a rapid and comprehensive adoption of DL approaches. Soon after DL permeated every aspect of the field, fundamentally transforming how medical images are analyzed across all applications.

In 2016, Geoffrey Hinton made the following comment about radiology and DL at the '2016 Machine Learning and Market for Intelligence Conference' in Toronto: "*People should stop training radiologists now – it's just completely obvious within 5 years deep learning is going to do better than radiologists. It might be 10 years, but we've got plenty of radiologists already.*"[1] While it is easy to take a Nobel Laureate's words out of context and accuse him of making a false prophecy, the more intriguing question is why his prediction has not (yet) come true. Given the rapid progress and amazing success of *generative* DNNs in the fields of NLP [250] and imaging [81], we want to delve deeper into the challenges of *discriminative* models in the medical domain.

Failure stories

While the real-world performance degradation of many AI systems remains undisclosed for commercial reasons, several notable failures have been documented. Breast cancer detection in mammography, despite being one of the oldest computer-aided medical imaging applications (first FDA approval in 1998 [221]), continues to underperform: A 2021 review [120] found that 94% (34/36) of AI systems were outperformed by individual radiologists, and none matched the accuracy of multiple radiologists' consensus. Commercial AI-based skin cancer detection via smartphones faces significant bias issues across different skin tones [86]. Google Health's diabetic retinopathy detection system trial in Thailand [25] struggled with poor lighting and image quality. The system rejected 21% of submitted images and showed reduced accuracy for those it did accept. In addition, cloud-based processing caused delays that reduced daily patient throughput. These anecdotal examples can be placed into a meta-perspective where evidence is given that the majority of research findings are false [179] and presented results are not reproducible [23].

[1] A recording of his words can be found at `https://www.youtube.com/watch?v=2HMPRXstSvQ`.

Proving health benefits

Medical imaging models will be involved in life-threatening decision-making processes, which requires clinicians to be trained with the AI system and have a basic understanding of how AI arrives at diagnostic or prognostic predictions in order to trust and validate the model reasoning [70]. Since model predictions may trigger a cascade of follow-up treatment, affect long-term well-being, or become critical to survival, they require approval of safety and efficacy. Randomized controlled trials, the gold standard in medicine, are challenging: the need to integrate with existing systems and meet stringent regulatory standards pose significant implementation challenges. On top of that, assessing the long-term impact of AI interventions on patient health outcomes requires extended periods of observation and follow-up, which can be resource-intensive and time-consuming – while the rapid evolution of the technology increases the risk of rapid obsolescence. These hurdles explain the relatively small number of peer-reviewed randomized controlled trials [193].

While these considerations may partially explain the 'gap' in AI success between the medical and non-medical domains, *within* the medical domain these obstacles have existed for medical devices for years, allowing the industry to build up expertise in handling such translational efforts. Implementing regulatory compliance and organizing control trials should therefore not be a roadblock even if it slows down translation. Obviously, there is a deeper divergence between *research* claiming 'superhuman' performance and actual patient benefit. One of these divisive factors is, that healthcare AI research occurs predominantly outside of actual clinical settings, limiting its real-world applicability [196]. Academic measures of success, such as publications and citations, differ from clinical impact measures, which focus on real-world adoption rates and patient outcomes. True success in medical AI is determined by the number of hospitals implementing the model and the quantifiable improvements in patient care, as opposed to the predictive performance of an algorithm [399]. This mismatch of incentives and measures of success leads to the prevalence of 'wrong and useless models' [242].

Data scarcity

While massive medical image databases exist in hospital Picture Archiving and Communication System (PACS) systems, the main challenge is not the existence of the data, but the data preparation process. For example, systems store many reports as free text, requiring sophisticated text mining to extract structured labels [231]. Willemink et al. [423] list eight steps involved in handling medical image data: Ethical approval, access, querying, de-identification, transfer, quality control, structuring, and labeling. Overall, the data curation process is costly and time-consuming, limiting access to large, diverse training datasets with expert annotations. As a result, medical imaging datasets lack behind the ones from natural image recognition in computer vision by two to three orders of magnitude [399]. Two other factors come into play, that complicate the use of typical

medical imaging datasets compared to general computer vision. First, the label noise introduced by uncertainty about a ground truth reference is much higher, and second, the classes tend to be much more imbalanced [231]. Compared to a clinician's perspective, typical datasets also contain a very small window of information about patients: medical history and the combination of multiple modalities adds a lot of valuable context to a given image that current systems are unable to incorporate.

Distribution shifts

Finally, there is a fundamental assumption in the translation of research results into clinical practice, which can be found in the precise mathematical formulations of error estimation: data should be *i.i.d.*, i. e., *independent and identically distributed*. But the data used during model development rarely match the local population and/or the local practice patterns [280]. In fact, to ensure the i.i.d. condition for model validation, the common practice in Machine Learning (ML) studies is to sample training and test data from the same data pool [399], which itself may not be representative. In part, this situation can be attributed to the disparate and heterogeneous data silos that result in few samples and thus sparse coverage of a 'general data distribution'. Castro et al. [54] advocate taking the causal relationships into account when making data selection and annotation decisions. Such relationships can be visualized using causal diagrams, i. e., directed acyclic graphs which map the cause-effect relationships of the variables involved [130]. A generic causal diagram for medical imaging is shown in Fig. 2.1[2].

The causal diagram shows four standard prediction tasks based on an image of a specific patient: Two of them are *causal predictions*, which follow the causal flow and require predicting image annotations (e. g., outlines of brain lesions in an MRI) or the referral (e. g., whether to perform a tumor resection or conduct chemotherapy alone for colorectal cancer). Contrary *anticausal predictions* must go against the causal direction and predict patient characteristics (e. g., gender or age from an X-ray) or diagnosis (e. g., categorization of cervical lesions in colposcopy based on biopsy). Changes in the environment, such as moving an AI system to a new hospital, can have downstream effects on the prediction tasks and are called *distribution shifts*[3]. A different population around the hospital or a given specialization that leads to selective referral of certain cases leads to different patient characteristics. In the causal case this scenario is called a *population shift*, while in the anticausal case it is called *prevalence shift*. Another example would be the difference in image acquisition by a newly introduced imaging device, leading to an *acquisition shift*. Each of these changes violates the assumption of identical data distribution and, if not being taken care of, reduces the performance of a deployed AI model [54].

[2]Precisely the given diagram is a *selection diagram* [291], that includes the special indicator variables for the *environment* and *sample selection*.

[3]Formal definitions of the sketched *distribution shifts* are given in Sec. 6.1.1.

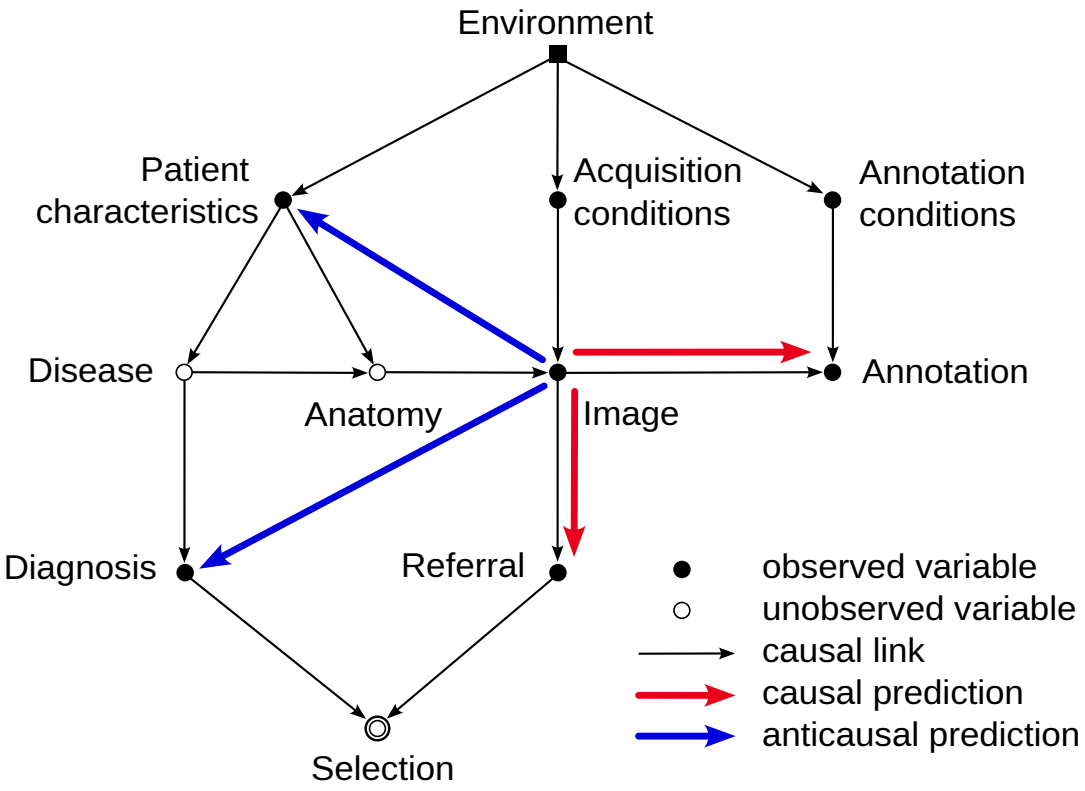


Figure 2.1: General causal diagram for medical imaging prediction tasks. External factors from the environment (top) can affect all entities in medical imaging: patient, image and annotation. Four typical prediction tasks are marked as colored arrows, starting from the medical image (model input) and pointing towards the entity to be predicted. The color indicates whether these tasks follow the causal direction (red) or not (blue). Ultimately a subset of samples is selected (bottom) as part of development datasets, which also may cause distribution shifts to ultimately observed deployment samples. Adapted from Castro et al. [54].

2.2 Datasets

We continue by giving an overview on the datasets that are used in this thesis. A concise overview is also provided in Tab. 2.1. We note the high variability in task size (min:170, q1:1572, q3:40 673, max:122 138), number of classes (min:2, q1:2, q3:5, max:257), Imbalance Ratio (IR) (min:1, q1:1.34, q3:14.27, max:171.33, see Def. 2.4) and imaging modalities.

Figure 2.2: Sample image from MURA [311].

Sonography The **dataset of breast ultrasound images** [95] contains 780 ultrasound images collected from 600 female patients (aged 25-75) at Baheya Hospital in Cairo, Egypt during 2018. The dataset is designed to aid in breast cancer detection and classification, with images categorized into three groups: normal breast tissue, benign tumors, and malignant tumors, each image having dimensions of approximately 500 × 500 pixels.

X-ray The **MURA** dataset [311] contains 40 561 radiographic images taken from multiple angles, representing 14 863 different medical studies of 12 173 patients. The images show seven different types of upper body x-rays: elbow, finger, forearm, hand, humerus, shoulder, and wrist. Between 2001 and 2012, board-certified radiologists at Stanford Hospital reviewed these images during routine diagnostic work and classified each study as either normal or showing abnormalities. The **Zhang Chest X-Ray Images** dataset [195] consists of chest X-ray images used for pneumonia diagnosis, collected from pediatric patients (ages 1-5) at the Guangzhou Women and Children's Medical Center. These anterior-posterior chest X-rays were gathered during routine clinical care and published with the goal of helping diagnose pneumonia – a serious lung infection that remains the leading cause of death from infectious disease in children under 5 years old worldwide. Parts of this dataset were also used in the **Kaggle COVID X-Ray** Dataset [164] for Covid-19 classification.

The **CheXpert** dataset [181] is a large collection of chest radiographs used for automated chest X-ray interpretation. It consists of 224 316 chest radiographs from 65 240 patients, including both frontal and lateral views. The dataset was collected from Stanford Hospital between October 2002 and July 2017. Each radiograph is labeled for the presence of 14 common chest radiographic observations. In this thesis we will treat this dataset as 13 separate tasks (ignoring the 'No Finding' observation). Since some of the attached labels are indicated as 'uncertain' or 'unmentioned' we will ignore corresponding samples for the respective tasks.

The **Shenzen Hospital CXR** dataset [185] is a collection of 662 high-resolution chest X-rays (approximately 3000 × 3000 pixels) captured at Shenzhen No.3 People's Hospital

in China during September 2012, using a Philips DR Digital Diagnost system. The dataset is nearly evenly split between normal cases (326) and cases showing tuberculosis (336), including both adult and pediatric frontal chest X-rays. Tuberculosis is a potentially deadly bacterial infection primarily affecting the lungs, caused by Mycobacterium tuberculosis. Chest X-rays are crucial for tuberculosis screening as they can reveal characteristic signs like upper lobe infiltrates, lung cavities, miliary patterns (tiny spots throughout the lungs), and pleural effusions (fluid around the lungs).

MRI Brain tumor detection and classification through medical imaging is crucial for effective treatment planning and patient outcomes, with MRI being the gold standard due to its excellent soft-tissue contrast and lack of radiation exposure. The **Brain Tumor Type Classification** dataset [61] is based on T1-weighted contrast-enhanced MRI images, collected from 233 patients at two Chinese hospitals between 2005 and 2010 and contains 3064 image slices showing three types of brain tumors: meningiomas (708 slices), gliomas (1426 slices), and pituitary tumors (930 slices). Complementary the **Kaggle Brain Tumor Classification** dataset [36] is a derivative from the data of the BRATS2015 challenge and comprises 3762 brain MRI. The images are categorized into two classes with present or no present brain tumor.

Figure 2.3: Sample image from CheXpert [181].

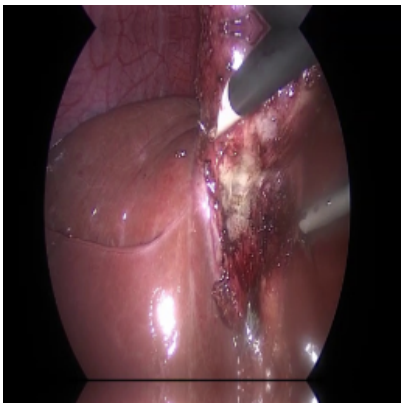


Figure 2.4: Sample image from Cholec80 [390].

CT The **COVID-CT-Dataset** [431] is a public medical imaging collection comprising 746 CT scans, with 349 images showing COVID-19 findings from 216 patients and 397 images without COVID-19 findings. The COVID-positive images were meticulously extracted from 760 COVID-19-related preprints published on medRxiv and bioRxiv between January 19th and March 25th 2020, using software to preserve image quality and extract associated clinical information. The negative cases were gathered from multiple medical imaging databases including MedPix, LUNA (lung cancer dataset), Radiopaedia, and PubMed Central, ensuring a diverse representation of non-COVID lung CT scans.

Laparoscopy The **Cholec80** dataset [390] is a comprehensive collection of laparoscopic cholecystectomy surgery videos created by the CAMMA (Computational Analysis and Modeling of Medical Activities) research group in

collaboration with the University Hospital of Strasbourg, IHU Strasbourg, and IRCAD. It comprises 80 high-quality videos of gallbladder laparoscopic surgeries performed by 13 surgeons. While the original publication comprised annotations for surgical phases and as well as the instruments present in the scene, multiple additional annotations have been released on top (e. g., surgical action recognition, critical view of safety, Semantic Segmentation (SemS), smoke detection). Because of its size and the diversity in annotations, the Cholec80 dataset has become a popular benchmark for research in computer-assisted interventions. We will leverage the original labels for the presence of 7 surgical instruments at one frame per second in the way of 7 binary image based tasks.

The **LapGyn4** dataset [222] is a comprehensive collection of laparoscopic gynecological surgery images, designed to support research in automated surgical video analysis and divided into four specialized subsets: (i) 'Surgical Actions' (30 000+ images): Documents general surgical activities and instrument usage during procedures, (ii) 'Anatomical Structures' (2700 images): Shows clear views of various pelvic organs, particularly useful for endometriosis treatment documentation, (iii) 'Specific Actions on Anatomy' (1000+ images): Focuses on particular surgical techniques (like suturing) performed on specific organs (uterus, ovary, vagina), and (iv) 'Instrument Count' (21 000+ images): Contains scenes showing varying numbers of visible surgical instruments (zero to three), useful for surgical phase identification. Some images of the instrument count dataset are taken from Cholec80.

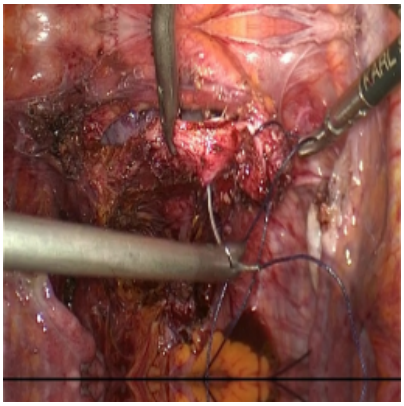


Figure 2.5: Sample image from LapGyn4 [222].

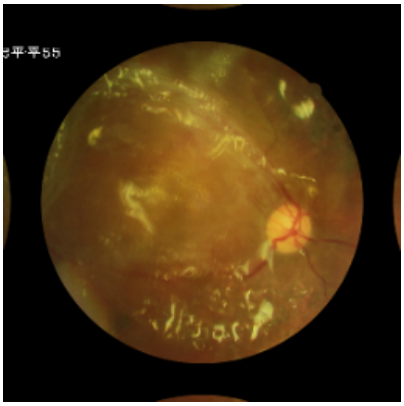


Figure 2.6: Sample image from DeepDRiD [232].

Fundus photography Fundus photography is a specialized medical imaging technique that captures detailed photographs of the interior surface of the eye, including the retina, optic disc, macula, and posterior pole. These images are crucial for documenting eye conditions, tracking disease progression, and aiding in the diagnosis of various eye disorders such as diabetic retinopathy, glaucoma, and age-related macular degeneration. The **DeepDRiD** (Deep Diabetic Retinopathy Image Dataset) [232] is a comprehensive collection of retinal images used for diabetic retinopathy research and machine learning applications. It is made up of 2000 regular fundus images from 500 patients and has annotations for diabetic retinopathy grading and image quality. We leverage five tasks from DeepDRiD: The binary task of overall image quality,

the 5-class problem of grading diabetic retinopathy, another 5-class problem of image clarity, yet another 5-class problem of the field definition and the 6-class problem of image artifact detection. The **APTOS2019** dataset [191] was created for a Kaggle competition to solve the same 5-class diabetic retinopathy grading problem and discussed at the 4th Asia Pacific Tele-Ophthalmology Society (APTOS) Symposium. We also incorporate another **cataract classification dataset from Kaggle** [55] that comprises about 600 images and distinguishes four categories: (i) normal (ii) cataract (iii) glaucoma (iv) retina disease.

Gastrointestinal endoscopy The **AIDA-E Barrett's esophagus** [13] (Analysis of Images to Detect Abnormalities in Endoscopy) dataset was collected at two Italian cancer institutes (IEO Milan and IOV Padova) during routine endoscopic surveillance of Barrett's Esophagus (BE) patients. The dataset contains 262 high-quality confocal images from 32 patients across 81 biopsy sites, captured using a Pentax confocal laser endoscope that enables real-time cellular imaging ("virtual histology") during endoscopy procedures. It was part of a challenge on Barrett's Esophagus Diagnosis, a precancerous condition where normal esophageal tissue is replaced by intestinal-type tissue with goblet cells. While traditional endoscopy with random biopsies has limited accuracy in detecting early cancer development, these confocal images were captured using fluorescein dye and have high resolution. Each image in the dataset is classified into one of three categories based on histological findings: (i) gastric metaplasia, (ii) proper Barrett's esophagus/intestinal metaplasia, or (iii) neoplasia.

The **Nerthus** dataset [299] contains 5525 annotated frames from 21 colonoscopy videos, each showing different levels of bowel cleanliness as defined by the Boston Bowel Preparation Scale (BBPS). The dataset was created at Bærum hospital in Norway to help develop automated systems for assessing bowel preparation quality, as this is crucial for successful colonoscopies but currently relies on subjective doctor assessments, with the goal of standardizing evaluation criteria and improving healthcare resource allocation.

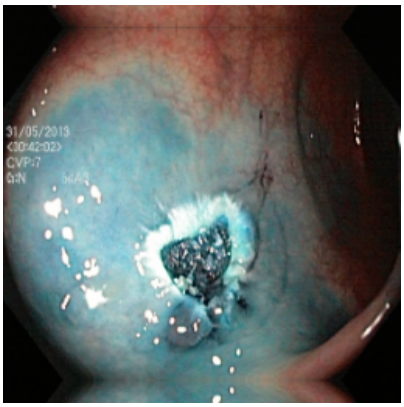


Figure 2.7: Sample image from Hyperkvasir [38].

The **Hyperkvasir** dataset [38] contains 10 662 gastrointestinal endoscopy images collected during routine examinations at the same Norwegian hospital between 2008 and 2016. The dataset evolved from an initial collection of 4000 images across 8 classes (called 'Kvasir'), was later doubled to 8000 images, and finally expanded to its current size covering 23 classes organized into four main categories: (i) 'Anatomical Landmarks' – Notable features in the upper and lower GI tract that help doctors navigate and confirm complete examination coverage, (ii) 'Mucosal View Quality' – Images showing different levels of mucosal visibility, classified using the Boston Bowel Preparation Scale (BBPS), (iii) 'Pathological Findings' – Documented

abnormalities and disease-related changes in the intestinal wall mucosa, categorized according to World Endoscopy Organization standards, and (iv) 'Therapeutic Interventions' – Images showing various treatment procedures like polyp removal, stenosis dilation, and bleeding ulcer treatment.

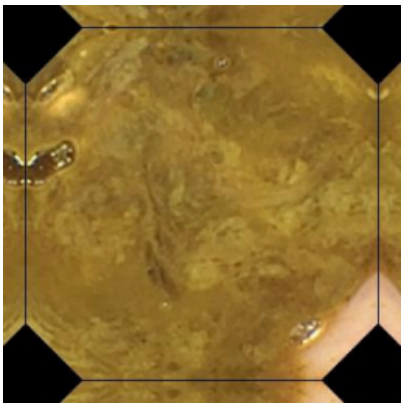


Figure 2.8: Sample image from Kvasir-Capsule [358].

The **Kvasir-Capsule** dataset [358] consists of 47 238 images collected through video capsule endoscopy, a technique where patients swallow a small capsule containing a camera and other electronics that records footage as it travels through their digestive system. The classes distribute across (i) seen anatomy, (ii)content of the bowel lumen, and (iii) the aspect of the mucosa and mucosal lesions (pathological findings). We removed 4 of the classes due to their very small size and split the dataset into three tasks according to the label clusters.

Laryngoscopy The **Laryngeal Cancerous Tissues** dataset [255] comprises 1320 tissue samples taken from laryngeal regions, including both healthy tissue and tissue showing early signs of cancer. These samples, each measuring 100 × 100 pixels, were carefully selected from narrow-band laryngoscopic examinations of 33 patients who were later diagnosed with laryngeal spinocellular carcinoma through histopathological testing. The dataset is evenly divided into four categories, with 330 samples in each: (i) Healthy tissue, (ii) tissue showing hypertrophic blood vessels, (iii) tissue with leukoplakia and (iv) tissue displaying intrapapillary capillary loops.

The **NBI-InfFrames** dataset [256] was developed to help researchers in surgical data science identify and classify the quality of endoscopic video frames. It contains 720 frames that were manually selected and annotated from narrow-band laryngoscopic videos. These videos came from 18 different patients who were subsequently diagnosed with laryngeal spinocellular carcinoma through histopathological testing. The dataset is equally distributed across four quality categories, with 180 frames in each: (i) 'Informative' – clear, usable frames, (ii) 'Blurred' – frames with motion blur or poor focus, (iii) 'Saliva/Specular' – frames obscured by saliva or containing light reflections, and (iv) 'Underexposed' – frames that are too dark.

Ophthalmic microscopy Cataract surgery is a common medical intervention that involves removing the eye's natural lens that has become cloudy (the cataract) and replacing it with an artificial intraocular lens. The surgery is typically performed under local anesthesia on an outpatient basis, takes about 30 minutes to complete, and has a high success rate in restoring vision clarity and improving quality of life for patients affected by cataracts. The **CatRelComp** dataset [127] was created to help train AI systems in

distinguishing between active surgical moments and idle periods during cataract surgery. The dataset contains 22 000 annotated video frames drawn from 22 cataract surgery videos recorded at Klinikum Klagenfurt in Austria (2017-2018), where each video contributed 1000 frames through uniform sampling evenly split between idle and action frames.

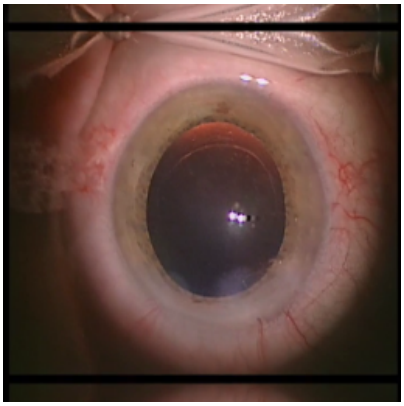


Figure 2.9: Sample image from CatRelComp [127].

Dermatoscopy In the field of medical imaging, dermatoscopy has emerged as a crucial non-invasive diagnostic technique that enhances the visualization of skin lesions through specialized illumination and magnification. Several significant datasets have been developed to advance research and clinical applications in this field. Among these, the **SKLIN2** [110] dataset provides 376 light fields across eight categories of skin lesions, offering a broad spectrum of conditions from melanoma to psoriasis. The **MEDNODE** [128] dataset, though smaller with 170 images, focuses specifically on melanoma and naevus cases, drawing from the University Medical Center Groningen's extensive dermatology archives. The **Derm7pt** [192] dataset stands out for its fine-grained classification approach, while the **PH2** [247] dataset offers 200 high-quality dermoscopic images captured under standardized conditions at Hospital Pedro Hispano in Portugal, complete with detailed medical annotations and dermoscopic criteria assessment. The **ISIC20** [328] dataset represents the most comprehensive collection, featuring images from over 2000 patients across multiple international institutions, including Hospital Clínic de Barcelona and Memorial Sloan Kettering Cancer Center. This dataset, which played a central role in the SIIM-ISIC Melanoma Classification Challenge, provides histopathologically verified diagnoses for malignant cases and expert-confirmed benign cases, making it particularly valuable for research and machine learning applications in dermatological diagnosis.

Non-medical datasets We will also work with some non-medical datasets from general computer vision that shall be mentioned qickly. **MNIST** [219] and its more modern extension **EMNIST** [73] are large collection of handwritten digits respectively letters. The two corresponding tasks are to categorize single handwritten digits respectively letters. **SVHN** [268] is a collection of digits from house numbers recorded for Google Street View that also strives to predict digits from those images. **CIFAR-10** and **CIFAR-100** [209] are two datasets of natural images comprising 10 respectively 100 classes of objects. Those images are like the previous ones of rather small resolution ($28 \times 28 \times 1$ for MNIST and $32 \times 32 \times 3$ for SVHN and CIFAR). Severly larger resolutions and number of classes are within the **Caltech101** [112] and its extension the **Caltech256** [149] datasets of natural objects that have been collected through Google Images. **ImageNet** [93]

marked a milestone in dataset scale, with the standard ImageNet-1k variant comprising 1000 classes derived from the WordNet database [249] and about 1.3 million training samples. **Stanford Dogs** [198] is a curated subset of the larger ImageNet-21k, focussing on categorizing 120 dog species. Finally, **ibean** [214] is a small scale dataset to predict diseases from images of bean plant leaves.

Table 2.1: **Task overview.** The experiments of this thesis cover a wide range of task sizes, number of classes, Imbalance Ratio (IR) (Def. 2.4), and imaging modalities.

task name	ID	# samples	# classes	IR	imaging domain	reference(s)
Nerthus	T01	5525	4	5.40	colonoscopy	[299]
HyperKvasir anatomical-landmarks	T02	4104	6	112.11	gastro & colonoscopy	[38]
HyperKvasir pathological-findings	T03	2642	12	171.33	gastro & colonoscopy	[38]
HyperKvasir quality-of-mucosal-views	T04	1925	3	8.76	colonoscopy	[38]
HyperKvasir therapeutic-interventions	T05	1991	2	1.01	colonoscopy	[38]
LapGyn4 anatomical structures	T06	2728	5	8.42	laparoscopy	[222]
LapGyn4 surgical actions	T07	30 682	8	10.90	laparoscopy	[222]
LapGyn4 instrument count	T08	21 424	4	1.12	laparoscopy	[222]
LapGyn4 anatomical actions	T09	4782	4	2.95	laparoscopy	[222, 390]
Cholec80 grasper presence	T10	89 910	2	1.22	laparoscopy	[390]

Continued on next page

Table 2.1: **Task overview.** The experiments of this thesis cover a wide range of task sizes, number of classes, Imbalance Ratio (IR) (Def. 2.4), and imaging modalities. (Continued)

task name	ID	# samples	# classes	IR	imaging domain	reference(s)
Cholec80 bipolar presence	T11	89 910	2	22.52	laparoscopy	[390]
Cholec80 hook presence	T12	89 910	2	1.32	laparoscopy	[390]
Cholec80 scissors presence	T13	89 910	2	55.16	laparoscopy	[390]
Cholec80 clipper presence	T14	89 910	2	30.11	laparoscopy	[390]
Cholec80 irrigator presence	T15	89 910	2	22.39	laparoscopy	[390]
Cholec80 specimenbag presence	T16	89 910	2	15.74	laparoscopy	[390]
Stanford dogs	T17	20 429	119	1.70	natural images	[93, 198]
SVHN	T18	73 257	10	2.98	natural images	[268]
Caltech101	T19	8677	101	25.81	natural images	[112]
Caltech256	T20	30 607	257	10.34	natural images	[149]
CIFAR10	T21	50 000	10	1.00	natural images	[209]
CIFAR100	T22	50 000	100	1.00	natural images	[209]
SKLIN2	T23	280	8	19.40	dermatoscopy	[110]
derm7pt	T24	616	5	13.69	dermatoscopy	[192]
MNIST	T25	60 000	10	1.24	handwritings	[219]
EMNIST	T26	112 800	47	1.00	handwritings	[73]
NBI-InfFrames	T27	720	4	1.00	laryngoscopy	[256]
Laryngeal cancerous tissue	T28	1320	4	1.00	laryngoscopy	[255]

Continued on next page

Table 2.1: **Task overview.** The experiments of this thesis cover a wide range of task sizes, number of classes, Imbalance Ratio (IR) (Def. 2.4), and imaging modalities. (Continued)

task name	ID	# samples	# classes	IR	imaging domain	reference(s)
CheXpert consolidation	T29	42 880	2	1.90	X-ray	[181]
CheXpert pneumonia	T30	8838	2	2.16	X-ray	[181]
CheXpert atelectasis	T31	34 704	2	25.13	X-ray	[181]
CheXpert pneumothorax	T32	75 789	2	2.90	X-ray	[181]
CheXpert pleural effusion	T33	121 583	2	2.43	X-ray	[181]
CheXpert pleural other	T34	3839	2	11.15	X-ray	[181]
CheXpert fracture	T35	11 552	2	3.60	X-ray	[181]
CheXpert support devices	T36	122 138	2	18.90	X-ray	[181]
CheXpert edema	T37	72 972	2	2.52	X-ray	[181]
CheXpert enlarged cardio-mediastinum	T38	32 436	2	2.00	X-ray	[181]
CheXpert cardiomegaly	T39	38 116	2	2.43	X-ray	[181]
CheXpert lung opacity	T40	112 180	2	16.00	X-ray	[181]
CheXpert lung lesion	T41	10 456	2	7.23	X-ray	[181]
Zhang Chest X-Ray Images	T42	5232	2	2.88	X-ray	[195]

Continued on next page

Table 2.1: **Task overview.** The experiments of this thesis cover a wide range of task sizes, number of classes, Imbalance Ratio (IR) (Def. 2.4), and imaging modalities. (Continued)

task name	ID	# samples	# classes	IR	imaging domain	reference(s)
Shenzhen Hospital CXR Set	T43	662	2	1.03	X-ray	[185]
kaggle COVID X-Ray dataset	T44	3091	2	1.38	X-ray	[164]
MURA wrist	T45	9752	2	1.45	X-ray	[311]
MURA shoulder	T46	8379	2	1.01	X-ray	[311]
MURA humerus	T47	1272	2	1.12	X-ray	[311]
MURA hand	T48	5543	2	2.74	X-ray	[311]
MURA forearm	T49	1825	2	1.76	X-ray	[311]
MURA finger	T50	5106	2	1.59	X-ray	[311]
MURA elbow	T51	4931	2	1.46	X-ray	[311]
ibean	T52	1167	3	1.02	natural images	[214]
CatRelComp	T53	18 000	2	1.00	ophthalmic microscopy	[127]
AIDA-E Barrett's esophagus	T54	262	3	5.73	confocal laser endomicroscopy	[13]
Dataset of breast ultrasound images	T55	780	3	3.29	sonography	[95]
kaggle cataract dataset	T56	601	4	3.00	fundus photography	[55]
kaggle Brain Tumor dataset	T57	3762	2	1.24	MRI	[36]
brain tumor type classification	T58	3064	3	2.01	MRI	[61]
COVID-CT-Dataset	T59	746	2	1.14	CT	[431]
MED-NODE	T60	170	2	1.43	dermatoscopy	[128]

Continued on next page

Table 2.1: **Task overview.** The experiments of this thesis cover a wide range of task sizes, number of classes, Imbalance Ratio (IR) (Def. 2.4), and imaging modalities. (Continued)

task name	ID	# samples	# classes	IR	imaging domain	reference(s)
PH2	T61	200	3	2.00	dermatoscopy	[247]
ISIC20	T62	32 701	2	55.28	dermatoscopy	[77, 328]
DeepDRiD dr level	T63	1200	5	7.50	fundus photography	[232]
DeepDRiD quality	T64	1200	2	1.08	fundus photography	[232]
DeepDRiD clarity	T65	1200	5	14.86	fundus photography	[232]
DeepDRiD field	T66	1200	5	114.43	fundus photography	[232]
DeepDRiD artifact	T67	1200	6	16.19	fundus photography	[232]
APTOS 2019 Blindness Detection	T68	3662	5	9.35	fundus photography	[191]
Kvasir-Capsule anatomy	T69	5718	2	2.74	capsule endoscopy	[358]
Kvasir-Capsule content	T70	38 466	4	76.99	capsule endoscopy	[358]
Kvasir-Capsule pathologies	T71	37 156	5	67.86	capsule endoscopy	[358]

2.3 Basic entities in biomedical image classification

Biomedical image classification aims to map medical images to probability distributions across predetermined categories of biological or medical significance. These categories may represent pathologies, anatomical structures, cellular phenotypes, tissue types, functional states, or other clinically relevant distinctions. In this section, we establish a formal framework and notation for addressing this fundamental task. We begin with the core mechanics – the 'inner loop' of a Lifelong Learning system (see Fig. 2.10) – maintaining a general formulation that accommodates diverse approaches. This foundation allows us to later specify concrete model implementations in Sec. 2.7 before expanding our

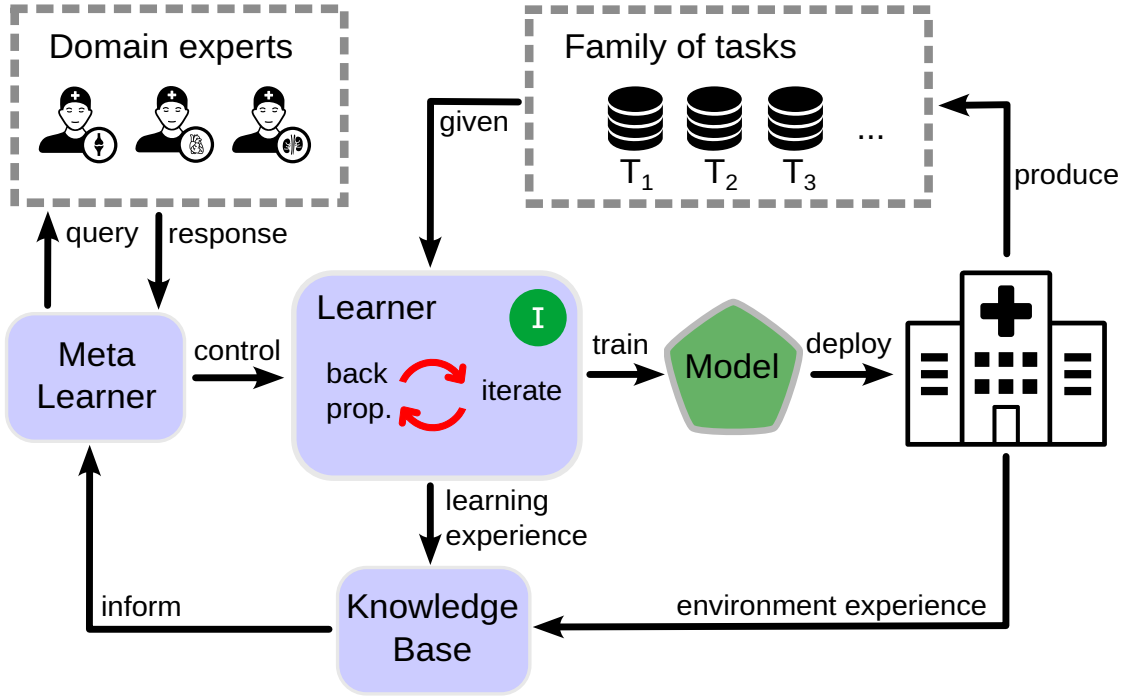


Figure 2.10: Inner learning loop. Anchoring of this section in the overall Lifelong Learning system (see Fig. 1.1). Given a task, the *Learner* trains a model using backpropagation – the 'inner loop' – marked in red.

perspective to explore metacognitive concepts in Sec. 2.8. We first define the essential components of biomedical image classification.

2.3.1 Images and tasks

Definition 2.1. An **image** x is an element of $\mathbb{R}^{h \times w \times c}$, with $h, w, c \in \mathbb{N}$ named **height**, **width** and **number of channels**. The set of all images is referred to as

$$\mathcal{X} := \bigcup_{h,w,c \in \mathbb{N}} \mathbb{R}^{h \times x \times c}.$$

The nature of h, w and c can differ substantially between imaging devices, their settings and the exact procedure. For example, hyperspectral images [234] may have from a handful to several hundred channels that correspond to different wavelengths [71]. In 3D medical imaging such as MRI or CT the channels correspond to the third spatial dimension. X-ray imaging produces 2D images with a collapsed $c = 1$. Whole-slide images from digital pathology usually have high resolutions along h and w reaching several thousand, while the famous ImageNet [93] dataset consists of natural RGB images

with dimensions 224 × 224 × 3.

Definition 2.2. A **task** $\mathcal{T}$ is a finite set of tuples (x, y), comprising each one image $x \in \mathcal{X}$ and a **label**[a] $y \in \mathbb{N}$. The two projections $X_\mathcal{T} := \{x \in \mathcal{X} | \exists y \in \mathbb{N} : (x, y) \in \mathcal{T}\}$ and $Y_\mathcal{T} := \{y \in \mathbb{N} | \exists x \in \mathcal{X} : (x, y) \in \mathcal{T}\}$ are called the images and labels of $\mathcal{T}$ respectively.

[a] As explained in Sec. 1.2, we will only focus on task category of Image-level Classification (ImLC) within this thesis. Some other task types comprise Semantic Segmentation (SemS), Object Detection (ObD), or Instance Segmentation (InS).

The cardinality of $\mathcal{T}$, denoted $|\mathcal{T}|$, is also called the **size** (or equivalently the **number of samples**) of $\mathcal{T}$. The cardinality of $Y_\mathcal{T}$, denoted $|Y_\mathcal{T}|$, is also called the **number of classes** of $\mathcal{T}$. Without loss of generality we can assume that the set $Y_\mathcal{T}$ of all labels in $\mathcal{T}$ is an initial segment of $\mathbb{N}$, i. e., $\{1, 2, 3, ..., C\}$ with $C := |Y_\mathcal{T}|$. If $C = 2$ the task is called **binary**, in case $C > 2$ it is called **multiclass**.[4] It is important to keep in mind, that the samples of a given task only show a small excerpt of the underlying problem. For some of the more theoretical considerations it will be thus helpful to model tasks as realizations of overarching data distributions.

Definition 2.3. Let $\mathcal{T}$ be a task with C classes. Then we associate with $\mathcal{T}$ a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, on which we have $|\mathcal{T}|$ i.i.d. random variable pairs $\{(X_i, Y_i)\}_{i \leq |\mathcal{T}|}$ with $(X_i, Y_i) \in \mathcal{X} \times \{1, ..., C\}$ having the same **joint distribution** $p_\mathcal{T}(X, Y)$. The elements $\{(x_i, y_i)\}_{i \leq |\mathcal{T}|}$ of $\mathcal{T}$ are **realizations** of $\{(X_i, Y_i)\}_{i \leq |\mathcal{T}|}$. The two marginal distributions $p_\mathcal{T}(X)$ and $p_\mathcal{T}(Y)$ are called the **image distribution** and the **class distribution**.

Definition 2.4. Given a task $\mathcal{T}$ and some label $k \in \mathbb{N}$ we call

$$\mathcal{P}_\mathcal{T}(k) := |\{(x, y) \in \mathcal{T} | y = k\}| \cdot |\mathcal{T}|^{-1}$$

the **prevalence** of class k in $\mathcal{T}$. The fraction $\mathrm{IR}_\mathcal{T} := \max_k \mathcal{P}_\mathcal{T}(k) / \min_k \mathcal{P}_\mathcal{T}(k)$ is called the **Imbalance Ratio (IR)** of $\mathcal{T}$.

If $\mathcal{T}$ is clear from context we may drop the index and simply write $\mathcal{P}(k)$ instead of $\mathcal{P}_\mathcal{T}(k)$. In case $\mathrm{IR}_\mathcal{T} = 1$ the task is called **balanced**, otherwise **imbalanced**.

[4] The case $|Y_\mathcal{T}| = 0$ implies $\mathcal{T}$ is the empty set. In case $|Y_\mathcal{T}| = 1$ the task may be called 'trivial'. Both cases are of no further interest in this thesis.

Proposition 2.5. Let $\mathcal{T}$ be a task with C classes, then

$$\sum_{i \leq C} \mathcal{P}(i) = 1.$$

Proof.

$$\begin{aligned}\sum_{i \leq C} \mathcal{P}(i) &= \sum_{i \leq C} |\{(x, y) \in \mathcal{T} | y = k\}| \cdot |\mathcal{T}|^{-1} \\ &= |\mathcal{T}|^{-1} \cdot \sum_{i \leq C} |\{(x, y) \in \mathcal{T} | y = k\}| \\ &= |\mathcal{T}|^{-1} \cdot |\{(x, y) \in \mathcal{T}\}| \\ &= 1\end{aligned}$$

□

Corollary 2.6. Let $\mathcal{T}$ be a balanced task with C classes, then $\mathcal{P}_{\mathcal{T}}(k) = C^{-1}$ for all $k \leq C$.

Proof. Since $\mathcal{T}$ is balanced, $1 = \mathrm{IR}_{\mathcal{T}} = \max_k \mathcal{P}_{\mathcal{T}}(k) / \min_k \mathcal{P}_{\mathcal{T}}(k)$, or equivalently

$$\max_k \mathcal{P}_{\mathcal{T}}(k) = \min_k \mathcal{P}_{\mathcal{T}}(k).$$

Thus, the prevalences for all classes must be equal and by Prop. 2.5, they sum to 1, which concludes the corollary. □

2.3.2 Models and their outputs

Definition 2.7. Let $\mathcal{T}$ be a task with C classes. A **model** φ for $\mathcal{T}$ is an algorithm that computes a mapping $\varphi : \mathcal{X} \mapsto \mathbb{R}^C$. The class of all models for $\mathcal{T}$ is denoted as $\Phi_{\mathcal{T}}$. With $\mathrm{Im}(\varphi) \subseteq \mathbb{R}^C$ we denote the image of φ under $\mathcal{T}$, i. e., $\{\varphi(x_i) | (x_i, y_i) \in \mathcal{T}\}$.

This definition clearly requires some explanation. For one, we require any model φ to 'accept' any image $x \in \mathcal{X}$. From the perspective of the model purpose this is to generalize across imaging specifics of $\mathcal{T}$. More technically this is usually realized with some *preprocessing*, that we interpret as part of the model computations[5]. Clearly, common preprocessing pipelines do not accept all kinds of images and may throw exceptions. Such exceptions may also be caused by models that exceed hardware limitations or

[5] A good argument in favor of this perspective is that during *model inference* usually preprocessing must also be performed. We will introduce preprocessing formally in Def. 2.83.

computations that exceed a given budget. In general, we will treat all those cases as **undefined** (or **invalid**) **output**. How to treat such cases is discussed as part of Chap. 4.

The next unusual choice is $\mathbb{R}^C$ as model output space. The most common perspectives for classification tasks are either to return a single class $y \in Y_{\mathcal{T}}$ (a **categorical model**) or a C-dimensional probability vector from the simplex $\Delta_{C-1} := \{p \in [0,1]^C | \sum_k p_k = 1\}$ (a **probabilistic model**). In practice though many models, and specifically NN, return primarily what is called **logits**[6]: a C-dimensional real-valued vector without restrictions. The standard approach to 'transform' logits to probabilities is to apply the following function [33]:

Definition 2.8. The **softmax** function $\sigma : \mathbb{R}^C \mapsto \Delta_{C-1}$ for $C \in \mathbb{N}, C \geq 1$ is given by:

$$\sigma(v)_i := \frac{e^{v_i}}{\sum_{j=1}^{C} e^{v_j}} \tag{2.1}$$

Making the distinction between 'raw' model logits and the more interpretable softmax-transformed will be important later on (see Sec. 4.1, Sec. 6.1). Note that our definition of a model does not exclude the possibility to output class probabilities directly, neither is the softmax function σ the sole option of transforming a generic output to class probabilities. We will regularly refer to a model φ that computes **class-probabilities** (i. e., $\varphi : \mathcal{X} \mapsto \Delta_{C-1}$), which may be interpreted as a softmax post-processed NN but is not restricted to these specifics.

Afterwards, the probabilities might be processed further to achieve a single class decision:

Definition 2.9. A **decision rule** ρ for $C \in \mathbb{N}, C > 1$ classes, is a mapping $\rho : \Delta_{C-1} \mapsto \{1, ..., C\}$. Two important decision rules are:

(i) the **argmax** operator, which is given by

$$\text{argmax}(p) := \min\{x \in \{1, ..., C\} | \forall k \in \{1, ..., C\} : p_k \leq p_x\}$$

(ii) the **threshold** (also **cutoff**) operator ρ_τ, which for $C = 2$ and $\tau \in [0, 1]$ is given by

$$\rho_\tau(p) := \begin{cases} 1 & \text{, if } p_1 \geq \tau \\ 2 & \text{, else.} \end{cases}$$

[6] Unfortunately 'logit' is an overloaded term. Originally the term was introduced by Berkson [29] in 1944 for the inverse of the 'logistic function' and an abbreviation for "**log**istic un**it**". The purpose of the function was to map probabilities $(0, 1)$ to $(-\infty, +\infty)$ and perform linear regression in this transformed domain before transferring the result back to probabilities. In DL people started calling the layer that feeds into the softmax the *logit layer* and later the *values* that feed into softmax the *logits*.

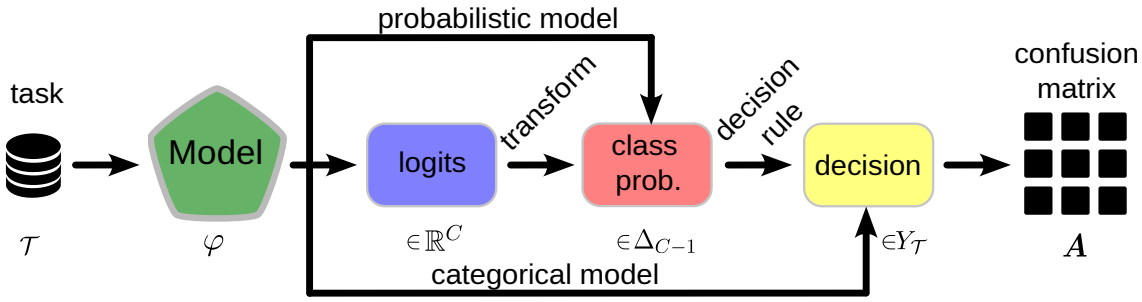


Figure 2.11: Flow of model outputs. We distinguish three kinds of models (see Def. 2.7): *categorical models* already compute decisions on images, *probabilistic models* provide class probabilities for all classes of a task and finally models that produce *logits* (such as most Neural Networks (NNs)). Via *transformation* (e. g., softmax (see Def. 2.8)) and *decision rule* (see Def. 2.9) it is for all those models possible to derive a *confusion matrix* (see Def. 2.11).

The different kinds of model outputs (logits, class-probabilities and categorical decisions) allow for different interpretations and reflect different needs that models may be asked for from individual applications (see Sec. 4.1). On the other hand, intervening in and modifying the post-processing pipeline by adjusting the decision rule or other transformations of model logits offers solutions to shifts in data distribution (see Sec. 6.1).

2.3.3 Performance assessment

Model assessment constitutes a critical step in the development and validation pipeline, requiring robust methodologies to quantify predictive performance and reliability.

Definition 2.10. A **performance measure** $\mu : \Phi_{\mathcal{T}} \mapsto \mathbb{R}$ for task $\mathcal{T}$ assigns any model φ for $\mathcal{T}$ a scalar value. If larger (respectively smaller) values of μ are perceived as better performance we call μ **positively (respectively negatively) oriented**.

Such performance measures are commonly also named as ‘metrics’; though they are not related to any of the properties of a metric space. Two things are important to note here: First, we do not require μ to necessarily depend on a prediction of some model φ, i. e., inference runtime in seconds, energy consumption during training in kWh, and number of learnable parameters are all valid performance measures. Though μ may very much evaluate φ on $\mathcal{T}$, compute predictions $\varphi(x)$ for images x in $X_{\mathcal{T}}$, optionally apply further transformations like the softmax (or even a decision rule), and finally compare the result with the corresponding label y of x.

The second important detail is that such evaluation usually requires images not previously shown to φ to avoid biased evaluation, also known as *data leakage*. Typically, a task is partitioned into multiple subsets, e. g., $\mathcal{T}_{\text{train}}$, $\mathcal{T}_{\text{val}}$, $\mathcal{T}_{\text{test}}$ that are used for different

purposes. We will provide more details about this in Sec. 2.7. For now, it suffices to note that our definition does not break apart as long as any inspected partition of $\mathcal{T}$ contains all classes. Because then any model φ for $\mathcal{T}$ is also a model for any such partition and vice versa, i. e., they all map to $\mathbb{R}^C$ for the same $C \in \mathbb{N}$.[7]

Note that if necessary via a transformation, e. g., the softmax function σ (see Def. 2.8), it is always possible to derive class probabilities from logits. Similarly, by leveraging a decision rule ρ, e. g., the argmax operator (see Def. 2.9) class probabilities can always be turned into categorical decisions (see Fig. 2.11). Hence, metrics that are based on categorical decisions made by φ are the most widely applicable class of performance measures. All of them are calculated based on the following entity.

Definition 2.11. For a task $\mathcal{T}$ with C classes and a model φ for $\mathcal{T}$ that computes categorical decisions, i. e., $\varphi : \mathcal{X} \mapsto \{1, ..., C\}$, we define the $C \times C$ **confusion matrix**

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1C} \\ a_{21} & a_{22} & \cdots & a_{2C} \\ \vdots & \vdots & \ddots & \vdots \\ a_{C1} & a_{C2} & \cdots & a_{CC} \end{bmatrix}$$

via $a_{ij} := |\{(x, y) \in \mathcal{T} | y = i, \varphi(x) = j\}|$.

The entry a_{ij} may be interpreted as the number of images in $\mathcal{T}$ that belong to class i and have been classified as class j by the model φ. Note that $\boldsymbol{A}$ is an integer matrix. Under mild assumptions we can conclude some simple properties of confusion matrices.

Proposition 2.12. Let $\boldsymbol{A}$ be the confusion matrix for task $\mathcal{T}$ with C classes and the categorical model φ for $\mathcal{T}$. If we assume that φ succeeds[a] in computation for all images of task $\mathcal{T}$ then the following hold:

(i) For any $k \leq C$,

$$\sum_{j \leq C} a_{kj} = \mathcal{P}(k) \cdot |\mathcal{T}|.$$

(ii) Moreover,

$$\sum_{i \leq C} \sum_{j \leq C} a_{ij} = |\mathcal{T}|.$$

[7] This assumption can be controlled easily in a research setup, but the deployment of an algorithm to an 'unknown' environment can lead to violations. Usually 'disappearing' classes do less harm than 'additional' ones (see Sec. 6.1). There are approaches to cope with these, e. g., *selective classification* may refuse to predict samples [125]. We refer to the literature for the details.

[a] Precisely it needs to succeed under the constraints given by the performance measure μ. Such constraints may be, for example, hardware limitations and/or a maximum computation time per image.

Proof. To prove (i) let $k \leq C$, then

$$\begin{aligned}\sum_{j\leq C} a_{kj} &\overset{2.11}{=} \sum_{j\leq C} |\{(x,y) \in \mathcal{T} | y = k, \varphi(x) = j\}| \\ &= |\{(x,y) \in \mathcal{T} | y = k, \varphi(x) \leq C\}| \cdot 1 \\ &= |\{(x,y) \in \mathcal{T} | y = k\}| \cdot |\mathcal{T}|^{-1} \cdot |\mathcal{T}| \\ &\overset{2.4}{=} \mathcal{P}(k) \cdot |\mathcal{T}|.\end{aligned}$$

From this (ii) follows straightforward

$$\sum_{i\leq C}\sum_{j\leq C} a_{ij} \overset{(i)}{=} \sum_{i\leq C} \mathcal{P}(i) \cdot |\mathcal{T}| \overset{2.5}{=} |\mathcal{T}|. \qquad \square$$

Definition 2.13. Let $\boldsymbol{A}$ be a confusion matrix for a task with C classes. For a class $k \leq C$ we define

(i) $\mathbf{TP}_k := a_{kk}$ the **True Positive (TP)** for class k,

(ii) $\mathbf{TN}_k := \sum_{i,j\leq C, i\neq k, j\neq k} a_{ij}$ the **True Negative (TN)** for class k,

(iii) $\mathbf{FP}_k := \sum_{i\leq C, i\neq k} a_{ik}$ the **False Positive (FP)** for class k,

(iv) $\mathbf{FN}_k := \sum_{j\leq C, j\neq k} a_{kj}$ the **False Negative (FN)** for class k.

Proposition 2.14. Let $\boldsymbol{A}$ be a confusion matrix for a task $\mathcal{T}$ with C classes and $k \leq C$, then under the same assumption as Proposition 2.12, it holds that

$$\mathbf{TP}_k + \mathbf{TN}_k + \mathbf{FP}_k + \mathbf{FN}_k = |\mathcal{T}|.$$

Proof. Follows immediately from Proposition 2.12 (ii). $\square$

Definition 2.15. Given a confusion matrix $\boldsymbol{A}$ for task $\mathcal{T}$ with C classes and the categorical model φ for $\mathcal{T}$ and $k \leq C$, we define the 2×2 **one-versus-the-rest confusion matrix**

$$\boldsymbol{A}^{(k)} := \begin{bmatrix} \mathbf{TP}_k & \mathbf{FN}_k \\ \mathbf{FP}_k & \mathbf{TN}_k \end{bmatrix}$$

This definition requires some elaboration. The special case of binary classification may arguably be perceived as canonical within the space of classification problems. Therefore and for its simplicity historically it has been studied the most, which lead to the development and dissemination of performance measures that are tailored to this specific use case. The natural extension of such performance measures μ to the multiclass case is through the aggregation[8] of all C evaluations of μ on the 'binarized' problem $\mathcal{T}^{(k)}$ and model $\varphi^{(k)}$ for $k \leq C$. The core idea is to merge all classes $Y_{\mathcal{T}} \setminus \{k\}$. Formally, we use $h_k : \mathbb{N} \mapsto \{1, 2\}$ with

$$h_k(n) := \begin{cases} 1 & \text{, if } n = k \\ 2 & \text{, else} \end{cases}$$

to define $\mathcal{T}^{(k)} := \{(x, h_k(y)) : (x, y) \in \mathcal{T}\}$ and $\varphi^{(k)}(x) := h_k(\varphi(x))$. Now the (regular) confusion matrix $\boldsymbol{A}$ for task $\mathcal{T}^{(k)}$ and model $\varphi^{(k)}$ equals exactly $\boldsymbol{A}^{(k)}$. If $C = 2$ from the beginning, then $\boldsymbol{A} = \boldsymbol{A}^{(1)}$, but $\boldsymbol{A}^{(2)}$ represents the 'flipped' matrix

$$\boldsymbol{A}^{(2)} = \begin{bmatrix} a_{22} & a_{21} \\ a_{12} & a_{11} \end{bmatrix},$$

as a result from changing the 'order' of classes.

For the rest of this chapter we will assume $\mathcal{T}$ to be a task with C classes and $\varphi : \mathcal{X} \mapsto \mathbb{R}^C$ a model for $\mathcal{T}$. Recall from the discussion following Def. 2.7, that a model output may be interpreted as either logits, class-probabilities or categorical decisions. In the following sections we present a variety of performance measures $\mu : \Phi_{\mathcal{T}} \mapsto \mathbb{R}$, adjacent with insights on their properties and relations. Unless stated otherwise we will assume that φ succeeds on all images of $\mathcal{T}$ (see Prop. 2.12). Special cases of 'failed' computations will be treated separately in Chap. 4.

2.4 Counting metrics

Counting metrics rely on categorical decisions made by φ and by the consideration prepending Def. 2.11 we can generally assume to be given a confusion matrix $\boldsymbol{A}$ by any model φ.

[8] In case of the arithmetic mean, the resulting performance measures are often called *macro average*, since averaging happens on the categorical level in contrast to *mirco average*, an approach that puts equal weight to all samples.

2.4.1 The Sensitivity perspective

Definition 2.16. Let $k \leq C$ be a class, then

(i) the **True Positive Rate (TPR)** (also **Sensitivity**, **Recall**, **Hit rate**) of class k is defined as

$$\mathbf{TPR}_k := \frac{\mathbf{TP}_k}{\mathbf{TP}_k + \mathbf{FN}_k},$$

(ii) the **False Negative Rate (FNR)** (also **Miss rate**, **Type II error**) of class k is defined as

$$\mathbf{FNR}_k := \frac{\mathbf{FN}_k}{\mathbf{TP}_k + \mathbf{FN}_k},$$

(iii) the **False Positive Rate (FPR)** (also **Type I error**) of class k is defined as

$$\mathbf{FPR}_k := \frac{\mathbf{FP}_k}{\mathbf{FP}_k + \mathbf{TN}_k},$$

(iv) the **True Negative Rate (TNR)** (also **Specificity**, **Selectivity**) of class k is defined as

$$\mathbf{TNR}_k := \frac{\mathbf{TN}_k}{\mathbf{FP}_k + \mathbf{TN}_k}.$$

In terms of popularity these metrics are used frequently in ML literature, with different domains establishing preferences for different names. We can conclude immediately that the value range of them is $[0, 1]$, and while TPR_k and TNR_k are positively oriented (see Def. 2.10), their 'counterparts' FPR_k and FNR_k are negatively oriented. Based on the considerations following Definition 2.15 we allow the notations $\mathbf{TPR} := \mathbf{TPR}_1$, $\mathbf{FNR} := \mathbf{FNR}_1$, $\mathbf{FPR} := \mathbf{FPR}_1$, $\mathbf{TNR} := \mathbf{TNR}_1$ as long as $C = 2$. For later reference, we formalize the following relationships:

Remark 2.17. Let $k \leq C$ be a class, then

(i)
$$\mathbf{TPR}_k = 1 - \mathbf{FNR}_k$$

(ii)
$$\mathbf{FPR}_k = 1 - \mathbf{TNR}_k$$

Definition 2.18. The **Accuracy (AC)** of a categorical model is defined as

$$\mathbf{AC} := |\mathcal{T}|^{-1} \cdot \sum_{i \leq C} a_{ii}.$$

Conversely, the **Error rate** is defined as

$$\mathbf{ER} := |\mathcal{T}|^{-1} \cdot \sum_{i,j \leq C, i \neq j} a_{ij}.$$

Remark 2.19. Obviously $\mathbf{AC} = 1 - \mathbf{ER}$.

AC is likely the most widespread classification metric used in biomedical imaging (see Maier-Hein et al. [240][9]) and beyond (e. g., as primary metric for famous computer vision datasets like ImageNet [93] and MNIST [219]). The reason for this is presumably the simplicity of the definition, that coveys the message 'How likely is the model correct?'. Once more the value ranges of AC and Error Rate (ER) are $[0, 1]$ and while AC is positively oriented, the ER is negatively oriented. We provide some more relations with previously introduced performance measures.

Proposition 2.20. (i)

$$\mathbf{AC} = |\mathcal{T}|^{-1} \cdot \sum_{i \leq C} \mathbf{TP}_i,$$

(ii)

$$\mathbf{AC} = \sum_{i \leq C} \mathcal{P}(i) \cdot \mathbf{TPR}_i.$$

Proof. (i) follows immediately from the definition of TP_k (see Def. 2.13), while (ii) can be shown via

$$\begin{aligned}
\sum_{i \leq C} \mathcal{P}(i) \cdot \mathbf{TPR}_i &\overset{2.12}{=} \sum_{i \leq C} |\mathcal{T}|^{-1} \sum_{j \leq C} a_{ij} \cdot \frac{\mathbf{TP}_i}{\mathbf{TP}_i + \mathbf{FN}_i} \\
&= |\mathcal{T}|^{-1} \cdot \sum_{i \leq C} \mathbf{TP}_i \cdot \frac{\sum_{j \leq C} a_{ij}}{a_{ii} + \sum_{j \leq C, j \neq i} a_{ij}} \\
&= |\mathcal{T}|^{-1} \cdot \sum_{i \leq C} \mathbf{TP}_i \\
&\overset{(i)}{=} \mathbf{AC}
\end{aligned}$$

□

[9] See also the results of our literature search conducted in Sec. 6.2.1.

Prop. 2.20 gives an alternative interpretation of AC as prevalence weighted sum of class-wise Sensitivities. If the weights of the TPRs are chosen to be equal instead, a derivative version of AC emerges.

Definition 2.21. The **Balanced Accuracy (BA)** [45] of a categorical model is defined as

$$\mathbf{BA} := C^{-1} \cdot \sum_{i \leq C} \mathbf{TPR}_i.$$

It follows immediately that BA is positively oriented and takes values in the range $[0, 1]$. BA and AC coincide for balanced tasks.

Proposition 2.22. For balanced $\mathcal{T}$ BA and AC are equal.

Proof.

$$\mathbf{AC} \overset{2.20}{=} \sum_{i \leq C} \mathcal{P}(i) \cdot \mathbf{TPR}_i \overset{2.6}{=} \sum_{i \leq C} C^{-1} \cdot \mathbf{TPR}_i \overset{2.21}{=} \mathbf{BA} \qquad \square$$

After Prop. 2.20 and Def. 2.21 a natural question would be if further 'weighted' Sensitivities could be of particular interest. Indeed, with the following definition we introduce a metric of such kind that will be of particular importance in Chap. 6.

Definition 2.23. Let $\{c_{ij}\}_{i,j \leq C}$ be a $C \times C$ real-valued matrix, called the **confusion costs** (also **confusion weights**). Then the **Expected Cost (EC)** [113] is defined as

$$\mathbf{EC} := |\mathcal{T}|^{-1} \cdot \sum_{i,j \leq C} c_{ij} a_{ij}.$$

The special case of $c_{ii} = 0 \; \forall i \leq C$ and $c_{ij} = 1 \; \forall i, j \leq C, i \neq$ is called **0-1-costs**.

The values c_{ij}, referred to as *costs* for confusions, can be leveraged to penalize certain class confusions more severe than others. A simple example for this are classes that have an ordinal structure, e. g., some severity classes for disease progress. In such a scenario penalizing confusions that are further off the reference class may be justified. Without restrictions on c_{ij} the value range of EC covers whole $\mathbb{R}$, but usually $c_{ij} \geq 0$ and hence EC falls within $[0, \infty)$. Expected Cost is negatively oriented.

Proposition 2.24. If $\{c_{ij}\}_{i,j \leq C}$ are the 0-1-costs, then $\mathbf{EC} = \mathbf{ER}$.

Proof.

$$\begin{aligned}\mathbf{EC} &= |\mathcal{T}|^{-1} \cdot \sum_{i,j\leq C} c_{ij}a_{ij} \\ &= |\mathcal{T}|^{-1} \cdot \sum_{i,j\leq C, i\neq j} a_{ij} \\ &= \mathbf{ER}\end{aligned}$$

□

Moreover, we can express EC in terms of prevalences and (a generalized form of) Sensitivities[10](since we cannot treat false predictions equally).

Proposition 2.25. For $i, j \leq C$ let $R_{ij} := a_{ij} / \sum_{k\leq C} a_{ik}$ be the fraction of all samples with reference class i that have been predicted as j. Then

$$\mathbf{EC} = \sum_{i\leq C} \mathcal{P}(i) \sum_{j\leq C} c_{ij} R_{ij}.$$

Proof.

$$\begin{aligned}\sum_{i\leq C} \mathcal{P}(i) \sum_{j\leq C} c_{ij}R_{ij} &= \sum_{i,j\leq C} c_{ij}\mathcal{P}(i) \frac{a_{ij}}{\sum_{k\leq C} a_{ik}} \\ &\overset{2.12}{=} \sum_{i,j\leq C} c_{ij}\mathcal{P}(i) \frac{a_{ij}}{\mathcal{P}(i)\cdot|\mathcal{T}|} \\ &= |\mathcal{T}|^{-1} \cdot \sum_{i,j\leq C} c_{ij}a_{ij} \\ &= \mathbf{EC}\end{aligned}$$

□

As the value of EC depends on the chosen costs and prevalences, it can be difficult to interpret. As a solution to this limitation, it might be usefule to use the normalized version of EC.

Definition 2.26. Let $\{c_{ij}\}_{i,j\leq C}$ be some confusion costs, then the **naive classifier** is a model that (independently of the input image x) always predicts class $\operatorname{argmin}_j \sum_{i\leq C} c_{ij}\mathcal{P}(i)$.

The naive classifier may be interpreted as the solution to the question 'Which class should be predicted to minimize the costs, while no information on the sample may be used?' With only ever predicting a single class it tries to minimize the EC. Entries in the confusion matrix off the j-th column are all zero. This simplest of all models is now used to normalize the EC.

[10]The fractions R_{ij} introduced by Prop. 2.25 generalize Sensitivity in the sense, that $\mathbf{TPR}_i = R_{ii}$.

Definition 2.27. Let $\{c_{ij}\}_{i,j\leq C}$ be some confusion costs, then the **Normalized Expected Cost (NEC)** [113] is given by

$$\mathbf{NEC} := \frac{\mathbf{EC}}{\min_j \sum_{i\leq C} c_{ij}\mathcal{P}(i)}.$$

While NEC is still negatively oriented and may take any real value, the interesting property about it is that values above one indicate an EC value that is worse than the naive classifier – a highly undesirable situation.

Another measure that allows the weighting of individual confusions was given by Cohen. For ease of notation and because it will be useful later we introduce a notation of predictive bias.

Definition 2.28. Let $\boldsymbol{A}$ be a confusion matrix and $k \leq C$, then the frequency of predicting k is defined as $\mathcal{B}_{\mathcal{T}}(k) := |\{(x,y) \in \mathcal{T} | \varphi(x) = k\}| \cdot |\mathcal{T}|^{-1}$ and called the **model bias**.

Similar to the case of prevalences, we will drop the task index for easier readability as long as the task is clear from context.

Proposition 2.29. For any $k \leq C$,

$$\sum_{j\leq C} a_{jk} = \mathcal{B}(k) \cdot |\mathcal{T}|.$$

Also

$$\sum_{\leq C} \mathcal{B}(k) = 1.$$

Proof. The proof closely follows the proof of Prop. 2.12. Let $k \leq C$, then

$$\begin{aligned}\sum_{j\leq C} a_{jk} &\overset{2.11}{=} \sum_{j\leq C} |\{(x,y) \in \mathcal{T} | y = j, \varphi(x) = k\}| \\ &= |\{(x,y) \in \mathcal{T} | y \leq C, \varphi(x) = k\}| \cdot 1 \\ &= |\{(x,y) \in \mathcal{T} | \varphi(x) = k\}| \cdot |\mathcal{T}|^{-1} \cdot |\mathcal{T}| \\ &\overset{2.28}{=} \mathcal{B}(k) \cdot |\mathcal{T}|.\end{aligned}$$

The second equation follows as

$$\sum_{k\leq C} \mathcal{B}(k) \overset{\text{(i)}}{=} \sum_{k\leq C} |\mathcal{T}|^{-1} \cdot \sum_{j\leq C} a_{jk} = |\mathcal{T}|^{-1} \sum_{k,j\leq C} a_{jk} \overset{2.12}{=} |\mathcal{T}|^{-1} \cdot |\mathcal{T}| = 1 \qquad \square$$

With the model bias as tool, we can now define Cohen's alternative performance measure for unequal confusion costs.

Definition 2.30. Let $\{c_{ij}\}_{i,j\leq C}$ be some confusion costs. Then the **Weighted Cohen's Kappa (WCK)** [76] is defined as

$$\mathbf{WCK} := 1 - \frac{|\mathcal{T}|^{-1} \cdot \sum_{i,j\leq C} c_{ij} a_{ij}}{\sum_{i,j\leq C} c_{ij} \mathcal{P}(i)\mathcal{B}(j)}.$$

The case of 0-1-costs is simply called **Cohen's Kappa (CK)** [75].[a]

[a] Noteworthy the special case of a binary task and 0-1-costs was already introduced by Myrick Haskell Doolittle in 1888 and became known as the 'Heidke skill score' in Meteorology [163].

The core idea of WCK is closely related to the NEC – in fact the EC can be identified in the numerator of the WCK formula. The main difference is the assumed baseline classifier in the denominator. While the naive classifier in NEC is based solely on the prevalences and costs, the WCK perspective takes additionally the model bias into consideration. One of the reasons for this is that it was designed as an 'inter-rater agreement' – not as a performance measure for discriminative models. The original perspective expects independent categorical decisions from two raters on the same data – a setting that is inherently symmetric (as long as we assume symmetric costs). Hence, what is now deemed prevalences would be the bias of one rater, while the model bias would be the bias of the second rater. The denominator now measures the (weighted) expected agreement by chance – a sensible baseline 'classifier' given the original perspective. But the discriminative and non-symmetric nature of our setting reveals the limitations of WCK. Although the theoretical value range of WCK has a lower bound of minus one and for 0-1-costs an upper bound of one, the values remain hard to interpret [92]. One of the more common WCK use cases have been ordinal categories, which explicitly requires some sort of confusion costs. The assignment of quadratically growing weights has though shown to produce 'paradoxical results' [414]. WCK is positively oriented. We prove the coincidence of WCK with other metrics under certain conditions.

Proposition 2.31. Let $\mathcal{T}$ be balanced and $\{c_{ij}\}_{i,j\leq C}$ the 0-1-costs, then

$$\mathbf{WCK} = \frac{C \cdot \mathbf{BA} - 1}{C - 1}.$$

Proof.

$$
\begin{aligned}
\mathbf{WCK} &\overset{2.30}{=} 1 - \frac{|\mathcal{T}|^{-1} \cdot \sum_{i,j \leq C} c_{ij} a_{ij}}{\sum_{i,j \leq C} c_{ij} \mathcal{P}(i) \mathcal{B}(j)} \\
&\overset{2.24}{=} 1 - \frac{\mathbf{ER}}{\sum_{i,j \leq C} c_{ij} \mathcal{P}(i) \mathcal{B}(j)} \\
&\overset{2.19}{=} 1 - \frac{1 - \mathbf{AC}}{\sum_{i,j \leq C} c_{ij} C^{-1} \mathcal{B}(j)} \\
&\overset{2.22}{=} 1 - \frac{1 - \mathbf{BA}}{C^{-1} \sum_{i,j \leq C, i \neq j} \mathcal{B}(j)} \\
&= 1 - \frac{1 - \mathbf{BA}}{C^{-1} \cdot (C-1) \cdot \sum_{j \leq C} \mathcal{B}(j)} \\
&\overset{2.29}{=} 1 - \frac{1 - \mathbf{BA}}{C^{-1} \cdot (C-1)} \\
&= 1 - \frac{C}{C-1} \cdot (1 - \mathbf{BA}) \\
&= \frac{C - 1 - C \cdot (1 - \mathbf{BA})}{C-1} \\
&= \frac{C \cdot \mathbf{BA} - 1}{C-1}
\end{aligned}
$$

□

We turn our attention to some metrics that are popular as a criterion for the performance of diagnostic tests in evidence-based medicine, but rather unused as performance measures of ML models.

Definition 2.32. For $k \leq C$ we define the **Positive Likelihood Ratio (LR+)** [376] as

$$\mathbf{LR}+_k := \frac{\mathbf{TPR}_k}{1 - \mathbf{TNR}_k}.$$

Similarly, the **Negative Likelihood Ratio (LR-)** is defined as

$$\mathbf{LR}-_k := \frac{1 - \mathbf{TPR}_k}{\mathbf{TNR}_k}.$$

The value ranges of both likelihood ratios can be derived from the value ranges of Sensitivity and Specificity, which are $[0, 1]$. In both cases the division allows for arbitrary large values within $[0, \infty)$. While LR+$_k$ is positively oriented, LR-$_k$ is negatively oriented. The positive (respectively negative) likelihood ratio expresses how many times more likely the positive (respectively negative) prediction of class k is for images of class k (TPR) versus than for those of the other classes combined (FPR). A value of 1 corresponds to a non-beneficial model. For diagnostic applications a value LR+ > 10 (respectively

LR$-$ < 0.1) is considered 'good' [356].

We will now define the last Sensitivity-focused counting metric of this thesis.

Definition 2.33. The **Youden's Index (J)** [293, 434] (also **(Bookmaker) Informdness** [303]) of a categorical model and class $k \leq C$ is defined as

$$\mathbf{J}_k = \mathbf{TPR}_k + \mathbf{TNR}_k - 1.$$

Informdness is positively oriented and its value range is $[-1, 1]$, as both TPR and TNR have value ranges within $[0, 1]$.[11] For the binary case we can link $\mathbf{J} := \mathbf{J}_1$ with another metric.[12]

Proposition 2.34. Let $C = 2$, then

$$\mathbf{BA} = \frac{\mathbf{J} + 1}{2}.$$

Equivalently $\mathbf{J} = 2\mathbf{BA} - 1$.

Proof.

$$\mathbf{BA} \overset{2.21}{=} C^{-1} \cdot \sum_{i \leq C} \mathbf{TPR}_i = \frac{\mathbf{TPR}_1 + \mathbf{TPR}_2}{2} \overset{2.15}{=} \frac{\mathbf{TPR}_1 + \mathbf{TNR}_1}{2} \overset{2.33}{=} \frac{\mathbf{J} + 1}{2} \qquad \square$$

2.4.2 The predictive value perspective

After these variations on metrics from the Sensitivity perspective, we will next present the perspective of 'predictive values'.

Definition 2.35. Let $k \leq C$ be a class, then

(i) the **Positive Predictive Value (PPV)** (also **Precision**) of class k is defined as

$$\mathbf{PPV}_k := \frac{\mathbf{TP}_k}{\mathbf{TP}_k + \mathbf{FP}_k},$$

(ii) the **False Omission Rate (FOR)** of class k is defined as

$$\mathbf{FOR}_k := \frac{\mathbf{FN}_k}{\mathbf{TN}_k + \mathbf{FN}_k},$$

[11] Be aware that for a binary task, or a single class, one may invert the value of J by inverting the (ovr) predictions. Hence a model with $\mathbf{J} = -1$ may be turned into an optimal classifier.

[12] There is also a multiclass variant for Bookmaker Informdness as defined by Powers [304], which we will not introduce in this thesis.

(iii) the **False Discovery Rate (FDR)** of class k is defined as

$$\mathbf{FDR}_k := \frac{\mathbf{FP}_k}{\mathbf{TP}_k + \mathbf{FP}_k},$$

(iv) the **Negative Predictive Value (NPV)** of class k is defined as

$$\mathbf{NPV}_k := \frac{\mathbf{TN}_k}{\mathbf{TN}_k + \mathbf{FN}_k}.$$

With respect to popularity these metrics are – except for PPV – used less frequently in ML literature. The value ranges are all $[0, 1]$, and while PPV_k and NPV_k are positively oriented (see Def. 2.10), once more their 'counterparts' FOR_k and FDR_k are negatively oriented. Similar to the notation allowed following Def. 2.16 we introduce the notations $\mathbf{PPV} := \mathbf{PPV}_1$, $\mathbf{FOR} := \mathbf{FOR}_1$, $\mathbf{FDR} := \mathbf{FDR}_1$, $\mathbf{NPV} := \mathbf{NPV}_1$ as long as $C = 2$. For reference, we formalize the following relationships:

Remark 2.36. Let $k \leq C$ be a class, then

(i)
$$\mathbf{PPV}_k = 1 - \mathbf{FDR}_k$$

(ii)
$$\mathbf{FOR}_k = 1 - \mathbf{NPV}_k$$

Instead of 'row-wise' relativization as done in TPR, the predictive values relativize the confusion matrix entries 'column-wise'. This perspective may be perceived as a 'dual' formulation of the performance assessment. Several analogies to previously shown identities exists for this perspective.

Definition 2.37. Let $\boldsymbol{A}$ be a confusion matrix and $k \leq C$, then the transposed matrix $\boldsymbol{A}^T$, with elements $a_{ij}^T := a_{ji}$ will be called the **dual confusion matrix**.

For the ease of readability we will denote $\mathbf{E}^T$ for any entity that is computed on $\boldsymbol{A}^T$ with the same formula as $\mathbf{E}$ is computed on $\boldsymbol{A}$.

Proposition 2.38. For any $k \leq C$, the prevalence $\mathcal{P}(k)$ (bias $\mathcal{B}(k)$) of class k in $\boldsymbol{A}$ equals the bias $\mathcal{B}^T(k)$ (prevalence $\mathcal{P}^T(k)$) of class k in $\boldsymbol{A}^T$, i. e., they are switched under transposition.

Proof. This follows straightforward

$$\mathcal{P}(k) \overset{2.12}{=} |\mathcal{T}|^{-1} \sum_{j \leq C} a_{kj} \overset{2.37}{=} |\mathcal{T}|^{-1} \sum_{j \leq C} a_{jk}^T \overset{2.29}{=} \mathcal{B}^T(k).$$

$\mathcal{B}(k) = \mathcal{P}^T(k)$ can be shown the same way. □

It is worth to spare some words on these results. Apparently *prevalences* $\mathcal{P}$ and *model bias* $\mathcal{B}$ behave as dual entities, but interestingly while *prevalences* $\mathcal{P}$ are an entity that is solely defined by the task $\mathcal{T}$ (see Def. 2.4) the *model bias* $\mathcal{B}$ is dependent on the model φ (see Def. 2.28) *and* the task $\mathcal{T}$. This insight will be important in Chap. 4 as well as Chap. 6. The next proposition gives some more dual relations.

Proposition 2.39. Let $\boldsymbol{A}$ be a confusion matrix and $k \leq C$, then

(i) The true positives $\mathbf{TP}_k$ (resp. true negatives $\mathbf{TN}_k$) of $\boldsymbol{A}$ are equal to the true positives $\mathbf{TP}_k^T$ (resp. true negatives $\mathbf{TN}_k^T$) of $\boldsymbol{A}^T$, i. e., they are preserved under transposition.

(ii) The false positives $\mathbf{FP}_k$ (resp. false negatives $\mathbf{FN}_k$) of $\boldsymbol{A}$ are equal to the false negatives $\mathbf{FN}_k^T$ (resp. false positives $\mathbf{FP}_k^T$) of $\boldsymbol{A}^T$, i. e., they are switched under transposition.

(iii) The Sensitivity $\mathbf{TPR}_k$ (resp. Specificity $\mathbf{TNR}_k$) of $\boldsymbol{A}$ are equal to the Precision $\mathbf{PPV}_k^T$ (resp. negative predictive value $\mathbf{NPV}_k^T$) of $\boldsymbol{A}^T$.

(iv) The false negative rate $\mathbf{FNR}_k$ (resp. false positive rate $\mathbf{FPR}_k$) of $\boldsymbol{A}$ are equal to the false discovery rate $\mathbf{FDR}_k^T$ (resp. false omission rate $\mathbf{FOR}_k^T$) of $\boldsymbol{A}^T$.

Proof. (i) and (ii) follow straight from Def. 2.13 and Def. 2.37. (iii) and (iv) can be shown straightforward using these results, as exemplary demonstrated via

$$\mathbf{TPR}_k \overset{2.16}{=} \mathbf{TP}_k/(\mathbf{TP}_k + \mathbf{FN}_k) \overset{\text{(i),(ii)}}{=} \mathbf{TP}_k^T/(\mathbf{TP}_k^T + \mathbf{FP}_k^T) \overset{2.35}{=} \mathbf{PPV}_k^T.$$ □

As the trace of a matrix is invariant under transposition, this proposition allows us to formulate an alternative characterization of AC.

Corollary 2.40.

$$\mathbf{AC} = \sum_{i \leq C} \mathcal{B}(i) \cdot \mathbf{PPV}_i.$$

Proof.

$$\begin{aligned}
\mathbf{AC} &\overset{2.18}{=} |\mathcal{T}|^{-1} \cdot \sum_{i \leq C} a_{ii} \\
&\overset{2.37}{=} |\mathcal{T}|^{-1} \cdot \sum_{i \leq C} a_{ii}^T \\
&\overset{2.18}{=} \mathbf{AC}^T \\
&\overset{2.20}{=} \sum_{i \leq C} \mathcal{P}^T(i) \cdot \mathbf{TPR}_i^T \\
&\overset{2.38,\, 2.39}{=} \sum_{i \leq C} \mathcal{B}(i) \cdot \mathbf{PPV}_i
\end{aligned}$$

□

In contrast to AC many other metrics are not invariant under their dual computation. One example is the following.

Definition 2.41. The **Markedness (MK)** [303] of a categorical model and class $k \leq C$ is defined as

$$\mathbf{MK}_k = \mathbf{PPV}_k + \mathbf{NPV}_k - 1.$$

For $C = 2$ we let $\mathbf{MK} := \mathbf{MK}_1$.

This definition might remind of Infordmness (Def. 2.33) and indeed, the relationship between the two is exactly the dual perspective.

Proposition 2.42. Let $k \leq C$, then $\mathbf{MK}_k^T = \mathbf{J}_k$ (and $\mathbf{J}_k^T = \mathbf{MK}_k$).

Proof. Follows immediately from Prop. 2.39. □

We are not aware of widespread usage of 'duals' from BA[13], EC, LR+ or LR- in the literature.

2.4.3 Combined perspectives

We will proceed with some metrics that combine the Sensitivity and Predictive value perspective (as WCK already does). For this we need to balance the two perspectives, which can be done via the following mean.

[13] Note that *Average Precision* as defined in Def. 2.56, is not the dual of *Balanced Accuracy*, i. e., it is not the average of class-wise *Precision*.

Definition 2.43. Let $n \in \mathbb{N}$ and $x_1, .., x_n$ be real numbers. The **weighted harmonic mean** of $x_1, .., x_n$ with positive real valued weights $w_1, .., w_n$ is defined as

$$H(x_1, .., x_n | w_1, .., w_n) := \sum_{i \leq n} w_i \cdot (\sum_{i \leq n} w_i x_i^{-1})^{-1}.$$

This immediately allows to define a very common metric in a slightly generalized form.

Definition 2.44. Let $k \leq C$ and $\beta > 0$ be a real, then the **F-beta** [396] score for class k is given by

$$\mathbf{F}_k(\beta) := H(\mathbf{TPR}_k, \mathbf{PPV}_k | \beta^2, 1).$$

On a side note for interested readers the naming of the 'F-measure' seems to be caused by an accident and has no deeper meaning [336]. Further the squaring of β can also be explained in the context as 'weighting' TPR β times as important as PPV, but we leave the derivation to be looked up in the literature (see Sasaki [336]). Much more common is the following formulation of the F-beta measure.

Proposition 2.45. Let $k \leq C$ and $\beta > 0$ be a real, then we can express the F-beta score as

$$\mathbf{F}_k(\beta) = \frac{(1 + \beta^2) \cdot \mathbf{TP}_k}{(1 + \beta^2) \cdot \mathbf{TP}_k + \beta^2 \cdot \mathbf{FN}_k + \mathbf{FP}_k}.$$

Note we can immediately conclude that the F-beta score has a value range of $[0, 1]$ as the numerator can not grow larger than the denominator. It also becomes obvious that F-beta is positively oriented.

Proof. Assume k and β as necessary by the proposition. Then

$$\begin{aligned}
\mathbf{F}_k(\beta) &\overset{2.44}{=} H(\mathbf{TPR}_k, \mathbf{PPV}_k|\beta^2, 1) \\
&\overset{2.43}{=} (1+\beta^2) \cdot \frac{1}{\beta^2 \cdot \mathbf{TPR}_k^{-1} + 1 \cdot \mathbf{PPV}_k^{-1}} \\
&= \frac{1+\beta^2}{\frac{\beta^2 \cdot \mathbf{PPV}_k + \mathbf{TPR}_k}{\mathbf{PPV}_k \cdot \mathbf{TPR}_k}} \\
&= \frac{(1+\beta^2) \cdot \mathbf{PPV}_k \cdot \mathbf{TPR}_k}{\beta^2 \cdot \mathbf{PPV}_k + \mathbf{TPR}_k} \\
&\overset{2.16}{=} \frac{(1+\beta^2) \cdot \frac{\mathbf{TP}_k}{\mathbf{TP}_k + \mathbf{FP}_k} \cdot \frac{\mathbf{TP}_k}{\mathbf{TP}_k + \mathbf{FN}_k}}{\beta^2 \cdot \frac{\mathbf{TP}_k}{\mathbf{TP}_k + \mathbf{FP}_k} + \frac{\mathbf{TP}_k}{\mathbf{TP}_k + \mathbf{FN}_k}} \\
&= \frac{(1+\beta^2) \cdot \frac{\mathbf{TP}_k \cdot \mathbf{TP}_k}{(\mathbf{TP}_k + \mathbf{FP}_k) \cdot (\mathbf{TP}_k + \mathbf{FN}_k)}}{\frac{\beta^2 \cdot \mathbf{TP}_k \cdot (\mathbf{TP}_k + \mathbf{FN}_k) + \mathbf{TP}_k \cdot (\mathbf{TP}_k + \mathbf{FP}_k)}{(\mathbf{TP}_k + \mathbf{FP}_k) \cdot (\mathbf{TP}_k + \mathbf{FN}_k)}} \\
&= \frac{(1+\beta^2) \cdot \mathbf{TP}_k}{(1+\beta^2) \cdot \mathbf{TP}_k + \beta^2 \cdot \mathbf{FN}_k + \mathbf{FP}_k}.
\end{aligned}$$

□

Corollary 2.46. Let $k \leq C$, then the **F1-Score (F1)** is given by

$$\mathbf{F1}_k := \mathbf{F}_k(1) = \frac{2 \cdot \mathbf{TP}_k}{2 \cdot \mathbf{TP}_k + \mathbf{FN}_k + \mathbf{FP}_k}.$$

The corollary follows immediately from Prop. 2.45. The special case of $\beta = 1$ is one of the most frequent metrics in medical imaging and the formula reveals also its self-duality (see Prop. 2.39). Closely related, though more common in the task of ObD (opposed to ImLC) is the following metric.

Definition 2.47. Let $k \leq C$, then the **Jaccard Index (JAC)** [183] (also **Intersection over Union (IOU)**[a]) for class k is given by

$$\mathbf{JAC}_k := \frac{\mathbf{TP}_k}{\mathbf{TP}_k + \mathbf{FN}_k + \mathbf{FP}_k}.$$

[a] According to the wikipedia entry of the Jaccard index it has been independently formulated by geologist Grove Karl Gilbert in 1884, botanist Paul Jaccard in 1912 and by T.T. Tanimoto in an IBM report in 1957.

The alternative name of IOU originates in a (set) theoretic perspective of measuring the overlap of two sets. Obviously JAC is positively oriented and has a value range of $[0, 1]$. It is rather uncommon as a performance measure of classification (one of the reasons is the following proposition) but more often used as a tool in InS. We can express the close

relation of JAC and F1 as follows:

Proposition 2.48. Let $k \leq C$, then

$$\mathbf{F1}_k = \frac{2 \cdot \mathbf{JAC}_k}{1 + \mathbf{JAC}_k}.$$

Vice versa

$$\mathbf{JAC}_k = \frac{\mathbf{F1}_k}{2 - \mathbf{F1}_k}.$$

Proof. Let $k \leq C$, then

$$\frac{2 \cdot \mathbf{JAC}_k}{1 + \mathbf{JAC}_k} \overset{2.47}{=} \frac{\frac{2\mathbf{TP}_k}{\mathbf{TP}_k + \mathbf{FN}_k + \mathbf{FP}_k}}{\frac{\mathbf{TP}_k + \mathbf{FN}_k + \mathbf{FP}_k}{\mathbf{TP}_k + \mathbf{FN}_k + \mathbf{FP}_k} + \frac{\mathbf{TP}_k}{\mathbf{TP}_k + \mathbf{FN}_k + \mathbf{FP}_k}} = \frac{2\mathbf{TP}_k}{2\mathbf{TP}_k + \mathbf{FN}_k + \mathbf{FP}_k} \overset{2.46}{=} \mathbf{F1}_k.$$

From here we deduce

$$\begin{aligned} \mathbf{F1}_k = \frac{2 \cdot \mathbf{JAC}_k}{1 + \mathbf{JAC}_k} &\Leftrightarrow \mathbf{F1}_k + \mathbf{F1}_k \cdot \mathbf{JAC}_k = 2 \cdot \mathbf{JAC}_k \\ &\Leftrightarrow \mathbf{F1}_k = 2 \cdot \mathbf{JAC}_k - \mathbf{F1}_k \cdot \mathbf{JAC}_k \\ &\Leftrightarrow \mathbf{F1}_k = \mathbf{JAC}_k(2 - \mathbf{F1}_k) \\ &\Leftrightarrow \frac{\mathbf{F1}_k}{2 - \mathbf{F1}_k} = \mathbf{JAC}_k \end{aligned}$$

□

While the F1 only considers TPR and PPV for one class, the next metric considers all classes at once.

Definition 2.49. The **Matthews Correlation Coefficient (MCC)** [146, 244][a] (also **Phi coefficient** [436]) of a categorical model is defined as

$$\mathbf{MCC} := \frac{\mathbf{AC} - \sum_{i \leq C} \mathcal{P}(i) \cdot \mathcal{B}(i)}{\sqrt{1 - \sum_{i \leq C} \mathcal{P}(i)^2} \cdot \sqrt{1 - \sum_{i \leq C} \mathcal{B}(i)^2}}.$$

[a] Matthews original formulation was for the binary case only. There exist multiple different generalizations to the multiclass case [303, 369]. We chose the one from Gorodkin [146] as the oldest and most established one.

The MCC is another rather comprehensive performance measure (compare with WCK or EC) and has a value range of $[-1, 1]$ (though the actual minimum possible value may vary). MCC is positively oriented and a value of zero is interpreted as the performance of 'random guessing'. Although not always easy to interpret in general [450], we want to derive some intuition on MCC and prove our claim it considers TPR_k *and* PPV_k for all classes k. For this we start with a small Lemma.

Lemma 2.50. Let $x \in \mathbb{R}$, then $2x \cdot (1-x) = 1 - x^2 - (1-x)^2$.

Proof.

$$1 - x^2 - (1-x)^2 = 1 - x^2 - 1 + 2x - x^2 = 2x - 2x^2 = 2x \cdot (1-x) \qquad \square$$

The Lemma will help us proof the more common and original binary formulation of MCC.

Proposition 2.51. Let $C = 2$, then

$$\mathbf{MCC} = \frac{\mathbf{TP} \cdot \mathbf{TN} - \mathbf{FP} \cdot \mathbf{FN}}{\sqrt{(\mathbf{TP}+\mathbf{FP})(\mathbf{TP}+\mathbf{FN})(\mathbf{TN}+\mathbf{FP})(\mathbf{TN}+\mathbf{FN})}}.$$

Proof.

$$\begin{aligned}
\mathbf{MCC} &\overset{2.49}{=} \frac{\mathbf{AC} - \sum_{i \le C} \mathcal{P}(i) \cdot \mathcal{B}(i)}{\sqrt{1 - \sum_{i \le C} \mathcal{P}(i)^2} \cdot \sqrt{1 - \sum_{i \le C} \mathcal{B}(i)^2}} \\
&\overset{C=2}{=} \frac{|\mathcal{T}|^{-1} \cdot (\mathbf{TP}+\mathbf{TN}) - \mathcal{P}(1) \cdot \mathcal{B}(1) - \mathcal{P}(2) \cdot \mathcal{B}(2)}{\sqrt{1 - \mathcal{P}(1)^2 - (1-\mathcal{P}(1))^2} \cdot \sqrt{1 - \mathcal{B}(1)^2 - (1-\mathcal{B}(1))^2}} \\
&\overset{2.50}{=} \frac{|\mathcal{T}|^2}{|\mathcal{T}|^2} \cdot \frac{|\mathcal{T}|^{-1} \cdot (\mathbf{TP}+\mathbf{TN}) - \mathcal{P}(1) \cdot \mathcal{B}(1) - \mathcal{P}(2) \cdot \mathcal{B}(2)}{\sqrt{2\mathcal{P}(1) \cdot (1-\mathcal{P}(1))} \cdot \sqrt{2\mathcal{B}(1) \cdot (1-\mathcal{B}(1))}} \\
&= \frac{|\mathcal{T}| \cdot (\mathbf{TP}+\mathbf{TN}) - \mathcal{P}(1) \cdot |\mathcal{T}| \cdot \mathcal{B}(1) \cdot |\mathcal{T}| - \mathcal{P}(2) \cdot |\mathcal{T}| \cdot \mathcal{B}(2) \cdot |\mathcal{T}|}{2 \cdot \sqrt{\mathcal{P}(1) \cdot |\mathcal{T}| \cdot \mathcal{P}(2) \cdot |\mathcal{T}|} \cdot \sqrt{\mathcal{B}(1) \cdot |\mathcal{T}| \cdot \mathcal{B}(2) \cdot |\mathcal{T}|}} \\
&\overset{2.14}{=} \frac{(\mathbf{TP}+\mathbf{FP}+\mathbf{FN}+\mathbf{TN}) \cdot (\mathbf{TP}+\mathbf{TN}) - (\mathbf{TP}+\mathbf{FN}) \cdot (\mathbf{TP}+\mathbf{FP}) - (\mathbf{FP}+\mathbf{TN}) \cdot (\mathbf{FN}+\mathbf{TN})}{2 \cdot \sqrt{(\mathbf{TP}+\mathbf{FP})(\mathbf{TP}+\mathbf{FN})(\mathbf{TN}+\mathbf{FP})(\mathbf{TN}+\mathbf{FN})}} \\
&= \frac{2 \cdot \mathbf{TP} \cdot \mathbf{TN} - 2 \cdot \mathbf{FP} \cdot \mathbf{FN}}{2 \cdot \sqrt{(\mathbf{TP}+\mathbf{FP})(\mathbf{TP}+\mathbf{FN})(\mathbf{TN}+\mathbf{FP})(\mathbf{TN}+\mathbf{FN})}} \\
&= \frac{\mathbf{TP} \cdot \mathbf{TN} - \mathbf{FP} \cdot \mathbf{FN}}{\sqrt{(\mathbf{TP}+\mathbf{FP})(\mathbf{TP}+\mathbf{FN})(\mathbf{TN}+\mathbf{FP})(\mathbf{TN}+\mathbf{FN})}}.
\end{aligned}$$

$\square$

This identity now allows to understand MCC as a summary over J (which itself is a summary over all TPRs in the binary case) and MK (which is for the same case a summary over all PPVs).

Corollary 2.52. Let $C = 2$, then

$$|\mathbf{MCC}| = \sqrt{\mathbf{J} \cdot \mathbf{MK}}.$$

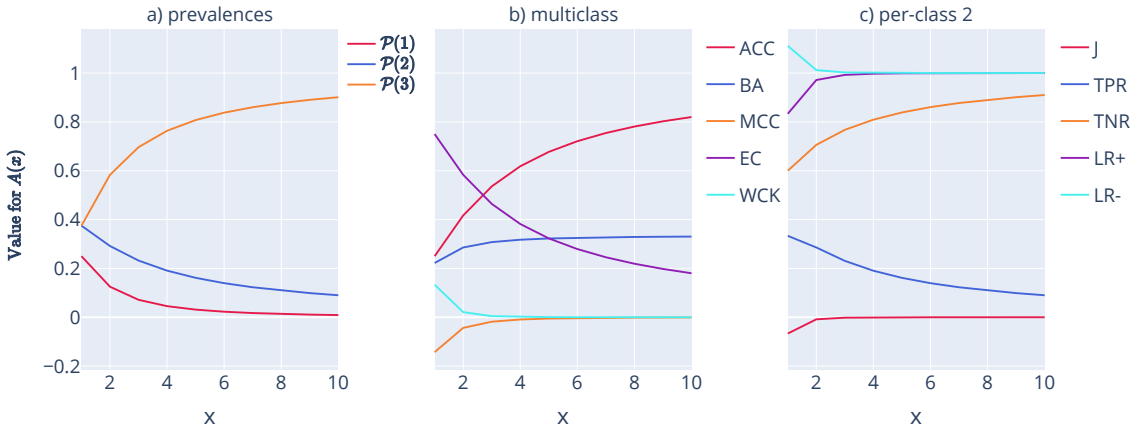


Figure 2.12: Behavior of counting metrics for Ex. 2.53. a) Class prevalences, b) Accuracy (AC), Balanced Accuracy (BA), Matthews Correlation Coefficient (MCC), Expected Cost (EC) (with 0-1-costs) and Weighted Cohen's Kappa (WCK) (with 0-1-costs) and c) Youden's Index (J), True Positive Rate (TPR), True Negative Rate (TNR), Positive Likelihood Ratio (LR+) and Negative Likelihood Ratio (LR-) for the family of confusion matrices given in Ex. 2.53. The latter per-class metrics are depicted for class 2.

Proof. To ease readability we will prove the equivalent squared equation

$$
\begin{aligned}
\mathbf{J}\cdot\mathbf{MK} &\overset{2.33,\,2.41}{=} (\mathbf{TPR}+\mathbf{TNR}-1)\cdot(\mathbf{PPV}+\mathbf{NPV}-1)\\
&\overset{2.17,\,2.36}{=} (\mathbf{TPR}-\mathbf{FPR})\cdot(\mathbf{PPV}-\mathbf{FOR})\\
&\overset{2.16,\,2.35}{=} \left(\frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FN}}-\frac{\mathbf{FP}}{\mathbf{FP}+\mathbf{TN}}\right)\cdot\left(\frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}}-\frac{\mathbf{FN}}{\mathbf{FN}+\mathbf{TN}}\right)\\
&= \frac{\mathbf{TP}(\mathbf{FP}+\mathbf{TN})-\mathbf{FP}(\mathbf{TP}+\mathbf{FN})}{(\mathbf{TP}+\mathbf{FN})(\mathbf{FP}+\mathbf{TN})}\cdot\frac{\mathbf{TP}(\mathbf{FN}+\mathbf{TN})-\mathbf{FN}(\mathbf{TP}+\mathbf{FP})}{(\mathbf{TP}+\mathbf{FP})(\mathbf{FN}+\mathbf{TN})}\\
&= \frac{\mathbf{TP}\cdot\mathbf{TN}-\mathbf{FP}\cdot\mathbf{FN}}{(\mathbf{TP}+\mathbf{FN})(\mathbf{FP}+\mathbf{TN})}\cdot\frac{\mathbf{TP}\cdot\mathbf{TN}-\mathbf{FN}\cdot\mathbf{FP}}{(\mathbf{TP}+\mathbf{FP})(\mathbf{FN}+\mathbf{TN})}\\
&= \frac{(\mathbf{TP}\cdot\mathbf{TN}-\mathbf{FP}\cdot\mathbf{FN})^2}{(\mathbf{TP}+\mathbf{FN})(\mathbf{FP}+\mathbf{TN})(\mathbf{TP}+\mathbf{FP})(\mathbf{FN}+\mathbf{TN})}\\
&\overset{2.51}{=} \mathbf{MCC}^2
\end{aligned}
$$

□

We want to close our introduction of counting metrics with some examples, that shed light on the different behavior of counting metrics.

Example 2.53. Fix $C = 3$ and a series of tasks $\{\mathcal{T}_x\}_{x\in\mathbb{N}}$, that evaluated by a model

φ produces the given family of confusion matrices

$$\boldsymbol{A}(x) := \begin{bmatrix} 0 & 1 & x \\ 1 & x & x^2 \\ x & x^2 & x^3 \end{bmatrix}.$$

$\boldsymbol{A}(x)$ is symmetric, i. e., self-dual. We depict the behavior of prevalences, multiclass metrics and per-class counting metrics (for class 2) on this family in Fig. 2.12.

Although the Sensitivity of class 1 remains zero independently of x the overall AC of $\boldsymbol{A}(x)$ surpasses 80% at $x = 9$. On the other hand BA approaches 1/3, meanwhile the vast majority of predictions are correct with increasing x. Because of the symmetry in $\boldsymbol{A}(x)$ we know $\mathbf{MK}_2 = \mathbf{J}_2$ (Prop. 2.42), $\mathbf{PPV}_2 = \mathbf{TPR}_2 = \mathbf{F1}_2$ (Prop. 2.39, Def. 2.44).

2.5 Curves and multi-threshold metrics

After the many counting metrics presented before, we turn our attention to performance measures μ that do not require the model φ to conduct categorical decisions. We will assume the model output as class probabilities though, e. g., via the softmax σ. Recall that $\boldsymbol{A}^{(k)}$ is the one-versus-the-rest confusion matrix merging all classes except for class $k \leq C$ (Def. 2.15). Also recall the threshold operator ρ_τ, which for $C = 2$ and $\tau \in [0, 1]$ yields a binary decision. We slightly extend these concept to the multiclass case:

Definition 2.54. Let φ be a model that produces probabilities, $\tau \in [0, 1]$, $\mathcal{T}$ be a task with C classes and $k \leq C$, then $\boldsymbol{A}^{(k|\tau)}$ shall be the confusion matrix for model $\rho_\tau \circ \tilde{h}_k \circ \varphi$ and binary task $\mathcal{T}^{(k)}$. Here the probability merging helper function $\tilde{h}_k : [0, 1]^C \mapsto [0, 1]^2$ is defined as

$$\tilde{h}_k(p) := (p_k, 1 - p_k).$$

Given a *sequence of thresholds* $\tau_1, ..., \tau_n$ this allows to compute a *sequence of performances* $\mu_1, ..., \mu_n$, where μ_i is a performance measure for task $\mathcal{T}^{(k)}$ on model $\rho_{\tau_i} \circ \tilde{h}_k \circ \varphi$, hence may access $\boldsymbol{A}^{(k|\tau)}$. A natural choice for $\tau_1, ..., \tau_n$ is the (strictly monotonous increasing) sorted list of probabilities $\{\varphi(x)_k : (x, y) \in \mathcal{T}\}$ – if necessary prepended by zero and followed by 1. For very large $\mathcal{T}$ subsampling from these thresholds might be appropriate.

Definition 2.55. The **Receiver Operating Characteristic (ROC)** [156] for class $k \leq C$ is a sequence of pairs $\{(a_i, b_i)\}_{i \leq N}$, where for a sequence of appropriately chosen thresholds $\tau_1, ..., \tau_n$ the elements are given by $a_i := \mathbf{FPR}, b_i := \mathbf{TPR}$ on

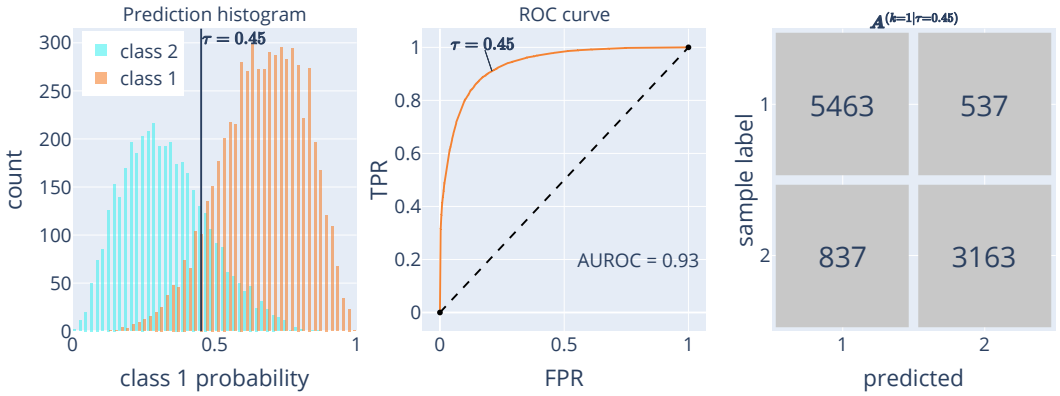


Figure 2.13: Receiver Operating Characteristic (ROC) visualization. **Left**: Histogram of model predictions, generated by a skewed beta distribution for 10 000 samples and a prevalence of 0.6. **Center**: ROC curve for class 1, the dashed line represents a naive classifier. **Right**: Confusion matrix at the specific threshold τ, which is also indicated in the other two subplots.

$\boldsymbol{A}^{(k|\tau_i)}$. The area under the curve on the interval $[0, 1]$ that results as linear interpolation between consecutive points is called **Area under the Receiver Operating Characteristic Curve (AUROC)** (also **Area under the curve (AUC)**).

The ROC curve originated during World War II for radar signal detection and distinguishing enemy objects from noise [241]. It was later adopted in a variety of domains and nowadays is a common metric in image classification [240]. The AUROC has an elegant interpretation as the probability of any randomly picked sample from class k to have a higher predicted probability as a randomly picked sample from any other class [111]. ROC curves also nicely visualize the J of a threshold as the vertical line between the curve and the diagonal representing an uninformed classifier [339]. The value range of AUROC is $[0, 1]$, it is positively oriented and a value of 0.5 may be interpreted as random guessing. The connections between predicted class probabilities, the ROC curve and corresponding confusion matrices at thresholds is visualized in Fig. 2.13.

Definition 2.56. The **Precision-Recall (PR) Curve** [228] for class $k \leq C$ is a sequence of pairs $\{(a_i, b_i)\}_{i \leq N}$, where for a sequence of appropriately chosen thresholds $\tau_1, ..., \tau_n$ the elements are given by $a_i :=$ **TPR**, $b_i :=$ **PPV** on $\boldsymbol{A}^{(k|\tau_i)}$. The summarization of this curve is achieved as a weighted mean of precisions at each threshold: **AP** $:= \sum_{1 \leq i \leq N} (a_i - a_{i-1}) \cdot b_i$ and called the **Average Precision (AP)**.

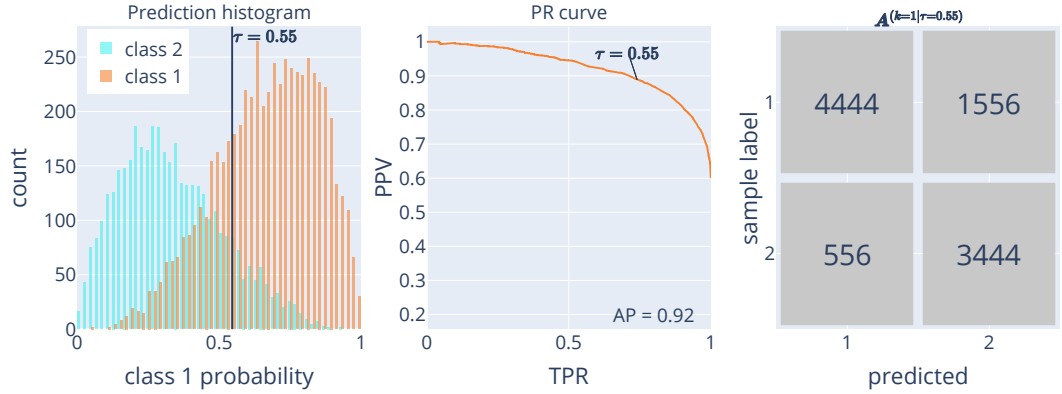


Figure 2.14: Precision-Recall (PR) visualization. **Left**: Histogram of model predictions, generated by a skewed beta distribution for 10 000 samples and a prevalence of 0.6. **Center**: PR curve for class 1. **Right**: Confusion matrix at the specific threshold τ, which is also indicated in the other two subplots.

For the PR-curve linear interpolation would yield too optimistic results, as the change in PPV must not be linear along variations in TPR [87]. While PR-curves and ROC-curves are connected, optimizing one does not necessarily guarantee to optimize the other [87]. Note that the TN of the confusion matrix $\boldsymbol{A}^{(k|\tau)}$ are not used during the computation of the coordinates of the curve. This is one of the reasons AP is commonly used in ObD tasks, where this entity is actually undefined [108]. Similar to AUROC the value range of AP is $[0, 1]$ and it is positively oriented, but there is no such fixed value for interpretation of a random classifier. Fig. 2.14 shows an example PR curve.

Definition 2.57. The **Decision Curve** [402] for class $k \leq C$ is a sequence of pairs $\{(\tau_i, b_i)\}_{i \leq N}$, where for a sequence of appropriately chosen thresholds $\tau_1, ..., \tau_n$ the elements are given by $b_i := \mathbf{NB}(\tau_i)$ on $\boldsymbol{A}^{(k|\tau_i)}$. Here the **Net Benefit (NB)** is given by

$$\mathbf{NB}(\tau) := |\mathcal{T}|^{-1}(\mathbf{TP} - \mathbf{FP} \cdot \frac{\tau}{1-\tau}).$$

Decision curves and net benefit are less commonly observed tools in image classification models. The goal of NB is to combine benefits (e. g., detecting disease) and harms (e. g., unnecessary procedures) on a single scale by using an 'exchange rate' that reflects clinical judgment on their relative importance (the *odds ratio* of the threshold). Decision curve plots allow clinicians to evaluate whether using a prediction model would provide clinical value compared to treating all or no patients [403]. This is achieved by incorporating the (range of) threshold(s) at which a clinician would recommend intervention - effectively

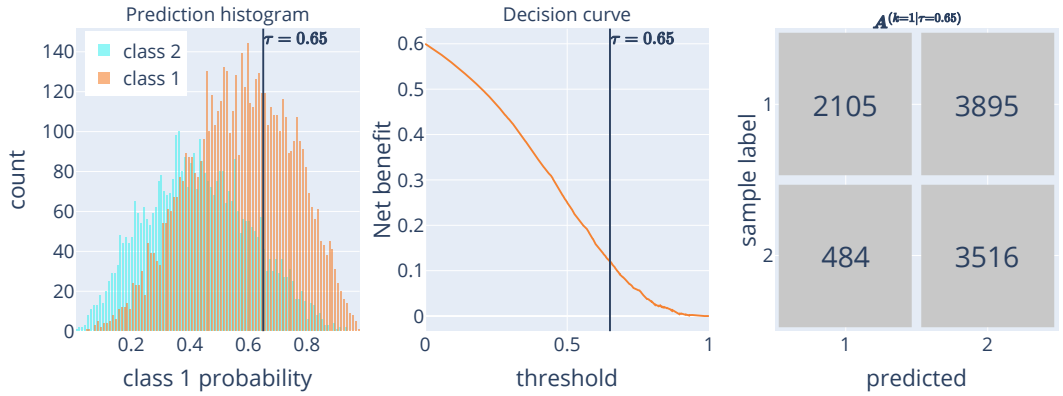


Figure 2.15: Decision curve visualization. **Left**: Histogram of model predictions, generated by a skewed beta distribution for 10 000 samples and a prevalence of 0.6. **Center**: Decision curve for class 1. **Right**: Confusion matrix at the specific threshold τ, which is also indicated in the other two subplots.

capturing their assessment of the relative costs of false positives versus false negatives in that specific clinical context. The value range of NB is $(-\infty, 1]$, and it is positively oriented. We conclude this section with a useful note on interpreting NB.

Remark 2.58. For some threshold $\tau \in (0,1)$ and $k \leq C$ it holds that on $\boldsymbol{A}^{(k|\tau)}$

$$\mathbf{NB}(\tau) = \mathcal{P}(1) \cdot \mathbf{TPR} - \mathcal{P}(2) \cdot (1 - \mathbf{TNR}) \cdot \frac{\tau}{1-\tau}.$$

Proof.

$$\begin{aligned}
\mathbf{NB}(\tau) &\overset{2.57}{=} |\mathcal{T}|^{-1}(\mathbf{TP} - \mathbf{FP} \cdot \frac{\tau}{1-\tau}) \\
&= \frac{\mathbf{TP}}{|\mathcal{T}|} - \frac{\mathbf{FP}}{|\mathcal{T}|} \cdot \frac{\tau}{1-\tau} \\
&= \frac{\mathbf{TP}+\mathbf{FN}}{|\mathcal{T}|} \cdot \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FN}} - \frac{\mathbf{FP}+\mathbf{TN}}{|\mathcal{T}|} \cdot \frac{\mathbf{FP}}{\mathbf{FP}+\mathbf{TN}} \cdot \frac{\tau}{1-\tau} \\
&\overset{2.12,\,2.16}{=} \mathcal{P}(1) \cdot \mathbf{TPR} - \mathcal{P}(2) \cdot \mathbf{FPR} \cdot \frac{\tau}{1-\tau} \\
&\overset{2.17}{=} \mathcal{P}(1) \cdot \mathbf{TPR} - \mathcal{P}(2) \cdot (1 - \mathbf{TNR}) \cdot \frac{\tau}{1-\tau}
\end{aligned}$$

□

2.6 Calibration

The performance measures presented so far all focus on the *discriminative* capabilities of the model φ – investigating how well *decisions* are done. A someway complementary question is for the *calibration* of the model φ – how well do the *predicted probabilities* perform?

2.6.1 Variants of calibration

To precisely capture the meaning of a *calibrated* or *reliable* classifier [393] we need to slightly dive into probability theory (see Def. 2.3).

Definition 2.59. Let φ be a probabilistic model for task $\mathcal{T}$. We call φ [393]

(i) **canonically calibrated** iff

$$\forall k \leq C, p \in \Delta_{C-1} : \mathbb{P}[Y = k | \varphi(X) = p] = p_k$$

(ii) **class-wise calibrated** (also **marginally calibrated**) iff

$$\forall k \leq C, q \in [0,1] : \mathbb{P}[Y = k | \varphi(X)_k = q] = q$$

(iii) **top-label calibrated** (also **confidence calibrated**) iff

$$\forall q \in [0,1] : \mathbb{P}[Y = \operatorname{argmax} \varphi(X) | \max \varphi(X) = q] = q.$$

Note that we require equality only as long as the conditional probabilities are defined (the condition has positive probability). Or phrased with the words of measure theory, the above equal signs are *almost surely* equal signs.

There is one important pitfall in the interpretation of model calibration, we want to stress from the beginning. The conditioning of the probabilities we used in the definitions of calibrated models (Def. 2.59) is always on $\varphi(X)$ – not X itself. So a perfectly calibrated model may not need to output $\mathbb{P}[Y = k|X]$ – the 'true posterior' probability, which will be shown in Ex. 2.63.

Proposition 2.60. Let φ be a probabilistic model for task $\mathcal{T}$. Then

(i) φ is canonically calibrated $\Rightarrow$ φ is class-wise calibrated,

(ii) φ is canonically calibrated $\Rightarrow$ φ is top-label calibrated.

Table 2.2: First counterexample calibration notions. The model φ is class-wise calibrated while not being top-label calibrated. Example from Chen et al. [59].

	$\varphi(X)_k$				$\mathbb{P}[Y = k\|X]$			
	k = 1	k = 2	k = 3	k = 4	k = 1	k = 2	k = 3	k = 4
X = 1	0.3	0.25	0.2	0.25	0.4	0.25	0.1	0.25
X = 2	0.3	0.5	0.2	0	0.2	0.5	0.3	0

Table 2.3: Second counterexample calibration notions. The model φ is top-label calibrated while not being class-wise top-label calibrated. Example from Vaicenavicius et al. [393].

	$\varphi(X)_k$			$\mathbb{P}[Y = k\|X]$		
	k = 1	k = 2	k = 3	k = 1	k = 2	k = 3
X = 1	0.6	0.1	0.3	0.7	0.2	0.1
X = 2	0.4	0.6	0	0.3	0.5	0.2

Proofs for Prop. 2.60 (i) and (ii) can be found in Gruber et al. [150]. Vaicenavicius et al. [393] state furthermore, that if $\mathcal{T}$ is a binary task, then the three notions of calibration coincide, but do not provide a proof for this. Importantly the coincidence of calibration notions is *not* given in general, as shown by the following examples.

Example 2.61. The examples will only attach mass to finitely many images with respect to $p(X, Y)$. For simplicity, we will identify these with an initial segment of $\mathbb{N}$. Let $\mathbb{P}[X = 1] = \mathbb{P}[X = 2] = 0.5$ and φ as well as $\mathbb{P}[Y|X]$ be defined as in

(i) Tab. 2.2. We observe φ to be class-wise calibrated while not being top-label calibrated: To show that φ is class-wise calibrated, we show for each class $k \leq C$ that for all $q \in \text{Im}(\varphi_k)$ holds that $\mathbb{P}[Y = k|\varphi(X)_k = q] = q$.

k=1: $\text{Im}(\varphi_1) = \{0.3\}$ and $\mathbb{P}[Y = 1|\varphi(X)_1 = 0.3] = \mathbb{P}[X = 1]\cdot 0.4 + \mathbb{P}[X = 2] \cdot 0.2 = 0.3$

k=2: $\text{Im}(\varphi_2) = \{0.25, 0.5\}$ and $\mathbb{P}[Y = 2|\varphi(X)_2 = 0.25] = \frac{\mathbb{P}[X=1]\cdot 0.25}{\mathbb{P}[X=1]} = 0.25$ as well as $\mathbb{P}[Y = 2|\varphi(X)_2 = 0.5] = \frac{\mathbb{P}[X=2]\cdot 0.5}{\mathbb{P}[X=2]} = 0.5$

k=3: $\text{Im}(\varphi_3) = \{0.2\}$ and $\mathbb{P}[Y = 3|\varphi(X)_3 = 0.2] = \mathbb{P}[X = 1]\cdot 0.2 + \mathbb{P}[X = 2] \cdot 0.2 = 0.2$

k=4: $\text{Im}(\varphi_4) = \{0.25, 0\}$ and $\mathbb{P}[Y = 4|\varphi(X)_4 = 0.25] = \frac{\mathbb{P}[X=1]\cdot 0.25}{\mathbb{P}[X=1]} = 0.25$ as well as $\mathbb{P}[Y = 4|\varphi(X)_4 = 0] = \frac{\mathbb{P}[X=2]\cdot 0}{\mathbb{P}[X=2]} = 0$

To prove that φ is not top-label calibrated, we can simply show that for $q = 0.3$ the necessary equation $\mathbb{P}[Y = \operatorname{argmax} \varphi(X) | \max \varphi(X) = q] = q$ fails, as $\mathbb{P}[Y = \operatorname{argmax} \varphi(X) | \max \varphi(X) = 0.3] = \mathbb{P}[Y = 1 | X = 1] = 0.4$.

(ii) Tab. 2.3. We observe φ to be top-label calibrated while not being class-wise calibrated: To show that φ is top-label calibrated, we show for each $q \in \operatorname{Im}(\max \varphi)$ holds that $\mathbb{P}[Y = \operatorname{argmax} \varphi(X) | \max \varphi(X) = q] = q$. As $\operatorname{Im}(\max \varphi) = \{0.6\}$ we only have $q = 0.6$ to verify this, which comes down to $\mathbb{P}[Y = \operatorname{argmax} \varphi(X) | \max \varphi(X) = 0.6] = \mathbb{P}[X = 1] \cdot \mathbb{P}[Y = \operatorname{argmax} \varphi(X) | X = 1] + \mathbb{P}[X = 2] \cdot \mathbb{P}[Y = \operatorname{argmax} \varphi(X) | X = 2] = 0.5 \cdot 0.7 + 0.5 \cdot 0.5 = 0.6$. To prove φ is not class-wise calibrated we can inspect $k = 1$ and $q = 0.6 \in \operatorname{Im}(\varphi_1)$. Here $\mathbb{P}[Y = 1 | \varphi(X)_1 = 0.6] = \mathbb{P}[Y = 1 | X = 1] = 0.7$.

We can immediately conclude that neither class-wise calibration nor top-label calibration imply canonical calibration (otherwise by Prop. 2.60 they would also imply each other, which has just been shown to be incorrect). An explicit counterexample for this was also given by Vaicenavicius et al. [393]. This shows that canonical calibration is the strongest notion. And indeed for a 'perfect' model it is necessary to be canonically calibrated.

Proposition 2.62. Let $\mathcal{T}$ be a task with C classes, then the **optimal classifier** $\varphi^\star$, given by $\forall k \leq C : \varphi^\star(x)_k := \mathbb{P}[Y = k | X = x]$ is canonically calibrated.

Proof. Let $k \leq C, p \in \Delta_{C-1}$, then $\mathbb{P}[Y = k | \varphi^\star(X) = p] = \mathbb{P}[Y = k | \forall_i \mathbb{P}[Y = i] = p_i] = p_k$. □

Unfortunately canonical calibration does not imply to be close to the optimal classifier, as shown by the next example.

Example 2.63. Let $\mathcal{T}$ be a task with C classes and class prevalences $\mathcal{P}$, then the constant model φ with $\varphi : x \mapsto \mathcal{P}$ is canonically calibrated; the image of φ is a singleton which perfectly matches the overall distribution of $\mathcal{T}$. Example from Chen et al. [59].

The model given in Ex. 2.63 is obviously a very uninformative trivial one (compare to the naive classifier defined in Def. 2.26) and does not internalize any discriminative value. As long as X contains any clue about Y (e. g., the image depicts the object associated with its class label) we would assume that $p_{\mathcal{T}}(Y|X)$ differs from the pure class distribution $p_{\mathcal{T}}(Y)$. With these considerations in mind we can agree to call calibration complementary to discrimination. Note that one of the fundamental differences is also the dependence

on a decision rule (see Def. 2.9).

2.6.2 Measuring miscalibration

So far we only defined how a calibrated model behaves. Unfortunately hardly any model will perfectly match any of our notions on calibration. The goal of a *calibration error* measure is to quantify the level of divergence between the two sides of the equal sign in the calibration definitions.

Definition 2.64. Let $\mathcal{T}$ be a task with C classes and φ a probabilistic model for $\mathcal{T}$. Then $r : \text{Im}(\varphi) \to \Delta_{C-1}$ with

$$r(p)_k := \mathbb{P}[Y = k | \varphi(X) = p]$$

is called the **(canonical) (re-)calibration function** [393]. Now let $d : \Delta_{C-1} \times \Delta_{C-1} \to [0, \infty)$ be a distance function, then the **Calibration Error (CE)** is defined as

$$\textbf{CE}_d := \mathbb{E}[d(r(\varphi(X)), \varphi(X))].$$

This definition may be interpreted from different angles. Assume to be given a 'fixed' classifier φ, then the canonical re-calibration function is the best possible post-processing function, such that $r \circ \varphi$ is canonically calibrated (though not necessarily accurate) [281]. If a model is already canonically calibrated, then r collapses to the identity function and (given d maps pairs with identical entries to zero) the CE thus turns zero. Measuring miscalibration can thus also be interpreted as measuring the difference between r and the identity function. Which implies that the computation of CE and the computation of the canonical re-calibration function are equally hard [281]. Similar formulations can also be found for class-wise and top-label calibration.

Definition 2.65. Let $\mathcal{T}$ be a task with C classes and φ a probabilistic model for $\mathcal{T}$. Then for $k \leq C$ the **marginal (re-)calibration function** $r : \text{Im}(\varphi) \to [0, 1]$ is given by

$$r(p) := \mathbb{P}[Y = k | \varphi(X)_k = p_k]$$

Now let $d : [0, 1] \times [0, 1] \to [0, \infty)$ be a distance function, the **Marginal Calibration Error (MCE)** [213][a] for class k is given by

$$\textbf{MCE}_d := \mathbb{E}[d(r(\varphi(X)), \varphi(X)_k)].$$

The **top-label (re-)calibration function** $r : \text{Im}(\varphi) \to [0, 1]$ is given by

$$r(p) := \mathbb{P}[Y = \text{argmax}\, p | \max \varphi(X) = \max p]$$

Now let $d : [0,1] \times [0,1] \to [0,\infty)$ be a distance function, the **Top-label Calibration Error (TCE)** [213] is given by

$$\mathbf{TCE}_d := \mathbb{E}[d(r(\varphi(X)), \max \varphi(X))].$$

[a] In the literature there is some ambiguity about the generic calibration errors and some approximations. In this case the MCE is also often referred to as 'class-wise calibration error' [150], or to the calibration error induced by the 'fixed partition' [393]. We will use the term class-wise calibration error for an aggregated estimator of the MCE in Def. 2.68. Also be aware that 'Maximum Calibration Error' [151], which is also abbreviated as MCE is unrelated to our definition and will not be of further interest in this thesis.

Unfortunately the true distribution $p_{\mathcal{T}}(X, Y)$ is unknown for the most cases and must be approximated by the available data from $\mathcal{T}$. Depending on the details of the approximation a bias (i. e., a degree of over or underestimation) and the convergence rate (i. e., how much additional samples improve the approximation) may be derived. Combined with the computational complexity of the estimator (both along the number of samples and the number of classes) various trade-offs arise. For most models the predicted probabilities will differ for each image in $\mathcal{T}$, thus the straight estimation of r would only be built upon a single realization. One approach to circumvent this, groups nearby probabilities into bins.

Definition 2.66. Let $\{B_i\}_{i \leq m}$ be a finite partitioning of $[0,1]$ (the **bins**), $\mathcal{T}$ be a task with C classes and φ a probabilistic model for $\mathcal{T}$. Then the **Expected Calibration Error (ECE)** [265] of φ is given by

$$\mathbf{ECE} := \sum_{i \leq m} \mathbb{P}[\max \varphi(X) \in B_i] \big| \mathbb{E}[\max \varphi(X) \,|\, \max \varphi(X) \in B_i] - \mathbb{P}[Y = \operatorname{argmax} \varphi(X) \,|\, \max \varphi(X) \in B_i] \big|.$$

The three terms in the sum are also often referred to as *bin-frequency*, *bin-wise mean confidence* and *bin-wise accuracy* [150]. It has been shown that $\mathbf{ECE} \leq \mathbf{TCE}$ [150, 393][14] and an abstract derivation of ECE from the general EC was given by Vaicenavicius et al. [393]. The ECE was "the first calibration estimator for a continuous one-vs-all multiclass mode [...] and is still the most commonly used measure to quantify calibration" [150]. Widespread usage may probably be partly due to the work by Guo et al. [151], who happened to introduce a simple model for the top-label re-calibration function.

[14] Kumar et al. [213] show in general that any binning scheme underestimates the calibration error.

Definition 2.67. Let $t \in \mathbb{R}_+$ be a positive real, then the function

$$f_{temp}(p) := \sigma(p/t)$$

is called **temperature scaling** [151][a].

[a] The original definition already incorporated the maximum of the softmax σ, as it would be required to model the *top-label re-calibration function*, but to evaluate temperature scaling in a setup that measures canonical or class-wise calibration sense we neglect it here (see Sec. 6.1).

Fitting the parameter t for the model $f_{temp} \circ \varphi$ such that the CE minimizes then corresponds to approximating the re-calibration function with f_{temp}. Apparently this is a very simple model, though Guo et al. [151] find it sufficient. Note though that their definition of miscalibration is given by the TCE and as proven by Gruber and Buettner [150], a minimal TCE implies minimal ECE. Nevertheless, triggered by these discoveries the topic of calibration in DNN gained popularity, which lead to further advances.

Definition 2.68. Let $\{B_i\}_{i \leq m}$ be a finite partitioning of $[0, 1]$ (the **bins**), $\mathcal{T}$ be a task with C classes and φ a probabilistic model for $\mathcal{T}$. Then the **Class-wise Calibration Error (CWCE)** [211, 213] of φ is given by

$$\textbf{CWCE} := C^{-1} \sum_{k \leq C} \sum_{i \leq m} \mathbb{P}[\varphi(X)_k \in B_i] \cdot |\mathbb{E}[\varphi(X)_k | \varphi(X)_k \in B_i] - \mathbb{P}[Y = k | \varphi(X)_k \in B_i]|.$$

The equal aggregation across classes is somewhat arbitrary as noted by Panchenko et al. [281] and weighted versions of the sum over the classes are also found in literature [213]. In both cases of ECE and CWCE we chose the L_1 norm as distance function, but the general L_p or specifically L_2 are also common [150, 213]. We also note that the probabilities and expectations in the computation of EC and CWCE are in practice obviously calculated over the respective task. Both ECE and CWCE are negatively oriented and have a value range of $[0, 1]$. But still neither of them captures canonical calibration – and both are non-differentiable. The latter is a property which is especially useful if a calibration error is used for the training of NN (see Def. 2.75).

Both of these issues are tackled with the following calibration error. Unfortunately a lot of the underlying theory which is based upon **Reproducing Kernel Hilbert Space (RKHS)** is out of scope for this thesis, thus we only give an abstract estimator and refer the reader to the details given by Widmann et al. [419].

Definition 2.69. The **Kernel Calibration Error (KCE)** [419] is given by

$$\mathbf{KCE} := (\mathbb{E}_{X,Y\ p_{\mathcal{T}},X',Y'\ p_{\mathcal{T}}}[e_Y - \varphi(X)^T k(\varphi(X),\varphi(X'))(e_{Y'} - \varphi(X'))])^{1/2},$$

with the matrix-valued kernel k and the canonical unit vectors e_i.

Widmann et al. [419] proofed that if k is a universal kernel, then $\mathbf{KCE} = 0 \Leftrightarrow \mathbf{CE} = 0$, i. e., φ is canonically calibrated. The interpretation of KCE is difficult though, as the value range depends on the kernel choice. Furthermore – although non-negative in expectation – concrete approximations of KCE may even turn negative. It is also possible to use **Kernel Density Estimation** [286, 326] to approximate the canonical calibration function. If the distance function d is chosen to be the L_p norm this gives our next calibration performance measure.

Definition 2.70. Let $\mathcal{T} = \{(x_i, y_i)\}_{i \leq N}$ be a task with C classes and φ a probabilistic model for $\mathcal{T}$. Then the **Expected Calibration Error Kernel Density Estimate ($\mathbf{ECE}^{\mathbf{KDE}}$)** [302] is given by

$$\mathbf{ECE}_p^{\mathbf{KDE}} := \frac{1}{N} \sum_{j \leq N} \left\| \frac{\sum_{i \neq j} K(\varphi(x_j), \varphi(x_i)) \cdot e_{y_i}}{\sum_{i \neq j} K(\varphi(x_j), \varphi(x_i))} - \varphi(x_j) \right\|_p^p,$$

where K is the **kernel** (e. g., a Dirichlet kernel [276]) and e_i is the i-th canonical unit vector.

$\mathrm{ECE}^{\mathrm{KDE}}$ tackles canonical calibration directly and Popordanoska et al. [302] also proved theoretical guarantees for the convergence, bias and computation complexity (along the number of samples). The value range of $\mathrm{ECE}^{\mathrm{KDE}}$ is $[0, 2]$, and it is negatively oriented. Though these properties may make $\mathrm{ECE}^{\mathrm{KDE}}$ appear more suitable for miscalibration estimation than KCE, it must be said that KCE in contrast is an unbiased estimator. Furthermore, it remains an issue that a canonically calibrated model may still be far from the optimal model (see Ex. 2.63).

2.6.3 Mixed assessment of calibration and discrimination

It is thus necessary to find an even stronger notion for a calibrated model – which we will find through the concept of *scoring rules*. The trade-off is that such scores do not only measure calibration but jointly incorporate discrimination performance.

Definition 2.71. A **scoring rule** [44, 150] for task $\mathcal{T}$ with C classes is a function $S : \Delta_{C-1} \times Y_{\mathcal{T}} \to \mathbb{R} \cup \{-\infty, \infty\}$. It measures the disagreement[a] between

a probabilistic prediction and some reference class. The corresponding **scoring function** $s : \Delta_{C-1} \times \Delta_{C-1} \to \mathbb{R} \cup \{-ty, \infty\}$ is given by $s(p, q) := \sum_{k \leq C} S(p, k) \cdot q_k$ and measures the expected disagreement between a probabilistic prediction and a probabilistic reference. A score is called **proper** if the **divergence** $d : \Delta_{C-1} \times \Delta_{C-1} \to \mathbb{R} \cup \{-\infty, \infty\}$, given by $d(p, q) := s(p, q) - s(q, q)$ is non-negative, i. e., for all $p, q \in \Delta_{C-1}$ holds that $s(p, q) \geq s(q, q)$. The score is called **strictly proper** if $d(p, q) = 0 \Rightarrow p = q$ for all $p, q \in \Delta_{C-1}$.

[a] The orientation of scoring rules is ambiguous in literature and usually one would expect a 'score' to be positively oriented. Following the standard perspective in ML of minimizing a loss function we will also interpret scoring rules to be negatively oriented. The general conclusions are not impaired by this, though some signs might change in derived formulas.

Note that the interpretation of the divergence may only be really meaningful for proper scoring rules, but be aware that in general the divergence of a proper scoring rule is not a true metric, specifically it must neither be symmetric nor does the triangle inadequately hold.[15] It can be shown that every proper scoring rule can be decomposed as follows:

Theorem 2.72. Let S be a strictly proper scoring rule for task $\mathcal{T}$ with C classes and φ a model for $\mathcal{T}$. Then

$$\mathbb{E}[S(\varphi(X), Y)] = S(\mathcal{P}, \mathcal{P}) - \mathbb{E}[d(\mathcal{P}, \mathbb{P}[Y|\varphi(X)])] + \mathbb{E}[d(\varphi(X), \mathbb{P}[Y|\varphi(X)])].$$

The three summands on the right are called (from left to right) *uncertainty*, *resolution* (also *sharpness*) and *reliability*.

The theorem was proven by Bröcker [44] in general for the multiclass case, while the decomposition for the binary case has been known for longer [90]. While the uncertainty is a constant only depending on $\mathcal{T}$, the resolution measures the divergence of φ with the uninformative naive classifier that only knows the class prevalences (see Ex. 2.63). Finally, the reliability term is a canonical calibration function (see Def. 2.64), since S is strictly proper, i. e., d is non-negative.[16] After all this theoretical background it is time to introduce an actual strictly proper scoring rule.

Definition 2.73. The **Logarithmic scoring rule** [131] (also **Ignorance scoring rule** [44]) for task $\mathcal{T}$ with C classes is the scoring rule defined by $S(p, i) := -\ln p_i$. The **Negative Log Likelihood (NLL)** (also **Cross Entropy**) for some model φ on

[15] We also note that the theory of scoring rules transfers to the continues case [131], opposed to the categorical domain $Y_{\mathcal{T}}$ we chose.

[16] Formally we switched the order of arguments from Def. 2.64 to the interpretation in Thm. 2.72, but defining a dummy $\tilde{d}(p, q) := d(q, p)$ resolves the issue.

$\mathcal{T}$ is given by

$$\begin{aligned}\textbf{NLL} &:= \mathbb{E}_{X,Y\sim\mathcal{T}} S(\varphi(X), Y) \\ &= |\mathcal{T}|^{-1} \sum_{(x_i,y_i)\in\mathcal{T}} S(\varphi(x_i), y_i) = -|\mathcal{T}|^{-1} \sum_{(x_i,y_i)\in\mathcal{T}} \ln \varphi(x_i)_{y_i}.\end{aligned}$$

The corresponding divergence to S is called **Kullback-Leibler Divergence (KLD)** [212] and given by

$$\textbf{KLD}(p,q) := s(p,q) - s(q,q) = \sum_{k\leq C} S(p,k)\cdot q_k - \sum_{k\leq C} S(q,k)\cdot q_k = \sum_{k\leq C} q_k \cdot \ln\frac{q_k}{p_k}.$$

The Logarithmic scoring rule is a strictly proper scoring rule [131] and the KLD is an important tool in information theory, which we will also use later to compare tasks (see Def. 5.13). In the literature the notation $\textbf{KLD}(p||q) := \textbf{KLD}(q,p)$ is also common, which might cause confusion by the switch of arguments. The NLL is very often used as a loss function in DL (see Def. 2.75). The value range of NLL is $[0,\infty)$, and it is negatively oriented.

Definition 2.74. The **Brier scoring rule** [43] (also **Quadratic scoring rule** [131]) for task $\mathcal{T}$ with C classes is the scoring rule defined by $S(p,i) := \sum_{k\leq C}(\delta_{ik} - p_k)^2$. Here δ_{ij} denotes the Kronecker delta (equalling 1 if $i=j$ and 0 otherwise). The **Brier Score (BS)** for some mode φ on $\mathcal{T}$ is given by

$$\begin{aligned}\textbf{BS} &:= \mathbb{E}_{X,Y\sim\mathcal{T}} S(\varphi(X), Y) \\ &= |\mathcal{T}|^{-1} \sum_{(x_i,y_i)\in\mathcal{T}} S(\varphi(x_i), y_i) = |\mathcal{T}|^{-1} \sum_{(x_i,y_i)\in\mathcal{T}} \sum_{k\leq C} (\delta_{y_i k} - \varphi(x_i)_k)^2.\end{aligned}$$

The **Brier Skill Score (BSS)** [131, 400] for some model φ on $\mathcal{T}$ is given by

$$\textbf{BSS} := 1 - \frac{BS}{BS_{naive}},$$

where $\textbf{BS}_{\textbf{naive}}$ is the BS of the naive probabilistic model given in Ex. 2.63, which constantly predicts the prevalences of $\mathcal{T}$.

The Brier scoring rule is a strictly proper scoring rule [131] and the specific decomposition along Thm. 2.72 for BS have been shown very early [91, 264]. The value range of BS is $[0,2]$ while it is negatively oriented (see the interpretation of the decomposition terms). One problem of BS is that, as long as $\textbf{BS} > 0$ we do not know how miscalibration relates to improper discrimination and in general by the dependence on the prevalences of $\mathcal{T}$

the interpretation remains difficult. Nevertheless, a result by Gruber and Buettner [150] shows that the square root of BS, the so-called **Root Brier Score (RBS)** is a robust upper bound of the canonical calibration error. To circumvent the interpretability issues of BS the BSS provides a normalization of BS along a naive classifier (similar to NEC in Def. 2.27). This makes BSS positively oriented and a value of 0 points towards naive performance.

2.7 Deep Neural Networks

Deep Learning (DL) is a subdomain of ML. ML is the study of algorithms that improve automatically through experience [254] and itself a subdomain of AI. The problem to properly define AI is more of philosophical nature and left to the reader [332]. We will solely focus on the case of *supervised classification* in ML, that can be described as follows:

Definition 2.75. Let $\mathcal{T}$ be a task with C classes, $\mathcal{H}$ a subclass of $\Phi_{\mathcal{T}}$ (called the **hypothesis space**) and $\mathcal{L} : \Delta_{C-1} \times Y_{\mathcal{T}} \to \mathbb{R}_{\geq 0}$ a computable **loss function**. The goal of **supervised learning** is to solve the following optimization problem

$$\operatorname{argmin}_{\varphi \in \mathcal{H}} |\mathcal{T}|^{-1} \sum_{(x_i, y_i) \in \mathcal{T}} \mathcal{L}(\varphi(x_i), y_i).$$

The term following argmin is also referred to as the **empirical risk**. Note though that the definition of supervised learning is not standardized in literature and the transitions to other paradigms, e. g., self-supervised learning are rather blurred [143]. As indicated before, it is good practice to partition $\mathcal{T}$ into multiple subsets, commonly $\mathcal{T}_{\text{train}}$, $\mathcal{T}_{\text{val}}$, and $\mathcal{T}_{\text{test}}$. The *training split* $\mathcal{T}_{\text{train}}$ may be used in automatic hypothesis updates, i. e., iterative modification on the current model. The *validation split* $\mathcal{T}_{\text{val}}$ is – automatically or manually – used in monitoring to prevent a phenomenon called *overfitting*, the extraction of residual variation [14]. Lastly, the *test split* $\mathcal{T}_{\text{test}}$ is kept untouched during the development of a model and only used in the end to estimate the empirical risk on unseen samples. Keeping this in mind, we will always assume that the task $\mathcal{T}$ used by a performance measure μ to assess a model φ has not been used during training or validation of φ.

Returning to the formulation of supervised learning, three questions may arise:

(I) How should one choose the loss function in order to determine a good model?

(II) How exactly should such the optimization be approached computationally?

(III) How should one design the hypothesis space to allow a good fit, while keeping the computational optimization feasible?

Of course these questions all have multiple possible answers and the research of about half a century will now be compressed into a few solutions within a couple of

paragraphs. We start with the first, question, which will turn out to be simple, because we already introduced NLL in Def. 2.73 as a strictly proper scoring rule, that through the decomposition given in Thm. 2.72 guarantees good discrimination as well as calibration. In practice there is also a weighted version we introduce here for later reference.

Definition 2.76. Let $\mathcal{T}$ be a task with C classes and $\{w_i\}_{i\leq C}$ a set of reals called the **class weights**, then the **weighted Cross Entropy** is the loss function given by

$$\mathcal{L}_{CE}(p, y) = -w_y \ln(p_y),$$

for $p \in \Delta_{C-1}$ and $y \leq C$.

The weights in Def. 2.76 allow us to adjust the focus of learning for the Stochastic Gradient Descent (SGD) and will be important in Sec. 6.1. We next turn to the second question: the *optimization* procedure.

Definition 2.77. Let φ_Θ be a model for task $\mathcal{T}$ with some parameters Θ and **initial values** Θ_0, $\{\gamma_i\}_{i\in\mathbb{N}}$ a sequence of positive reals called the **learning rates**, $\mathcal{L}$ a loss function and $\{\mathbb{B}_i\}_{i\in\mathbb{N}}$ a sequence of subsets from $\mathcal{T}$ (called the **(mini-)batches**). We define the sequence $\Theta_1, \Theta_2, \ldots$ iteratively via

$$\Theta_{i+1} := \Theta - \gamma_i \nabla_\Theta \sum_{(x_j,y_j)\in B_i} \mathcal{L}(\varphi_{\Theta_i}(x_i), y_i) = \Theta - \gamma_i \sum_{(x_j,y_j)\in B_i} \nabla_\Theta \mathcal{L}(\varphi_{\Theta_i}(x_i), y_i),$$

as long as $\mathcal{L}(\varphi_{\Theta_i}(x), y)$ is differentiable for all (x, y) in the batch $\mathbb{B}_i$. This process is called **Stochastic Gradient Descent (SGD)**.

SGD was already used by Rosenblatt [325] to train the first NN in the 1950s, while the gradient descent method was proposed already by Cauchy et al. [56] more than a century earlier. The popularization of backpropagation [330] took until the late 1980s and was tied to efficient calculations of the derivatives according to the chain rule. A lot of research went into appropriate choices of the ingredients for SGD and the 'convergence guarantees' under various conditions. Surprisingly, although there are no theoretical guarantees SGD often works very well on modern NNs and tasks despite them having a non-convex loss space. This is also due to enhancements like keeping some momentum [143], weight decay regularization or per-parameter learning rates as given in Adam [200]. The learning rates $\{\gamma_i\}_{i\in\mathbb{N}}$, the initialization Θ_0 and the sampling strategy to generate $\{\mathbb{B}_i\}_{i\in\mathbb{N}}$ are examples for **hyperparameters** – parameters of the overall learning procedure while not being part of the model itself. Finally, we want to discuss the hypothesis space and will introduce modern DNN.

Definition 2.78. A **feedforward neural network** is a model φ composed of finitely many **layers** (also **blocks** or **modules**) $\{f^{(i)}\}_{i\leq D}, D \in \mathbb{N}$: $\varphi = f^{(D)} \circ .... \circ f^{(1)}$. Each layer is a parametrized mapping $f_w^{(i)} : \mathbb{R}^{n_i} \to \mathbb{R}^{m_i}$ such that $m_i = n_{i+1}$ and the parameters w are called **weights**. The first layer $f^{(1)}$ is also called **input layer**, while the last layer $f^{(D)}$ is called **output layer**. All other layers are usually referred to as **hidden layers** and their number directly corresponds to the model **depth** D. In contrast, the **width** of a layer corresponds to its output dimension m_i and the maximum internal dimension $\max m_i$ is called the **model width**. The outputs of a layer are referred to as **features** or **logits**.

Such a NN is usually called 'deep', i. e., a DNN, if D is sufficiently large, although there is no common threshold which determines this. AlexNet [210], which was an essential milestone in image classification, consists of eight layers and is usually perceived as 'deep'. But there is some ambiguity in determining the depth of a NN, since sometimes multiple layers are grouped into 'blocks', which are by definition layers themselves. It becomes obvious, that so far we have been rather vague on the nature of the *layers* of a feedforward neural network, but we will give some important examples on the following pages. Noteworthy the design given by Def. 2.78 of a NN corresponds well to the needs given by the SGD methods as the derivatives for the weights may be derived by the chain rule. The naming of such models is inspired by biology: The features at a certain layer in the model translate to the activation level of neurons as found, for example, in the central nervous system of vertebrates. These signals are forwarded to the next 'layer' of neurons that determine their own activation based on the received signals. We continue with formally giving the first example of an 'artificial neural network'.

Definition 2.79. A **fully-connected layer** (also **dense layer** or **linear layer**) is a mapping $f : \mathbb{R}^n \to \mathbb{R}^m$ given by

$$f(x) := \alpha(\boldsymbol{W}x + b),$$

here $b \in \mathbb{R}^m$ is called the **bias**, $\boldsymbol{W}$ is an $m \times n$ matrix (confusingly also sometimes called **weights**, although the bias is also considered a learnable parameter) and $\alpha : \mathbb{R} \to \mathbb{R}$ (applied element-wise) is an **activation function**. A **Multilayer Perceptron (MLP)** is a feedforward neural network whose layers are all fully-connected.

As mentioned before the MLP was the earliest kind of NN [325].[17] Various *universal approximation theorems* (see for example [297]) guarantee under different constraints for the width respectively depths of neural networks, that for non-polynomial activation

[17] The presented MLP by Rosenblatt [325] had a single hidden layer which was initialized with random values and not trained though.

functions, (almost) any continuous function may be approximated through a MLP. Bound by this non-polynomial requirement early activation functions included the hyperbolic tangent tanh and the logistic function $(1 + e^{-x})^{-1}$, but the very simple **rectified linear unit (ReLU)** $\max(0, x)$ became popular in the 2010s [129][18] One of the problems with fully-connected layers is their *density*, as they have comparably many learnable parameters ($(n + 1) \cdot m$), which becomes infeasible for large inputs. A solution to this problem is given by the next kind of model.

Definition 2.80. A **convolutional layer** [143] is a mapping $f : \mathbb{R}^{h \times w \times c} \to \mathbb{R}^{h' \times w' \times c'}$, where given by

$$f(x) := \alpha(x \star \boldsymbol{K} + b),$$

where $\boldsymbol{K} \in \mathbb{R}^{h^\star \times w^\star \times c^\star}$ is called the **kernel**, $b \in \mathbb{R}^{h' \times w' \times c'}$ is again called bias and $\alpha : \mathbb{R} \to \mathbb{R}$ (applied element-wise) is an activation function. The operator $\star$ is called **convolution** and given by

$$(x \star \boldsymbol{K})_{ijk} := \sum_{m \leq h^\star} \sum_{n \leq w^\star} \sum_{o \leq h^\star} x_{i-m, j-n, k-o} \cdot \boldsymbol{K}_{mno}.$$

A NN that has at least one convolutional layer is called a **Convolutional Neural Network (CNN)**.[a]

[a] We omitted some details in the common convolution of NN, like 'stride', 'padding' and 'dilation' that have an influence on the shape of the output ($h' \times w' \times c'$).

The core idea of convolutional layers is to leverage the grid structure of images and make the layer 'equivariant' to translation. That means if g is some translation (moving the pixels of an image by a constant offset in each dimension) then $f \circ g = g \circ f$. The *kernel* behaves like a small sliding window on the image grid treating perceived inputs (e. g., a common range of $h \times w$ in kernel size is 3×3 up to 7×7) independent of the original position in the image. One could say 'a dog stays a dog, independently of whether it is placed in the lower right corner or in the upper left corner of an image'. Moreover, the convolutional operation may be realized with a matrix operation[19] rendering the overall layer very similar to the fully-connected one – and making it efficient for computation on modern GPUs. The main difference is that the matrix equivalent of the convolution has 'sparse weights' in the sense that many of them are 'tied together' – often referred to as 'shared parameters' [143]. The true number of learned parameters is thus determined by

[18] Obviously ReLU is not differentiable at $x = 0$, but choosing an arbitrary value (e. g., 0 or 1) for the derivative at this point works well in practice. Multiple *soft* (i. e., smoothed) variants like GELU [165] or Mish [253] have also been proposed to circumvent the non-differentiability.

[19] In practice most implementations use the related 'cross-correlation' operator instead, but call the layer 'convolutional' after all (e. g., *pytorch* [15].)

the kernel (plus the ones from the bias), which can be several orders of magnitude lower than the corresponding fully-connected layer. Although the 'receptive field' of a feature – the elements of the inputs that influenced its output – is drastically smaller compared to a fully connected layers, stacking multiple convolutional layers in a model increases the overall receptive field of an output logit. Further translational stabilization is usually achieved by a parameter-free **pooling layer** which aggregates small windows of an image, either by averaging or taking the maximum thus neglecting *smaller* activations, to increase robustness to noise [41].[20] Another noteworthy ingredient are **BatchNorm layers** [180], an alternative for **dropout layers** [366] to regularize the network as it achieves 'scale independence' of parameters and input[21], hence allowing for larger learning rates. These advances, joint by the progress in computational hardware allowed the successful training of models up to about 20 to 30 layers, but surprisingly adding more layers lead to a *degradation* of model performance. He et al. [161] noted that in theory adding *identity layers* should allow deeper models, but the identity function is hard to learn given the matrix multiplication and non-linear activations from the usual layers above. Instead, it is easier for a layer f to learn that the 'residual', i. e., the difference in input and output should be zero. So if g is the desired function to learn, the layer only learns $f(x) = g(x) - x$ and afterwards the output is recast via $f(x) + x$.

Definition 2.81. A **residual layer** (also **skip connection** or **residual block**) [161] is a mapping $f : \mathbb{R}^n \to \mathbb{R}^m$, encapsulating an internal block $\tilde{f} : \mathbb{R}^n \to \mathbb{R}^m$ via

$$f(x) := \tilde{f}(x) + \boldsymbol{W}x,$$

and $\boldsymbol{W}$ being an $m \times n$ **projection matrix**. If $n = m$ the default is to use the identity matrix and the residual layer adds no parameters to the model, otherwise the elements of $\boldsymbol{W}$ are learned. A NN that has at least one residual layer is called a **Residual Network**.

Surprisingly this 'simple' ingredient allowed the authors to successfully train models with hundreds of layers and the family of resulting **ResNets** [161] to stay competitive for many years [422]. One of the important ingredients for this is according to Wightman et al. [422] an appropriate sampling strategy (see Def. 2.77), for which we want to introduce some general terminology.

Definition 2.82. Recall the class of all images is denoted by $\mathcal{X}$. A computable mapping $t : \mathcal{X} \to \mathcal{X}$ is called an **image transformation**. An **augmentation policy** is an algorithm that samples finite sequences of image transformation.

[20] And they can also be leveraged to apply models on varying image sizes.

[21] That means if $\boldsymbol{W}x$ is a computation within a layer, applying BatchNorm **BN** achieves $\mathbf{BN}(\boldsymbol{W}x) = \mathbf{BN}((a\boldsymbol{W})x)$ for scalars a, while not affecting the backpropagation.

The goal of data augmentation is to 'synthetically' increase the number of samples in a task $\mathcal{T}$. Obviously special care has to be taken, such that the sampled transformation $\{t_i\}_{i\leq n}$ ensure that the *augmented* sample $(t_n(...(t_1(x))), y)$ realistically matches to the underlying distribution $p_{\mathcal{T}}(X, Y)$ of the original task[22]. A simple example of a transformation is **horizontal flipping**, mirroring the image along a vertical central line. Using this very transformation might produce realistic samples for, e. g., an image of a horse in the CIFAR10 dataset [208], but would terribly hurt the model in telling apart the letters p and q in the EMNIST dataset [74], as it would create a 'corrupted' sample with a wrong label. A special sequence of transformations is usually leveraged to *unify* the images of a dataset.

Definition 2.83. The **preprocessing** of a task $\mathcal{T}$ refers to applying a finite sequence of deterministic transformations $\hat{t} := t_n \circ ... \circ t_1$ on $\mathcal{T}$ such that a new dataset $\hat{\mathcal{T}} := \{(\hat{t}(x_i), y_i) : (x_i, y_i) \in \mathcal{T}\}$ is produced.

Preprocessing is used for various purposes, e. g., rescaling, denoising, anonymization, etc. The resulting task shares the same number of classes as well as number of samples. Usually the convention is to understand preprocessing as a separate preliminary step before the actual model is applied[23], but formally we can attach a *preprocessing layer* to any NN. Such a layer has no trainable parameters, but the choice of the preprocessing transformations is another example of a *hyperparameter*. We finally formalize this idea with the last definition of this section, that naturally brings us to the final section of this chapter.

Definition 2.84. A **Trainer** (also **Learner**) $\mathfrak{T}$ is an algorithm that tries to solve a supervised learning problem. Formally it takes a task $\mathcal{T}$ and some **hyperparameters** ω to compute a model $\varphi = \mathfrak{T}(\mathcal{T}, \omega)$ that sufficiently solves the supervised learning problem described by ω and φ. The entirety of hyperparameters ω is also sometimes referred to as the **training pipeline** and the process of executing the trainer is called **model training**.

By design the hyperparameters ω must comprise the ingredients given in Def. 2.75, i. e., a hypothesis space (often referred to as a **neural architecture**) and a loss function, furthermore it includes in our case the ingredients from Def. 2.77, i. e., a sampling strategy (usually comprising also an augmentation policy) and the learning rates (usually through an algorithm called **learning rate scheduler**). But as we noted the preprocessing is

[22] Interestingly also out-of-distribution samples might help. Interesting examples are CutMix [437] and MixUp [442], which may act as regularization and reduce overconfidence [52].

[23] Although often computed separately and only once for efficiency reasons, preprocessing is usually an implicit part of the model, as it poses certain formatting constraints to the input samples. The library of *nnU-Net* [182], for example, ships with a dedicated preprocessing module.

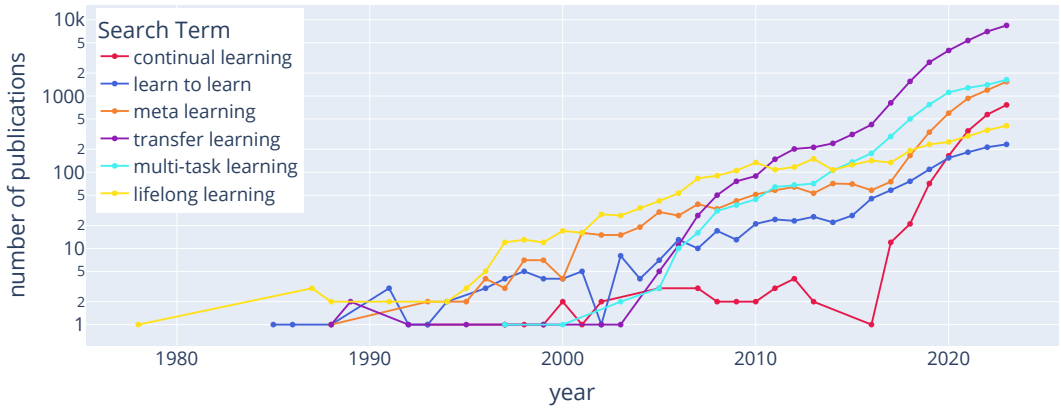


Figure 2.16: Publication counts for meta learning and related terms. Y-axis is log-scaled. Data was extracted through `scopus.com` and publications were restricted to the domains of computer science and math and restricted up to year 2023. Accessed on 11/08/24.

also a hyperparameter. The choice of ω is crucial for training a successful model φ [305]. To measure the success usually a separate $\mathcal{T}_{\text{test}}$ is kept aside during training and used to compute the fitness of φ via some performance measure μ (see Def. 2.10). The process of selecting ω is called **hyperparameter optimization** and various techniques like grid search, random search [28], Bayesian optimization [349, 362] as well as human driven ('manual') search – often in conjunction with cross-validation [31] – are commonly applied to determine 'the best' ω. Even if such hyperparameter search methods are employed, selecting the most important hyperparameters and defining an appropriate search space is difficult and results in multiple iterations that lead to a slow and resource intense workflow.

2.8 Lifelong Learning

We are finally prepared to formally differentiate the various terms surrounding 'Lifelong Learning' from Chap. 1. The shared difference compared to our definition of a supervised classification task (see Def. 2.75) is that there are now other tasks to be learned 'in parallel'. Before we dive into the details, it is worth noting that the terms and clear definitions we will use in the following are not uniformly agreed and used in the literature – especially the distinction between *Lifelong Learning* (Def. 2.90) and *Continual Learning* (Def. 2.88). We will mention most of the ambiguities and point towards comprehensive review papers whenever possible. As can be seen in Fig. 2.16 it was the AI hype of the 2010s that pushed all of these terms above the 100 publications per year mark. The general setting may thus be view as a comparatively young research domain and no long-term, stable nomenclature has yet emerged to fully resolve past entanglements. For the remainder of

this section we will gradually build up our way in the definition of Lifelong Learning and end with a detailed comparison of each learning paradigm.

Before we start, we want to motivate the idea of 'combining' multiple tasks from a human perspective. Our brains – the learning model – do not treat each of the tasks we are faced in isolation, but rather decompose complex tasks (e. g., driving a car) into chunks that we can relate with previously acquired skills (e. g., traffic light detection, prediction of object motion, interpretation of symbols on the speedometer). In the specific case of image classification most of the readers would not need any training samples for a 'cat versus dog' classification task, as the concepts of 'cats' and 'dogs' have been previously acquired. The *Imagenet large scale visual recognition challenge* comprises 1000 sometimes rather fine-grained classes and requires some familiarization for human raters to reduce prediction error. Nevertheless the required 'training samples' are orders of magnitude lower compared to ML models learning from scratch [331]. It takes a lot of prior knowledge to interpret some imaging modalities in medicine. Detecting polyps from gastrointestinal endoscopy [38], identifying cancer among skin lesions in dermoscopy [328] or examining pathologies in chest X-rays [181] requires years of training for domain experts. But these tasks are also not learned in isolation, and instead accompanied by learning concepts of physiology and other related, though simpler tasks. We admit, that the mentioned 'concepts', their 'relatedness' and 'difficulty' are rather fuzzy in these explanations. The topic of 'task relatedness' will be of major interest in this thesis and treated in Chap. 5. We start with a basic paradigm of combining multiple tasks.

Definition 2.85. Let $m \in \mathbb{N}$ and $\{\mathcal{T}_i\}_{i\leq m}$ be m tasks, $\mathcal{H}$ a subclass of $\bigcap_{i\leq m} \Phi_{\mathcal{T}_i}$, $\{\mathcal{L}_i\}_{i\leq m}$ a sequence of computable loss function and $\{w_i\}_{i\leq m}$ a sequence of reals (called **task weights**). The task of (supervised) **Multitask Learning** [80, 435, 447] is to solve the following optimization problem

$$\operatorname{argmin}_{\varphi\in\mathcal{H}} \sum_{i\leq m} w_i |\mathcal{T}_i|^{-1} \sum_{(x_j,y_j)\in\mathcal{T}_i} \mathcal{L}_i(\varphi(x_j), y_j).$$

A strict reading of the definition requires that all tasks $\mathcal{T}_i$ must share the same number of classes, but we interpret the shared hypothesis space in the sense that the model knows for which tasks it should perform the prediction (it predicts a from an image x and the task index i, i. e., it should actually be denoted $\varphi(x, i)$). Typically, a multitask model comprises a 'shared' part, which we call **backbone**, and m task specific parts, which we call **task heads**. Each head performs predictions for one specific task, so the implicitly given task index i from the definition above is used to select the appropriate head and the corresponding prediction in order to compute the loss. This approach is named **hard parameter sharing** [329, 397], because the backbone parameters are shared across tasks. Conversely, in **soft parameter sharing** [329, 397] each task has its own 'full' model,

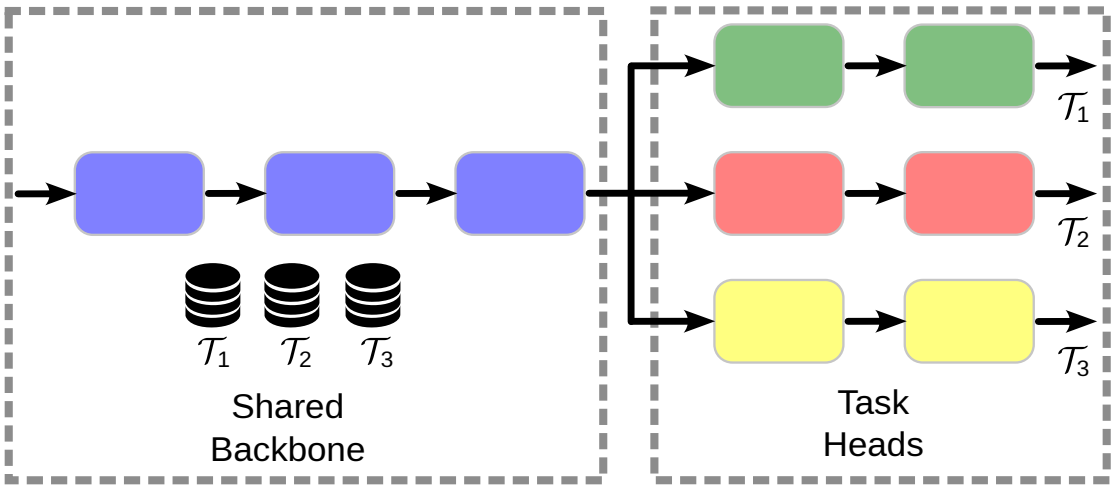


Figure 2.17: Hard parameter sharing for Multitask Learning. Images are first fed through a shared backbone until at some point features traverse individual task heads. Inspired by Vandenhende et al. [397].

while restrictions are made with the 'deviation' of the parameters in the backbones. A typical NN imposing hard parameter sharing comprises a sequence $f^{(b)} \circ .... \circ f^{(1)}$ for the backbone and m sequences $\{f_i^{(h_i)} \circ .... \circ f_i^{(1)}\}_{i \leq m}$. To predict for task i the full 'forward pass' comprises $\varphi = f_i^{(h_i)} \circ .... \circ f_i^{(1)} \circ f^{(b)} \circ .... \circ f^{(1)}$ (see Fig. 2.17). Usually the depths of the heads (h_i) are rather small compared to the depth of the backbone (b).

The task weights $\{w_i\}_{i \leq m}$ allow for a different focus on the tasks. Sometimes there is an outstanding 'pivot task' that should be solved, while the final application does not require good performance on the other 'auxiliary tasks'. Then the weights – though usually being all positive during training – of these tasks might be set to zero during the evaluation of the model. The other common perspective cares equally about the performance on all tasks and would set $w_i = w_j \forall i, j \leq m$. In both cases the intention is to use the training signal by the other tasks learned in conjunction as a regularizer for the learning of each individual task. This inductive bias is hoped to lead to models that generalize better [329]. But great care has to be taken to ensure these tasks will not 'interfere' with each other and lead to a suboptimal overall performance (compared to solving each classification task separately). Selecting appropriate tasks [367] and dedicated learning strategies for them [435] are ongoing research questions. The 'transfer of knowledge' between tasks is rather implicit in the formulation of Multitask Learning. An *explicit* formulation of knowledge transfer is given by our next definition.

Definition 2.86. Let $m \in \mathbb{N}$ and $\{\mathcal{S}_i\}_{i \leq m}$ be m tasks (called **source tasks**). Furthermore, there is a classification problem for task $\mathcal{T}$ (typically called the **target task**). **Transfer Learning** [278, 378, 415, 451] describes the solving of the

classification problem for $\mathcal{T}$, while making use of (any knowledge derived from) $\{\mathcal{S}_i\}_{i\leq m}$.

Interestingly the definition of Transfer Learning has been rather stable since the early survey from Pan et al. [278]. The only difference is that more recent work allows multiple source 'domains', while originally $m = 1$ [451]. The latter is still the most common case and transfer learning – if not distinctly called 'multi-source' – by default refers to that scenario. A special subcase of Transfer Learning is the concept of **Domain Adaption** [411] that assumes that (the majority of) the data given from the targets task is 'unlabeled', i. e., while images are given, no labels are provided to the learner. The definition explicitly mentions the transfer of knowledge, but keeps the kind of knowledge open. This allows Multitask Learning and Transfer Learning to overlap: A Multitask Learning problem that uses hard parameter sharing and focuses on a pivot task is one example how to perform Transfer Learning. Conceptually closely related, the most prominent and widespread technique of Transfer Learning replaces the simultaneous learning by a sequential one.

Definition 2.87. Assume $\mathcal{S}$ and $\mathcal{T}$ are tasks and a DNN $\varphi = f^{(D)} \circ .... \circ f^{(1)}$ has been trained to solve $\mathcal{S}$. Let $1 < d < D$, then solving the supervised classification task for $\mathcal{T}$ where the d initial layers of all models in the hypothesis space match the ones from model φ and the weights of the SGD are initialized with the ones from φ, is called **fine-tuning** of φ to $\mathcal{T}$.

Typically, d is close to D so a big part of the model (also called **backbone** in this case) is transferred. Although fine-tuning is only one form of Transfer Learning, the two terms are sometimes used interchangeable [199]. We will give more details on the usage and success of fine-tuning in Sec. 3.2 and continue with defining further learning paradigms for now – the sequential learning nature of fine-tuning directly leads to the next one.

Definition 2.88. Let $m \in \mathbb{N}$ and $\{\mathcal{T}_i\}_{i\leq m}$ be a sequence of m tasks, $\mathcal{H}$ a subclass of $\bigcap_{i\leq m} \Phi_{\mathcal{T}_i}$, $\{\mathcal{L}_i\}_{i\leq m}$ a sequence of computable loss functions. The task of (supervised) **Continual Learning** [88, 410] (also (**(Task) Incremental Learning**)) is defined across multiple time steps $t \leq m$. At each step t the system is only allowed access to task $\mathcal{T}_t$ for model training and solve the current overall optimization problem

$$\text{argmin}_{\varphi\in\mathcal{H}} \sum_{i\leq t} |\mathcal{T}_i|^{-1} \sum_{(x_j,y_j)\in\mathcal{T}_i} \mathcal{L}_i(\varphi(x_j), y_j).$$

It is assumed that due to hardware limitations the system is not (or only in a very restrictive sense) allowed to inspect samples from previous tasks. Thus while it may be

easy to minimize the empirical risk for the current task the main challenge is to further reduce, or at least not increase the empirical risk for previous tasks. The straightforward combination of hard parameter sharing and fine-tuning, which adds a task specific head for each new task and finetunes the head with the backbone from the previous iteration, will suffer from a phenomenon called **catastrophic forgetting** [201]. Fine-tuning a DNN will in most cases reduce the performance of the model on any previous source task. Related to Continual Learning are also **Curriculum Learning** [413], which aims to optimize the order (and composition) of (sub)tasks to be learned in order to solve the supervised problem of a given target task, and **Online Learning** [350], which assumes that the data of each iteration is from the same distribution, or phrased differently all the $\{\mathcal{T}_i\}_{i\leq m}$ are subsets of a larger overall target task.

Definition 2.89. Let $m \in \mathbb{N}$ and $\{\mathcal{T}_i\}_{i\leq m}$ be m tasks and $\mathfrak{T}$ be a trainer. **Meta Learning** [172] refers to the problem of solving

$$\operatorname{argmin}_\omega \sum_{i\leq m} |\mathcal{T}_i|^{-1} \sum_{(x_j,y_j)\in\mathcal{T}_i} \mathcal{L}_i(\mathfrak{T}(\mathcal{T}_i,\omega)(x_j), y_j).$$

To measure the performance of a *Meta Learner* the proposed solution $\omega^\star$ is usually applied to previously unseen (target) tasks, in contrast to the standard classification problem, Multitask Learning and Transfer Learning, where usually held out samples of the same task are used for evaluation. The trained model $\mathfrak{T}(\mathcal{T}, \omega^\star)$ for such a target task $\mathcal{T}$ is then evaluated again on its held-out test set. In that sense Meta Learning shares the idea of Continual Learning that 'new' tasks show up and the experience gained from previous tasks needs to be leveraged in order to solve it. In contrast to Continual Learning (and Multitask Learning) it is though not a *single model* that needs to solve all tasks. It is the learning strategy itself, in form of the hyperparameters that is optimized through a *Meta Learner*. Specifically the invention of Model-Agnostic Meta Learning (MAML) by Finn et al. [117] sparked a lot of interest in Meta Learning and shaped the focus on learning a good weight initialization. This was particularly in the context of **Few-Shot Learning** [361], the special case of Meta Learning that evaluates on tasks with very few samples per class (mostly 5 or 10). An alternative term used for the setting above is also **AutoML** [177], that stems from the desire to ease the process of performing model training. While some use the terms interchangeable [398] others note that AutoML may also comprise 'data cleaning' which is outside the scope of Meta Learning [172]. It may be once more due to the impact of MAML that 'bi-level optimization' is sometimes perceived as a key ingredient of Meta Learning and, e. g., classic hyperparameter optimization techniques like random search and Bayesian optimization are not considered Meta Learning by some researchers [172]. The definition given for Meta Learning limits the information flow from the source tasks to a target task to the hyperparameters ω – a rather abstract form and limitation that is not unambiguous in literature. Envisioning a 'true' *Meta Learner*

surely involves task dependent information flow, as well as backward flow from newly observed tasks to previous ones. We capture this intuition in the long promised definition of Lifelong Learning.

Definition 2.90. Let $\{\mathcal{T}_i\}_{i\in\mathbb{N}}$ be a sequence of tasks, $\{\mathcal{H}_i \subseteq \Phi_{\mathcal{T}_i}\}_{i\in\mathbb{N}}$ a sequence of hypothesis spaces and $\{\mathcal{L}_i\}_{i\in\mathbb{N}}$ a sequence of computable loss functions. The task of (supervised) **Lifelong Learning** [60, 283] is defined across multiple time steps $t \in \mathbb{N}$. At each step t the system faces a new task $\mathcal{T}_t$ for model training, gets feedback from the open environment (e. g., on the behavior of previously deployed models $\{\varphi_i\}_{i<t}$) and is obliged to solve (respectively improve upon) all the current overall optimization problems

$$\text{argmin}_{\varphi_i\in\mathcal{H}_i} |\mathcal{T}_i|^{-1} \sum_{(x_j,y_j)\in\mathcal{T}_i} \mathcal{L}_i(\varphi_i(x_j), y_j)$$

for $i \leq t$.

Chen et al. [60] elaborate extensively about the requirements for such a system. They specifically require the system to comprise a *Knowledge Base*, which accumulates the insights gained over the lifetime of the running system. In contrast to all previous definitions Lifelong Learning *interacts* with the environment. Two kinds of interactions – one with human experts and one with previously unseen deployment data – will be the focus of Chap. 4 and Chap. 6. Lifelong Learning builds upon an intrinsic temporal perspective of knowledge and tasks – closely to Continual Learning and implicitly also Meta Learning. In contrast, it is though not limited to keep a *single* model or restrict knowledge transfer uni-directional. It focuses on multiple tasks at once as Multitask Learning but less static and more individually per task. A successful Lifelong Learning system *must* use Transfer Learning to inform the model training with the experience from the *Knowledge Base*. We can directly derive that the individual learning per task is a necessary baseline to compare any Lifelong Learner against. More details on evaluation strategies are given in Chap. 4, but for now we note that the system must be tested with various combinations of previous source tasks and target tasks, as well as a multitude of environmental feedback.

3 State-of-the-art Pipelines for Image Classification

After presenting the basics of biomedical image classification in the previous chapter, this chapter focuses on current practices, published work, and gaps in the literature with respect to our three research questions from Sec. 1.2. Here, Sec. 3.1 presents related work on validation practices of predictive models in medical image classification. Next, Sec. 3.2 compares existing approaches for training DNN in sparse data settings. Then, Sec. 3.3 gives an overview of the literature on distribution shifts during model deployment. We conclude the chapter with Sec. 3.4, which summarizes the previous sections, compares the state of the art with our research questions, and highlights gaps in the literature.

3.1 Model validation

The validation of ML models for medical image classification presents unique challenges that require careful consideration of performance measures and evaluation protocols. This section reviews current practices and identifies limitations in model validation approaches.

Challenges in performance measure selection

Selecting appropriate performance measures is critical for meaningful model evaluation in biomedical imaging. Specifically, reporting every possible metric can overwhelm analysis and fail to address application-specific needs [408]. Simultaneously, choosing performance measures based solely on popularity has been shown to be misleading across multiple medical imaging applications [145, 203, 240, 275, 321, 392]. Thomas et al. [384] caution against the narrow focus on single target measures, invoking *Goodhart's law*[1] to highlight how this approach can lead to skewed model development and suboptimal research resource allocation. Instead, comprehensive evaluation requires diverse metrics to assess accuracy, robustness, and generalizability, while benchmarking

[1] We provide it here in a simplified version: "*When a measure becomes a target, it ceases to be a good measure.*" [370]

against solid baselines and accounting for data uncertainty. Khan et al. [196] emphasize that clinical adoption necessitates early involvement of diverse stakeholders to ensure seamless integration into existing clinical workflows and practices. This collaborative approach should extend to metric selection, involving domain experts (e. g., physicians, ML researchers, biostatisticians) and stakeholders (e. g., companies, insurers, regulators) to ensure alignment with real-world medical applications and enhance evaluation credibility [384].

Comparative literature on performance measures

Several studies have assessed comprehensive suites of performance measures for classification tasks in general or specifically medical imaging. Sokolova et al. [363] examined multiclass counting metrics for various 'invariances' – changes in the confusion matrix that do not alter metric values. By analyzing whether performance measures are affected by these invariances, they identified similarities between metrics and discussed use cases where different invariances are desirable. Steyerberg et al. [368] reviewed newly proposed performance measures, contextualizing them with traditional metrics and emphasizing the importance of complementary model assessment for both discrimination and calibration. Taha et al. [377] unified definitions for 20 performance measures in semantic segmentation for 3D medical image analysis, some of which are applicable to 2D image classification. They focused on theoretical categorization of metrics, extending definitions for probabilistic annotations and providing efficient reference implementations. Their protocol for recommending performance measures via specific combinations of metric properties, data properties, and algorithm requirements is particularly valuable. Hossin et al. [173] discussed various classification metrics in the context of 'Prototype Selection', focusing on model generalizability and important factors to consider when designing new performance measures. Grandini et al. [147] analyzed multiclass counting metrics, highlighting their suitability for different use cases in model development while discussing both advantages and disadvantages. More recently, Varoquaux et al. [400] presented and discussed a set of performance measures for binary and multiclass image classification in medical imaging, particularly addressing the prevalence dependency of some metrics while advocating for calibration measures.

Analyses of individual metrics

Some studies have critically examined particular performance measures. Christen et al. [69] reviewed the F1, discussing its shortcomings for many classification use cases. Similarly, Sebastiani [346] demonstrated that F1 fails to meet intuitive expectations when analyzed against a set of desirable metric properties. Ferrer [113] conducted an in-depth analysis on the generalizability of EC, comparing it with numerous other metrics. In a series of publications, Giuseppe Jurman and Davide Chicco have advocated for the MCC through direct comparisons with various other performance measures [63–68, 188]. How-

ever, the unrestricted appropriateness of MCC has also faced counterarguments [450].

Universal recommendations

Various recommendations for best practices in benchmarking AI models for medical image analysis have been formulated [39, 260, 271, 284, 300, 399]. These guidelines address appropriate data handling, statistical meaningfulness, the preference of complementary performance measures, reporting standards, and baseline selection in the general sense. However, given the variety of applications in medical imaging (see Sec. 2.2), these recommendations lack specific instructions or criteria for use-case-specific metric selection.

Evidence for inappropriate validation

The guidelines also contrast the empirical insights by Maier-Hein et al. [240], who assessed 150 biomedical image analysis competitions and revealed a concentration on few performance measures and other prevalent flaws in determining 'best' algorithms. Evidence for incomplete reporting of validation results is also given by Hicks et al. [167], who recalculate and interpret non-reported performance measures for five exemplary studies in gastroenterology. In a systematic literature review O'Shea et al. [272] found only 36% out of 186 studies in radiological cancer diagnosis provided a rationale for their choice of performance measure(s). An even larger meta analysis of 503 studies on diagnostic accuracy across multiple imaging domains and applications by Aggarwal et al. [8] concludes "poor design, conduct and reporting of studies".

Implementation and aggregation

Beyond metric selection, the implementation and aggregation of performance measures must be robust to variations in models or data, including handling missing values, sample inter-dependencies, and algorithmic non-determinism [420]. Although conveniently available through standardized interfaces in multiple libraries, (e. g., *torchmetrics* [94] or *scikit-learn* [292]) many metrics require configuration, e. g., a cutoff (see Def. 2.9). Additionally, beyond the aspect of purely ranking models, human perception and interpretation of performance measures play critical roles, requiring a balance between comprehensive evaluation and interpretability to facilitate model adoption. While multiple metrics are necessary to characterize performance comprehensively, their presentation must remain accessible to stakeholders with varying levels of technical expertise. This balance between thoroughness and clarity remains a significant challenge in the field.

3.2 Training in sparse data settings

Medical image analysis faces persistent challenges due to data scarcity, which stems from privacy concerns, expensive annotation processes, and the rarity of certain conditions (see Sec. 2.1). Addressing these limitations has given rise to several methodological approaches that aim to maximize model performance despite limited labeled data availability.

Overview of approaches

Crowd sourcing [236] leverages collective human intelligence by distributing annotation tasks across numerous individuals, potentially reducing costs and accelerating data labeling. However, in medical contexts, this approach often requires specialized domain knowledge, raising questions about annotation quality and consistency when performed by non-experts [309].

Data augmentation artificially expands limited datasets through transformations (see Def. 2.82) that preserve class identity while introducing variation. More advanced approaches include generative models (e. g., Generative Adversarial Networks (GANs) [144]) that create realistic synthetic medical images, helping models generalize better across patient populations and acquisition settings.

Active Learning [235] strategically selects the most informative samples for annotation, optimizing the learning process with minimal labeled data. This approach iteratively identifies uncertain or boundary cases where model predictions lack confidence, prioritizing them for expert review. By focusing annotation efforts on the most valuable samples, active learning can significantly reduce annotation costs while maintaining model performance.

Unsupervised Learning [314] methods extract patterns and representations from data without relying on explicit labels. In medical imaging, techniques such as autoencoders and contrastive learning can leverage large unlabeled datasets to learn meaningful representations, which can then be fine-tuned for specific diagnostic tasks with limited labeled data.

Multitask Learning (see Def. 2.85) leverages shared representations to simultaneously address multiple related objectives, potentially improving performance across all tasks compared to individual modeling. By combining data from multiple tasks, the system can develop more robust and generalizable representations, partially mitigating data scarcity for individual tasks [447].

Federated Learning [225] enables collaborative model training across multiple institutions without centralizing sensitive patient data. This approach distributes the training process, allowing models to learn from diverse tasks while maintaining data privacy and regulatory compliance. In medical imaging, federated learning addresses ethical and legal constraints around data sharing, though it introduces challenges related to harmonizing heterogeneous data distributions, establishing trust between participating institutions, and managing computational synchronization across sites [319].

Self-configuring methods, such as *Automated Machine Learning* [162] and *nnU-Net* [182], automate parts of the model training pipeline for specific applications. These frameworks analyze dataset properties to automatically determine, e. g., optimal preprocessing steps, network configurations, and training strategies, without requiring manual tuning. In biomedical image segmentation, *nnU-Net* has demonstrated state-of-the-art performance across diverse tasks by applying a consistent set of heuristics, though its application remains limited to segmentation problems and relies on predefined rules.

Task Transferability Estimation

The concept of Transfer Learning dates back to the 1970s [42] and involves any knowledge transfer from a source task to a target task (see Def. 2.86), with its most prominent application being pretraining on large-scale datasets [310, 451]. While utilizing off-the-shelf pretrained models is a common practice to speed up model convergence (see Def. 2.87), the vast number of available architectures, pretraining schemes, and datasets, makes selecting the most suitable option a time-consuming process. This premise also holds for the nascent field of FMs [37]. A key challenge in medical imaging Transfer Learning is the domain gap between general computer vision datasets and medical images. It has been shown that simply relying on benchmarks for pretrained models, such as ImageNet [93], does not translate well to the medical domain [310].

These limitations of generic Transfer Learning approaches have fueled research in *Task Transferability Estimation*, which aims to quantify the potential of knowledge transfer between tasks [5]. While the research problem has been known for decades [27] and many methods have been proposed, challenges with respect to data privacy considerations, robustness in realistic heterogeneous data settings, and the avoidance of negative transfer remain [98, 445, 451].

Existing Task Transferability Estimation methods, which attempt to assess model suitability for a target task, often have only been applied in unrealistically homogeneous settings [4, 96, 233, 269], lack large-scale validation (12 tasks in [257], 9 tasks in [158], 8 tasks in [381], 7 tasks in [313], 5 tasks in [12]), or are incompatible with data privacy requirements [433]. Others are not suited for scalability, either because of computational complexity [62, 115, 270, 294, 381, 388][2] or the assumption that tasks share underlying images [438].

3.3 Dataset shifts in algorithm deployment

Dataset shifts, where training and deployment data distributions differ, pose significant challenges in ML. In medical image analysis, these shifts are particularly critical due to

[2] A scheme that is targeted by many publications concerns the search for a matching pretrained model from a 'model hub' and involves infering predictions on the target task for *every* model in the model hub. Because of the large file sizes of model weights, we consider this approach not scalable.

their impact on diagnostic reliability.

Fundamental concepts

The problem of dataset shifts in ML was early summarized by Quinonero-Candela et al. [306], who categorized six groups of shifts. Among these, two are particularly relevant: *covariate shifts*, where the distribution of images $p(X)$ changes but $p(Y|X)$ remains stable, and *prior probability shifts*[3], where the label distribution $p(Y)$ changes but $p(X|Y)$ stays consistent. Given a causal model in the form of $p(X|Y)P(Y)$ they present Bayes rule as a potential solution when target prevalences $p(Y)$ are known.

The inconsistent terminology in dataset shift research was addressed by Moreno-Torres et al. [261], who clarified definitions and discussed not only covariate and prior probability shifts but also *concept shift*, where $p(X|Y)$ respectively $p(Y|X)$ change, while $p(Y)$ respectively $p(X)$ remain stable. These concept shifts present particular challenges due to the variety of potential causes, as illustrated by Gama et al. [122]. Despite these complexities, modeling these causal relationships has gained popularity in healthcare applications [334].

Dataset shifts in healthcare applications

Zhang et al. [440] provide compelling evidence for the importance of addressing dataset shifts when deploying ML systems in healthcare. They identify several causes including "institutional differences (such as local clinical practices, or different instruments and data-collection workflows), epidemiological shifts, temporal shifts (for example, changes in physician and patient behaviours over time) and differences in patient demographics (such as race, gender and age)". Their work references examples from literature demonstrating performance deterioration after deployment under these shifts. For instance, Zech et al. [439] investigated confounding factors of cross-institutional prevalence shifts in pneumonia detection in X-ray images, finding reduced discriminative performance when models were applied across institutions.

Subbaswamy et al. [372] advocate focusing on understanding the data generation process in medical AI. Castro et al. [54] translate this concept more rigorously into the language of causality for medical imaging, proposing a generic causal diagram for imaging workflows (see Fig. 2.1). Their categorization of shifts focuses on medical imaging processes, preferring the term *prevalence shifts*, which we follow in our work. Moreover, they provide specific recommendations for each shift type, suggesting Bayes rule or sample reweighting (see Def. 6.7) strategies for prevalence shift situations. Arjovsky et al. [18] raise an important point that many approaches for handling covariate shift become infeasible when prevalence shifts occur simultaneously, highlighting the complexity of real-world deployment scenarios.

[3] A synonym for prevalence shifts. For a full list of synonyms used in literature see Def. 6.4.

Analysis of prevalence shifts

One of the earliest works on handling prevalence shifts is by Saerens et al. [333]. They motivated sample reweighting and proposed an algorithm to estimate unknown prevalences in a new environment – a problem later termed *quantification* [26] (see Def. 6.6).

Zhang et al. [443] assessed this problem more thoroughly, detailing the required assumptions for their proposed quantification method and demonstrating how sample reweighting can address prevalence shifts. Similar theoretical derivations were later provided by Lipton et al. [230] for their approach of estimating prevalences and adjusting NNs post-hoc. Although they validated their approach on image classification, their MLP and the datasets used (MNIST and CIFAR10) are not representative of medical image analysis.

Dockès et al. [99] provide an overview of when and how dataset shifts affect predictions in biomedical tasks, discussing both detection and correction techniques. For prevalence shifts specifically, they suggest updating probabilities according to Bayes rule but do not address prevalence estimation.

Ovadia et al. [277] experimentally demonstrated that temperature scaling (see Def. 2.67) can mitigate increased miscalibration for covariate shifts. However, Alexandari et al. [11] showed that temperature scaling alone is insufficient in the context of prevalence shifts and proposed *Bias-Corrected Temperature Scaling* along with a method for quantification. Their evaluation included three imaging tasks (one medical), though they only reported miscalibration on the non-shifted test set. This solution was later placed within a theoretical framework and independently verified by Garg et al. [124].

Recent trends

In a systematic benchmark of 24 quantification methods, Schumacher et al. [345] assessed performance across 40 tabular datasets, providing recommendations while acknowledging the dynamic nature of the field. Indeed, since their work, additional quantification methods have been proposed, such as those by Moreo et al. [263], demonstrating the continued evolution of approaches to address prevalence shifts in deployment environments.

3.4 Summary of open challenges

We briefly summarize the previous sections, before drawing conclusions and putting our research questions into perspective with literature.

(RQ1) Validation of image classifiers

As revealed by our review on prevailing practices and existing research on performance measure selection (see Sec. 3.1) there is a strong need for a systematic decision guide for performance measure selection. Current approaches frequently rely on popular but

potentially misaligned metrics, leading to disconnects between reported performance and clinical utility. **(RQ1)** addresses this gap by seeking methodologies to systematically translate clinical objectives into appropriate validation metrics. This requires not only understanding the mathematical properties of various metrics but also establishing processes for capturing and incorporating domain expertise into validation frameworks. A driving question is which properties of a dataset and clinical application are most relevant for decision-making about performance measures.

(RQ2) Enabling knowledge transfer

Measuring task similarity represents a fundamental challenge in establishing effective knowledge transfer between medical imaging applications. Unlike traditional Transfer Learning methods that rely on direct access to source data, privacy-preserving task similarity should aim to capture essential characteristics of learning tasks without requiring raw data exchange. **(RQ2)** focuses on developing robust mechanisms for knowledge transfer that respect the unique constraints of biomedical applications. An ideal similarity measure would predict transfer performance accurately while remaining computationally efficient and respecting patient confidentiality requirements. Current approaches either lack validation in heterogeneous medical imaging contexts, fail to scale efficiently, or cannot operate within privacy constraints – limitations our research aims to address through a novel transferability estimation method.

(RQ3) Understanding prevalence shifts

The detection and compensation of prevalence shifts in deployment environments has a critical gap in current research. While theoretical frameworks exist for adjusting predictions under known prevalence changes, the interplay between prevalence quantification, re-calibration, and downstream decision rules represents a complex optimization problem that requires careful investigation. **(RQ3)** seeks to develop a systematic approach for maintaining model performance under changing class distributions, a common scenario in clinical practice. Moreover a large-scale evaluation with DNNs on applications of medical imaging is necessary. By addressing these challenges, we aim to enable robust model deployment across diverse healthcare settings with varying disease prevalences.

Towards Lifelong Learning systems

The translational struggles of current approaches in biomedical imaging highlight the need for systems capable of Lifelong Learning (see Def. 2.90) – continuously and automatically growing knowledge as more data and tasks are incorporated over time. Such systems would ideally combine the strengths of Transfer Learning, Federated Learning, and automated optimization while addressing their respective shortcomings. Recent work by Soltoggio et al. [364] presents a compelling vision for the future of AI where independent learning units share knowledge throughout their lifetimes, creating a "society of AI

systems" that collectively holds more knowledge than any single agent. This concept of *Shared Experience Lifelong Learning* is particularly relevant for medical applications, where the constant evolution of illnesses, pathogens, and diagnostic technologies limits the effectiveness of single-task models in isolation.

A truly effective Lifelong Learning system for medical imaging needs to address the three fundamental challenges identified in our research questions:

(i) Align technical optimization with clinical utility through systematic translation of clinical objectives into appropriate validation metrics.

(ii) Quantify relationships between tasks to maximize knowledge reuse without compromising patient privacy, with mechanisms that scale efficiently as the *Knowledge Base* grows over time.

(iii) Adapt to changing data distributions across clinical environments, particularly compensating for prevalence shifts over time.

These capabilities could lead to long-term scalability of AI which would reduce energy consumption and carbon footprint of systems [364], while simultaneously addressing the unique challenges of biomedical image classification.

Part III

Learning to learn

4 Application-specific Validation of Image Classification Algorithms

Disclosure

Parts of the results of this chapter have been published at the *Medical Imaging with Deep Learning (MIDL) conference* [316], the *Medical Imaging Meets NeurIPS workshop* [315] and in *Nature Methods* [238]. See App. A for full disclosure.

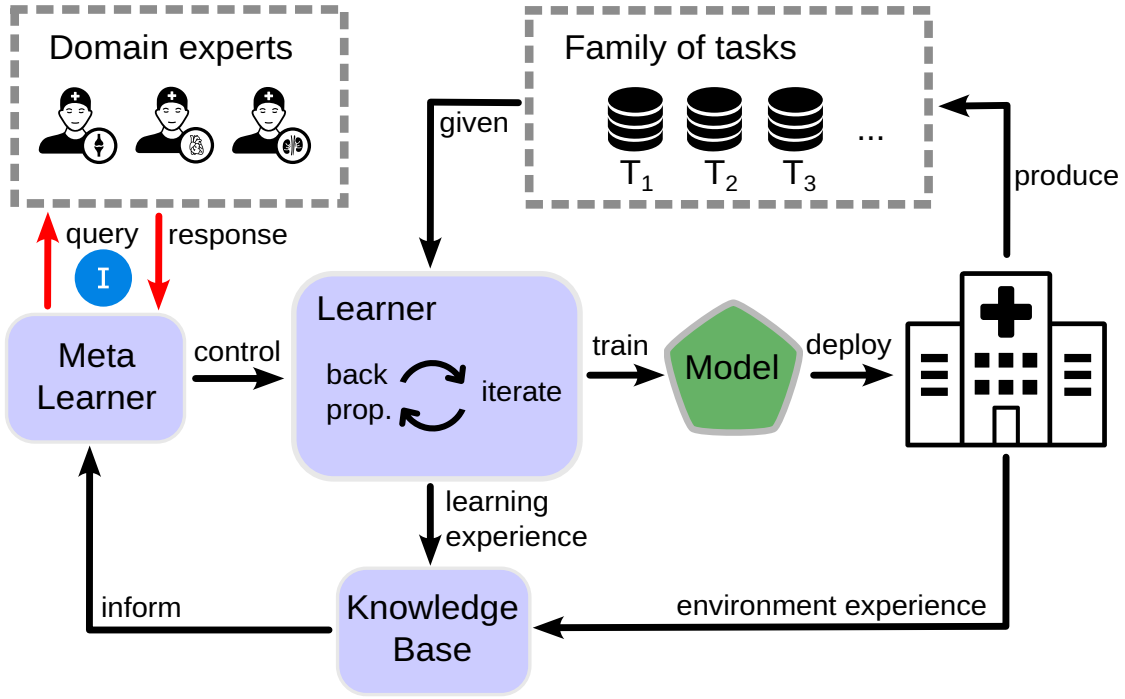


Figure 4.1: Reward-learning loop. Anchoring of this chapter in the overall Lifelong Learning system (see Fig. 1.1). Given a task, the *Meta Learner* interacts with domain experts to determine the optimal performance measure for the *Learner* to train a model. Loop is highlighted in red.

This chapter addresses the first research question of designing a **reward-learning loop** as part of the **Design** phase in the AI lifecycle (see Fig. 4.1):

Research Question 1

How can clinical objectives be systematically translated into appropriate AI model validation metrics?

In Sec. 2.1 we elaborated on the issues with *clinical alignment* of performance evaluation and the implications for AI models in medical imaging. Sec. 3.4 summarized the gap in the literature regarding comprehensive guidelines for metric selection. In this chapter we want to present our approach for such a *metric recommendation framework*. The vision in mind is a Lifelong Learning system that, when faced with a new task, iteratively poses a minimal set of questions to domain experts. The answers are used to compile a set of performance measures that are used to evaluate the task. Sec. 4.1 will present our international initiative *Metrics Reloaded* to provide such recommendations. Next, Sec. 4.2 will present the resulting metric selection process along with some concrete examples. Finally, Sec. 4.3 concludes the chapter with a discussion of our results.

4.1 Methods

This section presents our methodology for answering **(RQ1)** by following these steps: First, we provide background information on the international consortium formed to better understand and apply metrics in Sec. 4.1.1. After that, Sec. 4.1.2 summarizes all the ImLC metrics identified by the consortium. We also provide a detailed analysis of the relationships between the presented metrics, which helps to find complementary sets of metrics. Next, we present the main characteristics of tasks that determine appropriate metrics, summarized as the *problem fingerprint* in Sec. 4.1.3.

4.1.1 The *Metrics Reloaded* initiative

Our methodology for developing the *Metrics Reloaded* framework recommendations involved establishing an international consortium of experts and conducting a structured, multi-stage Delphi process. The Delphi technique, a systematic method for reaching consensus through iterative questionnaires and feedback [47], is particularly valuable in medical research for developing best practices when evidence is sparse or contradictory [267]. The expert panel initially consisted of 30 specialists from 25 institutions worldwide, drawn from three key research initiatives: the Biomedical Image Analysis Challenges (BIAS) initiative, the Medical Open Network for Artificial Intelligence (MONAI) Working Group for Evaluation, Reproducibility and Benchmarks, and the Medical Image Computing and Computer Assisted Interventions (MICCAI) Special Interest Group for Challenges (formerly MICCAI board working group). To enhance domain coverage and breadth of

expertise, the consortium has expanded over time to include additional experts in biology, medicine, epidemiology, biomedical image analysis, statistics, mathematics, computer science, and meta-research topics. The consortium was further strengthened by the inclusion of key figures from the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network [355].

The final expert consortium included a total of 73 researchers, composed of 73% male and 27% female, working at 65 different institutions. Of these experts, 52% were professors, followed by 37% who were postdoctoral researchers. The median h-index of the group was 34, with a mean of 27, a minimum of 6, and a maximum of 113. The median academic age of the researchers was 18 years, with a mean of 19, a minimum of 3, and a maximum of 42. The experts originated from 18 countries and covered 5 continents. The geographic distribution of the expert consortium showed a predominant European representation (73%), with significant contributions from Germany (35%) and the United Kingdom (12%). North American experts constituted 21% of the consortium, primarily from the United States (13%) and Canada (8%). The remaining participants came from South America (3%), Australia (2%), and Asia (1%). In terms of academic background, 66% of the experts had technical expertise, 7% were clinical experts, 3% had a biological background, and 24% had mixed expertise. Regarding the institutions, 88% of them provided staff data. Of these, 58% of the institutions had between 1000 and 10 000 employees, 25% were even larger, with 10 000 to 100 000 employees, and 16% had fewer than 1000 employees. Only 2% of the institutions exceeded 100 000 employees.

Within this international consortium, we established several distinct working groups:

(1) A three-member **core team**, took responsibility for the overall coordination of the project. This included managing the Delphi process, developing the framework structure based on expert input, designing and analyzing the surveys, and facilitating the workshops. To maintain objectivity, the core team generally abstained from participating in the voting procedures.

(2) An **extended core team** provided support to the primary coordinators, assisting with Delphi process administration, survey development, and workshop logistics.

(3) A series of **expert groups** that worked on specific aspects of the framework in between the surveys. Each expert group was led by up to two leads, who organized the communication within the expert group and between the expert group and the core team. The final expert groups were:

   (i) The **Image-level Classification (ImLC)** expert group, which focused on the ImLC task as described in this thesis.

   (ii) The **Semantic Segmentation (SemS)** expert group, focused on the task of SemS of images.

   (iii) The **Object Detection (ObD) and Instance Segmentation (InS)** expert group, focused on the two tasks of ObD and InS.

(iv) The **biomedical** expert group, consisting of clinicians and other domain experts, whose purpose was to ensure that the recommendation framework would meet application-specific needs and to identify scenarios for evaluating the framework.

(v) The **cross-topic** expert group, that addressed task-independent metric recommendations, such as metric aggregation.

(vi) The **calibration** expert group, established at a later date after calibration was identified as an additional important topic.

Overall, the process involved six distinct stages, including five workshops and nine surveys prior to the final Delphi consensus vote. After the surveys were completed, the core team carefully analyzed the results, discussed them with expert groups or individual experts as needed, and integrated feedback to iteratively refine the framework. The major stages of the compilation and consensus-building process are detailed in the following paragraphs.

1. Initialization The project began with a kick-off workshop aimed at defining the precise scope of the recommendation framework. We prepared for this workshop by conducting a preliminary survey focused on collecting relevant literature and documenting both theoretical and practical cases where metrics failed in classification, segmentation, and detection tasks. Following the workshop discussions, we implemented a three-survey sequence that yielded four key outcomes: (1) establishment of a unified terminology, (2) definition of inclusion criteria focusing on classification tasks at both the image/object and pixel levels, (3) development of a curated list of relevant metrics, the final version of which is presented in Sec. 4.1.2, and (4) generation of initial problem fingerprints, which were subsequently refined into the final fingerprints presented in Sec. 4.1.3.

2. Drafting The second Delphi workshop focused on forming specialized expert groups to lead distinct task forces. We formed five initial groups: three dedicated to specific problem categories (ImLC, SemS, and ObD/InS), supplemented by a biomedical expert group and a cross-topic expert group. Each problem category group was tasked with developing recommendations to address common evaluation pitfalls in their type of task [317]. The cross-topic group focused on broader metric-related challenges beyond metric selection, including aggregation methods, reporting standards, implementation considerations, statistical analyses, ranking procedures, and bias assessment. The biomedical expert group ensured the clinical relevance of the framework and identified key biomedical use cases. To facilitate the work of the task forces, we distributed group-specific surveys. We adopted a flexible approach, allowing each group autonomy in developing their recommendations without imposing methodological constraints. Preliminary results were presented by the expert group leaders to the core team during the third Delphi workshop.

3. Consolidation After the expert groups completed their initial draft recommendations, the *Metrics Reloaded* core team undertook a harmonization process, working closely with each group to integrate and standardize their contributions. The resulting decision trees encapsulating the core recommendations were presented and reviewed during the fourth Delphi workshop.

4. Revision The members of the consortium and their respective teams carried out an internal validation of the decision trees through practical application. The *Metrics Reloaded* core team, working closely with the expert groups, integrated the feedback collected through surveys into the framework. The comprehensive first draft was then presented and evaluated at the fifth Delphi workshop.

5. Crowdsourcing To ensure broad community input, we launched an extensive public feedback campaign following the release of the framework on arXiv [237]. The campaign combined social media outreach, targeted mailings, and an online survey. The survey, which received responses from 186 researchers, provided options for both quick feedback and detailed assessments of the framework's usefulness and comprehensiveness. Of the respondents, 82 provided written feedback, with 58 opting for detailed responses. This crowdsourcing effort yielded several significant outcomes: seven substantive contributors were invited to join the consortium, leading to the creation of a new expert group focused on calibration recommendations. Community feedback also guided the development of metric cheat sheets and the implementation of the web toolkit [1]. The survey responses informed our selection of biomedical use cases and led to enhancements to the framework, including the addition of new classification metrics such as the EC. The revised framework, incorporating these community-driven improvements, underwent a final consortium review through another survey, from which the core team compiled the final recommendations for Delphi-based consensus building.

6. Consensus The finalization of the framework involved an accelerated Delphi process to reach consensus on its ten core components. The calibration recommendations underwent two rounds of revisions in response to consortium feedback, ultimately achieving strong support with only a single dissenting vote. The remaining nine sub-processes demonstrated robust consensus in their first round, with disagreement rates ranging from 0% to 7%. Consortium members were given the opportunity to review and veto final minor changes, primarily formatting and style adjustments. No vetoes were exercised.

Disclosure

Importantly, while the scope of the *Metrics Reloaded* framework also comprises ObD, SemS and InS, we will focus solely on the results for the *ImLC* parts. The remainder of this chapter will describe the methodological steps taken by the *ImLC* expert group (under my leadership and in close collaboration with the core team) and the recommendations that were finally generated. This restriction may sometimes lead to omission of individual elements in the lists presented in this thesis, but for better reference we will mostly keep the wording and numbering of the original publication [238].

Methodologically the work of the ImLC group can be broken down as depicted in Fig. 4.2. Therein the following components reside:

Use case The application requesting the recommendation. It comprises the abstract 'intentions' of the domain experts, some actual data in the form of a task $\mathcal{T}$ to be evaluated, and also the nature of the data distribution $p_{\mathcal{T}}(X, Y)$ (see Def. 2.3).

Metrics A selection of potentially suitable performance measures for the use case. The goal is to recommend a subset of metrics that captures the intent of the use case, while avoiding pitfalls.

Properties The metric definitions result in certain properties of metrics that deem them appropriate or inappropriate under various conditions. Desirable properties can capture the degree of fit with the associated use case intentions, while undesirable properties can lead to misinterpretation or incorrect conclusions about the fitness of a model.

Pitfalls A structured collection of 'incorrect conclusions' that may be drawn from performance measures under certain conditions. Any recommendation for a use case should avoid suggesting metrics that are susceptible to the pitfalls that apply to the conditions of the use case.

Problem fingerprints The structured responses to a series of binary and multiple choice questions, that attempt to capture the conditions of the use case and the intentions of the domain experts. Only a subset of these questions need be answered to obtain recommendations.

Recommendation interview The actual process of asking the domain experts questions to obtain metric recommendations.

We will describe the full set of metrics and the summary of relevant properties in Sec. 4.1.2, and the pitfalls and problem fingerprints in Sec. 4.1.3 as part of the methodology. The recommendation process will be described as part of the results in Sec. 4.2.1. Some concrete use cases are also described in Sec. 4.2.3.

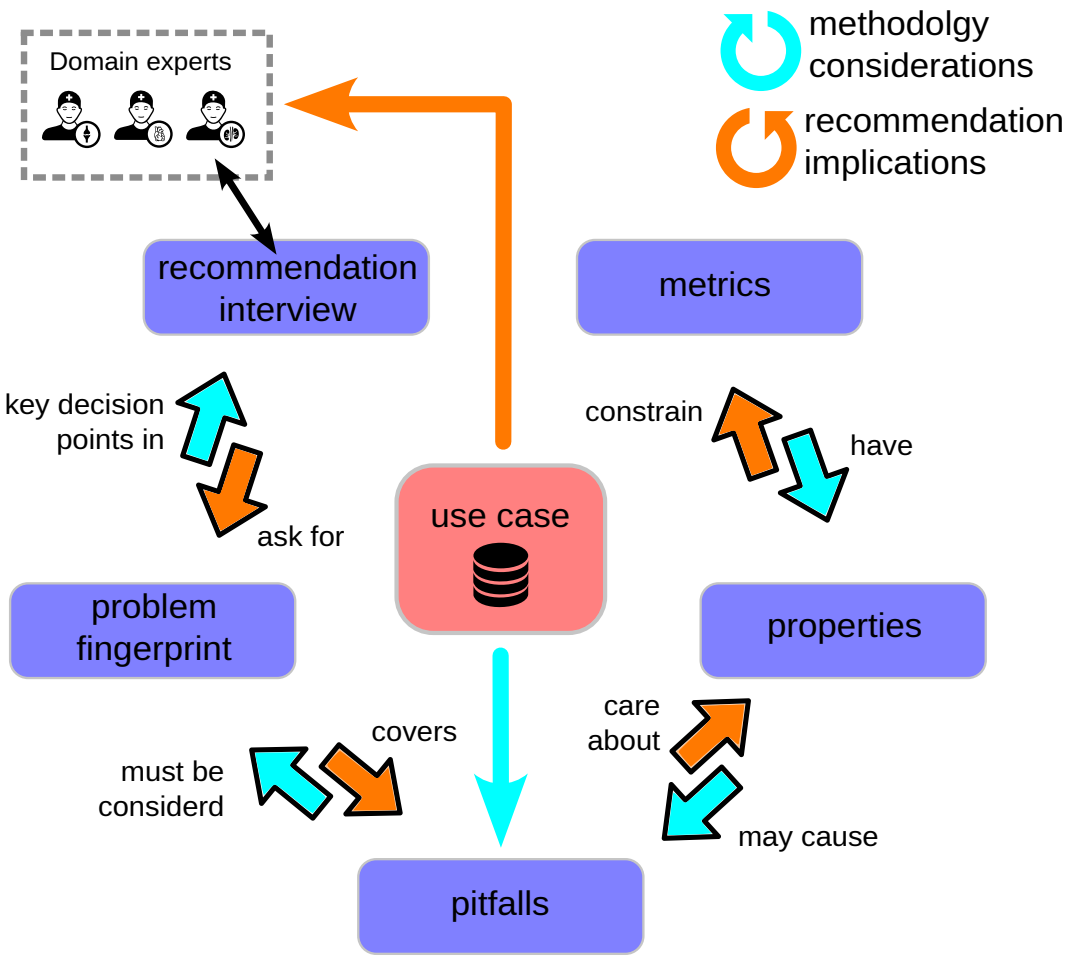


Figure 4.2: Methodological steps for recommendation generation. Methodologically, it is necessary to understand metric properties and how they combine with use case characteristics to create pitfalls in metric suitability. Relevant pitfalls are covered by the problem fingerprint, which determines the key decisions during the recommendation interview with domain experts. Since the domain experts know the use case characteristics, it is possible to select appropriate metrics in the forward pass to avoid pitfalls.

4.1.2 Metric pool

The list of candidate performance measures within the framework evolved through the iterations described above. Initially, metrics from established international biomedical competitions [240], complemented by previous studies [147, 363, 377], were considered. While some metrics were deemed less important by the experts in the surveys, others were added, e. g., during the crowdsourcing. The final *pool* of metrics is given in Tab. 4.1

with references to the precise definitions given in Chap. 2. The categorization[1] into *counting metrics*, which are based on a single confusion matrix, *multi-threshold metrics*, which are based on multiple confusion matrices via a sweep over thresholds, and finally *calibration metrics*, which are based solely on class probabilities, has already been used along with the introduction of metrics in Sec. 2.4, Sec. 2.5 and Sec. 2.6.

To understand why a particular metric is susceptible to a particular pitfall, we need to dissect metrics and identify properties that, in conjunction with the given use case, are the *cause* for a particular pitfall. We thus enrich the task pool overview with some additional properties to better explain their nature.

One of these properties, which is a major pitfall for model assessment in this chapter (and will also be vital in Chap. 6), is the influence of *prevalences* $\mathcal{P}$ on the performance measure μ. The chain of thought here is as follows: If μ depends on $\mathcal{P}$ (which itself depends only on $\mathcal{T}_{\text{test}}$), but $\mathcal{P}$ is assumed to be not 'representative' for the final application (e. g., because the data collection for $\mathcal{T}_{\text{test}}$ was biased, hence the realizations $\mathcal{T}_{\text{test}}$ are not i.i.d. from the actual problem distribution $p_{\mathcal{T}}(X, Y)$ as given in Def. 2.3), then the value of μ on the application data $\mathcal{T}_{\text{app}}$ is likely to differ from the value of μ on $\mathcal{T}_{\text{test}}$. Some metrics avoid or compensate for this influence by weighting classes equally and focussing mainly on TPR or similar aspects of the confusion matrix. Metrics that incorporate the *predictive perspective* as given in Def. 2.35 are prevalence dependent, as discussed after Prop. 2.38.

Another key property is whether a performance measure is inherently multiclass or relies on a one-versus-the-rest confusion matrix (Def. 2.15) to be defined for a particular class. We will refer to the former as **multiclass** metrics and the latter as **per-class** metrics. The ability to weight classes for multiclass metrics and the ability to plug in individual confusion costs also differentiate metrics.

Table 4.1: **Metric pool.** The recommended classification metrics along with the reference of their definition in this thesis, their value range and orientation (↑ positively, ↓ negatively), whether they are applicable to multiclass tasks without ovr reduction (Def. 2.15), depend on the task prevalence and allow confusion costs. Table inspired by Table SN 2.1 from Maier-Hein et al. [238].

Metric	Acr.	Def.	Value range	Multi-class	Class weight-ing	Prev. dep.	Confu-sion costs
Counting metrics							
Accuracy	AC	2.18	$[0, 1]$ ↑	yes		yes	
Balanced Accuracy	BA	2.21	$[0, 1]$ ↑	yes	equal		

Continued on next page

[1] The identification and naming of metric categories for each problem category was also a task of the expert groups.

Table 4.1: **Metric pool.** The recommended classification metrics along with the reference of their definition in this thesis, their value range and orientation (↑ positively, ↓ negatively), whether they are applicable to multiclass tasks without ovr reduction (Def. 2.15), depend on the task prevalence and allow confusion costs. Table inspired by Table SN 2.1 from Maier-Hein et al. [238]. (Continued)

Metric	Acr.	Def.	Value range	Multi-class	Class weighting	Prev. dep.	Confusion costs
Matthews Correlation Coefficient	MCC	2.49	$[-1, 1] \uparrow$	yes	equal	yes	
Weighted Cohen's Kappa	WCK	2.30	$[-1, 1] \uparrow$	yes	via costs	yes	yes
Expected Cost	EC	2.23	$(-\inf, \inf) \downarrow$	yes	via costs	yes	yes
True Positive Rate	TPR	2.16	$[0, 1] \uparrow$				
True Negative Rate	TNR	2.16	$[0, 1] \uparrow$				
Positive Likelihood Ratio	LR+	2.32	$[0, \inf) \uparrow$				
Positive Predictive Value	PPV	2.35	$[0, 1] \uparrow$			yes	
Negative Predictive Value	NPV	2.35	$[0, 1] \uparrow$			yes	
F-beta	F-beta	2.44	$[0, 1] \uparrow$			yes	via β
Net Benefit	NB	2.57	$(-\inf, 1] \uparrow$			yes	via τ
Multi-threshold metrics							
Area under the Receiver Operating Characteristic Curve	AUROC	2.55	$[0, 1] \uparrow$				
Average Precision	AP	2.56	$[0, 1] \uparrow$			yes	
Calibration metrics							
Brier Score	BS	2.74	$[0, 2] \downarrow$	yes		yes	
Root Brier Score	RBS	2.74	$[0, \sqrt{2}] \downarrow$	yes		yes	
Negative Log Likelihood	NLL	2.73	$[0, \inf) \downarrow$	yes		yes	
Expected Calibration Error	ECE	2.66	$[0, 1] \downarrow$	yes		yes	
Class-wise Calibration Error	CWCE	2.68	$[0, 1] \downarrow$	yes	equal		
Kernel Calibration Error	KCE	2.69	$[0, \inf) \downarrow$	yes		yes	
Expected Calibration Error Kernel Density Estimate	ECE^{KDE}	2.70	$[0, 2] \downarrow$	yes		yes	

As demonstrated in Sec. 2.4, many counting metrics are closely related and hence do not necessarily provide 'complementary' information about the performance of a categorical classifier. Recommending multiple related metrics may distort the overall

picture and suppress less represented properties. In addition to the properties from Tab. 4.1, these relationships were an important criterion for including or recommending specific metrics. We summarize the relationships in Tab. 4.2. Thereby we distinguish the following categories of relationships:

1. Cyan: Given one metric the other is computable without any further information from the confusion matrix.
2. Orange: This reflects mutual computability under strong assumptions: We fix $\beta = 1$ for the F-beta score, 0-1-costs for WCK and EC and assume $\mathcal{T}$ is a balanced binary task.
3. Black: This relation captures cases where one metric is a generalization (or instantiation) of another. In other words, for some fixed 'parameters' of one metric it coincides with another (parameter-free) metric.
4. Yellow: All other noteworthy relationships are covered in this category. Most of the time, one metric is an 'ingredient' in the calculation of another.

4.1.3 Problem fingerprints

In order to provide tailored metric recommendations, we need to be aware of the pitfalls that should be avoided for the particular use case. Although published separately [317], the collection, analysis, and systematic categorization of pitfalls related to performance measures was a key driver for progress in *Metrics Reloaded*. Notably, the collected pitfalls targeted not only the appropriate *selection* of a performance measure, but also the correct *identification of the problem category*, i. e., in our case, whether the formulation of the medical questions in terms of an ImLC problem is appropriate, and the correct *application* of a performance measure, which concerns, e. g., the choice of parameters, appropriate aggregation, and adequate reporting. For the scope of this thesis, we focus primarily on the following pitfalls[2] [317]:

[P2.1] **Disregard of the domain interest.** These are the aspects of the use case that relate to the intentions of the domain experts. It comprises:

(i) **Importance of confidence awareness** e. g., evaluating only counting metrics, although the probability estimates of a classifier are required in the application.

(ii) **Importance of comparability across datasets** e. g., comparing models that have been assessed via AC on different tasks with different prevalences.

[2]We will discuss some other pitfalls in Tab. 4.3.

Table 4.2: Metric relations. We reference noteworthy relationships between metrics from Chap. 2 above the diagonal. Below the diagonal we use colors to code the kind of relationship between metrics. Cyan: mutually computable, Orange: mutually computable with standard choice of metric parmeters for binary and balanced tasks, Black: generalization/instantiation, Yellow: other notable relationship. Table inspired by Fig SN 2.1 from Maier-Hein et al. [238].

	AC	BA	CK	EC	ER	F-beta	F1	J	JAC	LR+	MCC	MK	NB	NPV	PPV	TNR	TPR	WCK
AC	x	2.22			2.19						2.49				2.40		2.20	
BA		x						2.34									2.21	2.31
CK			x															2.30
EC				x	2.24												2.25	
ER					x													
F-beta						x	2.46								2.44		2.44	
F1							x		2.48									
J								x			2.52	2.42				2.33	2.33	
JAC									x									
LR+										x						2.32	2.32	
MCC											x	2.52						
MK												x		2.41	2.41			
NB													x			2.58	2.58	
NPV														x		2.39		
PPV															x		2.39	
TNR																x		
TPR																	x	
WCK																		x

(iii) **Unequal severity of class confusions** e. g., treating all confusions equally for an ordinal or hierarchical structure of classes.

(iv) **Importance of cost benefit analysis** e. g., ignoring the clinically justifiable threshold for false positives in relation with true positives.

[P2.3] **Disregard of the properties of the dataset.** These are the aspects of the use case that relate to the data of the task. This includes:

(i) **High class imbalance** e. g., performing evaluation with a metric that does not reflect the relative performance compared with the naive classifier.

[P2.4] **Disregard of the properties of the algorithm output.** These are the aspects of the use case that relate to the model. These are:

(i) **Availability of predicted class scores** e. g., approximate AP with a single point on the curve.

[P3.2] **Inadequate metric aggregation.** These are the aspects of the combination of metric values into model scores. These are:

(i) **Non-independence of test cases** e. g., having multiple samples from the same patient or hospital.

(ii) **Possibility of invalid prediction** e. g., incorrectly formatted outputs, canceled prediction computations due to resource constraints, or failed model inference by throwing an exception.

Straightly related to these pitfalls are the categories for the problem fingerprints.[3]

[FP2] **Domain interest-related properties.** These are the aspects of the use case that relate to the intentions of the domain experts. It comprises:

[FP2.5] **Penalization of errors:** addresses **[P2.1] (iii)** and **[P2.3] (i)**

[FP2.6] **Decision rule strategy:** addresses **[P2.1] (iv)**

[FP2.7] **Calibration of predicted class scores:** addresses **[P2.1] (i)**

[FP4] **Dataset-related properties.** These are the aspects of the use case that relate to the data of the task. This includes:

[FP4.1] **High class imbalance:** addresses **[P2.3] (i)**

[FP4.2] **Provided class prevalences reflect the population of interest:** addresses **[P2.1] (ii)**

[FP4.3] **Non-independence of test cases** addresses **[P3.2] (i)**

[3] The given numbering is an example for selective relevance from the *Metrics Reloaded* framework: The fingerprint families **FP1 - Problem category** and **FP3 - Target structure-related properties** are left out as we assume a whole image problem to be given (see Def. 2.2).

[FP5] **Algorithm output-related properties.** These are the aspects of the use case that relate to the model. These are:

[FP5.1] **Availability of predicted class scores:** addresses **[P2.4] (i)**

[FP5.3] **Possibility of invalid algorithm output:** addresses **[P3.2] (ii)**

A granular overview on the problem fingerprint elements and individual descriptions are given in Figs. 4.3, 4.4, and 4.5.

We provide more details for some fingerprint items that are not sufficiently self-explanatory:

FP2.5.5: Compensation for class imbalances

While AC remains the dominant metric for multiclass classification (see Sec. 6.2.1 and Maier-Hein et al. [240]), it exhibits significant weaknesses when faced with class imbalances. Consider a binary classification scenario with the following confusion matrix TP=0, FP=1, FN=1, TN=10 000, which paradoxically yields an AC of approximately 1. This seemingly perfect score masks three critical methodological pitfalls that fundamentally undermine the metric's reliability in imbalanced scenarios:

Misleading metric values due to missing reference value for naive classifier: In the given example, the near-perfect score hides the fact that the same performance could have been achieved by a naive classifier that always predicts the dominant class (see Def. 2.26). In general, in balanced scenarios, the AC of a naive classifier is $1/C$ (follows from Prop. 2.22 and Def. 2.21, where the TPRs will necessarily sum to one), which serves as an important anchor when interpreting the metric scores. However, when class imbalances are present, no such interpretation can be made, and the naive reference depends on the class prevalences.

Misleading metric values due to unequal contribution of classes to the metric score: In the example provided, the near-perfect score masks the fact that all samples of the positive class (here: one sample) were misclassified. While all classes contribute similarly to the AC metric in balanced scenarios, frequent classes dominate the performance score in imbalanced settings (see Prop. 2.20). While none of the rare cases are correctly classified, the metric achieves a near perfect score due to the very good performance on the dominant class. Prevalence-independent metrics (see Tab. 4.1) are based on the equal contribution of each class irrespective of prevalence.

Misleading metric values due to missing consideration of predictive values: In our example, the near-perfect score hides the fact that the PPV of this system is zero. Generally, in balanced scenarios, high AC scores imply high predictive values (see Cor. 2.40), which are important indicators of the utility of a classification system in

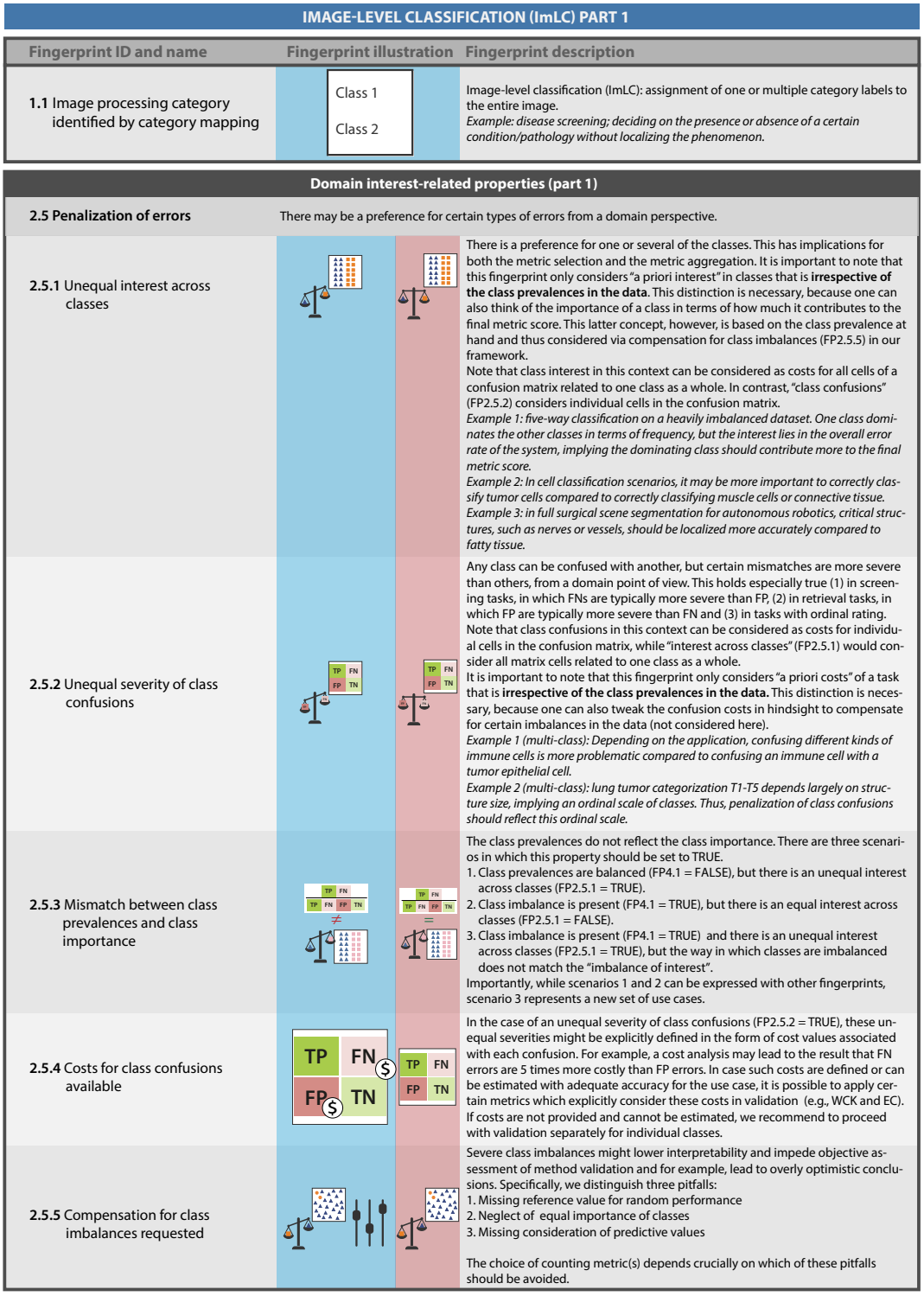


Figure 4.3: Elements of the problem fingerprint with corresponding questions (part I). The columns display from left to right: fingerprint ID and name, an illustration for the corresponding property in the blue column (for binary tasks the red column displays a counterexample), and a detailed description in the last column. Originally published in Maier-Hein et al. [238].

IMAGE-LEVEL CLASSIFICATION (ImLC) PART 2

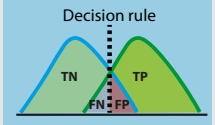
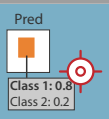
Fingerprint ID and name	Fingerprint illustration	Fingerprint description
Domain interest-related properties (part 2)		
2.6 Decision rule applied to predicted class scores Options: - Target value-based - Optimization-based - Argmax-based - Cost-benefit-based - No decision rule applied		Modern algorithms output continuous class scores. Making a classification decision requires identifying a decision rule applied to the scores, which amounts to setting a cutoff value in binary tasks. A product of this process is the (decision rule-specific) confusion matrix. This matrix enables the computation of popular single-threshold counting metrics, such as Sensitivity, PPV and F_1 Score. Depending on domain interest the decision rule can be set in multiple ways: Target value-based (only for binary tasks): The cutoff represents the threshold for which a specific target metric value (e.g., Sensitivity = 0.95) is achieved. Importantly, this threshold has to be determined on a separate data split. Other metric values (e.g., Specificity) are then reported for this specific threshold. We use the notation Metric@(TargetMetric = TargetValue) (e.g., Specificity@Sensitivity = 0.95) in this case. This cutoff strategy is limited to binary classification problems. Optimization-based: The decision rule is inferred by optimizing a target metric, such as the F_1 Score in the binary case, on a separate data split. Argmax-based: Especially in multi-class scenarios and if no target value is defined, no separate data split for optimization is available, or there are concerns w.r.t generalization of data-based decision rule optimization, a common option is to follow the principle of a Bayes classifier and pick the class with the highest predicted class score. Cost-benefit-based: In case the predicted class scores are calibrated (see FP2.7), and there is either a task related risk-cutoff (only for binary classification tasks, e.g. "only treat patients with cancer risk >10%"), or explicit costs for misclassification errors provided, one can apply this decision rule directly to the scores without data-driven optimization. Notably, in binary classification tasks provided risk cutoffs are often based on a cost ratio of TP versus FP (e.g., not more than 10 FPs per 1 TP should be treated). In most cases, no specific risk cutoff can be determined, thus model performance is plotted over a reasonable range of risk scores ("decision curve analysis"). No decision rule applied: Examples for no interest in validating a method at a specific decision rule are 1) focus on general methodological performance across many tasks and data sets without application interest, 2) concerns regarding the comparability of results based on a single decision rule that is fixed across varying study cohorts (see also FP4.2), or 3) focus on the probabilistic predictions to obtain and communicate personalized risk factors of indi-
2.7 Calibration of predicted class scores		When validating classification methods - particularly those with applications that involve direct human read-out - it is often crucial for the predicted class scores themselves to be interpretable. A system is well-calibrated if the predicted class scores (i.e., the output of the model) reflect the true probabilities of the outcome (formal definition see App. C.6).
2.7.1 Calibration assessment requested		This property should be set to TRUE if the predicted class scores should reflect the true probabilities of the outcome. An obvious requirement for this assessment is that predicted class scores are available (FP5.1 = TRUE). Methods subject to validation in this context are either classification models whose inherent calibration quality is assessed, or re-calibration methods, which are typically accuracy-preserving (bijective) transformations on the classifier outputs aiming to improve calibration quality.
2.7.2 Comparative calibration assessment requested Options: - Comparison of re-calibration methods on the same classifier - Comparison of calibration performance across classifiers - Comparison of overall performance across classifiers - No comparative assessment		Comparison of re-calibration methods on the same classifier: The potential benefit of one or more re-calibration methods is to be assessed and compared. The desired validation output is a ranking of re-calibration methods (including the performance of "no re-calibration") from which the best method can be selected. Comparison of calibration performance across classifiers: This comparison of classification models potentially includes re-calibration methods applied on their outputs. The desired validation output is a ranking of methods according to calibration quality. Comparison of overall performance across classifiers: Overall performance refers to the joint assessment of discrimination performance and calibration quality. This comparison of classification models potentially includes re-calibration methods applied on their outputs. The desired validation output is a single ranking naturally weighting both aspects. No comparative assessment: If the interest lies in understanding the reliability of predicted class scores for one given model, no metrics for comparative assessment are required.
2.7.3 Assessment of interpretability of model outputs requested Options: - Assessment of calibration error in isolation - Joint assessment of calibration and discrimination - No assessment of interpretability required		There is an interest in understanding the reliability of predicted class scores for a given model as a basis for interpreting and communicating results. The desired validation output is a single score that captures the interpretability of the output. We differentiate two notions of interpretability that correspond to different families of calibration-sensitive metrics. Assessment of calibration error in isolation: The user may be interested in assessing the calibration quality of a model "in isolation" (i.e., without associated discrimination power). Quantifying this property is often a desirable aspect of validating a model. Knowing that predicted class scores are well-calibrated allows making specific statements about individual output scores such as "80% of outputs with score 0.8 belong to the true class". Importantly, calibration metrics do not assess whether outputs match the true posterior probabilities. Hence, the "risk statements" depend on the model at hand (and its discrimination power) and thus cannot associate a true model-independent risk to a data sample (e.g., patient). Joint assessment of calibration and discrimination: An alternative approach to assessing the interpretability of outputs is to compare the predicted class scores directly to the reference and thus to quantify the overall performance of a model (discrimination and calibration) in one joint score. This assessment can also be interpreted as measuring whether scores match the true posteriors, e.g. the risks of individual patients. A disadvantage of this strategy is that the calibration error is conflated with discrimination performance, thus prohibiting statements about the reliability of particular scores such as "80% of outputs with score 0.8 belong to the true class".

Figure 4.4: Elements of the problem fingerprint with corresponding questions (part II). The columns display from left to right: fingerprint ID and name, an illustration for the corresponding property in the blue column (for binary tasks the red column displays a counterexample), and a detailed description in the last column. Originally published in Maier-Hein et al. [238].

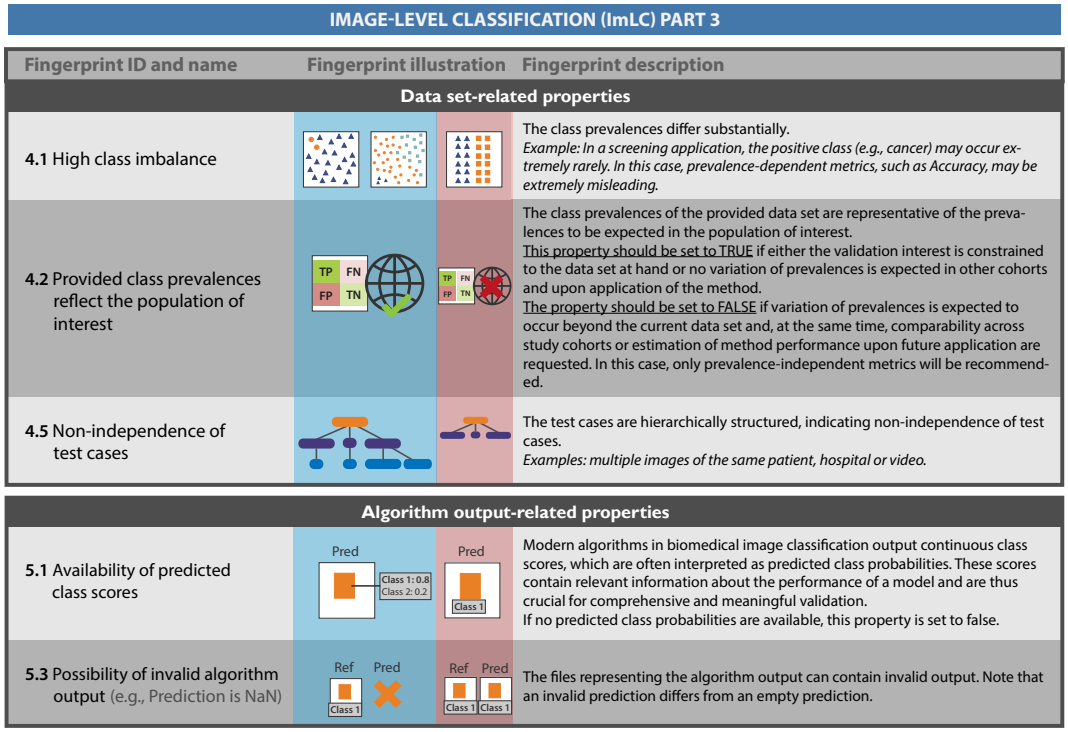


Figure 4.5: Elements of the problem fingerprint with corresponding questions (part III). The columns display from left to right: fingerprint ID and name, an illustration for the corresponding property in the blue column (for binary tasks the red column displays a counterexample), and a detailed description in the last column. Originally published in Maier-Hein et al. [238].

practice. This is not necessarily the case in imbalanced scenarios, as seen in the provided example, where the PPV is zero despite a high AC. To compensate for this effect, metrics that combine the Sensitivity and predictive value perspectives (see Sec. 2.4.3) can be considered, which explicitly assess the predictive performance of a classifier.

FP2.6: Decision rule applied to predicted class scores

The choice of whether to apply a decision rule during validation depends critically on the research objective: to emphasize either the quality of discrete classification decisions or the nuanced information contained in continuous probability predictions. While certain domains, such as cell InS, have standardized decision-rule-based validation [51], contemporary clinical research increasingly advocates decision-rule-agnostic approaches [260]. These perspectives argue that fixed decision rules often suffer from critical limitations: they tend to be over-optimized for specific datasets, produce non-transferable results across different study cohorts (see Chap. 6), and may not capture the complex cost-benefit dynamics of different clinical applications [403]. Moreover, continuous class probabilities are recognized as more informative for patient communication and clinical decision making [424][4]. Acknowledging this methodological tension, our approach provides flexibility by making the application of decision rules optional and allowing users to encode their validation preferences through a configurable problem fingerprint. The fingerprint thus supports multiple decision rule strategies to accommodate different research needs.

Target-value based (only for binary tasks) Certain clinical or research contexts require reaching a predefined performance threshold for a particular metric (e. g., requiring a sensitivity of 0.95). In such scenarios, we introduce the notation **Metric@(TargetMetric = TargetValue)**, such as 'Specificity@(Sensitivity = 0.95)', which represents the Specificity (=TNR) corresponding to the target Sensitivity (=TPR). A critical methodological consideration is that this cutoff determination must be performed using a separate, dedicated data partition to ensure methodological rigor and to avoid overfitting.

Optimization-based In the absence of a predefined target value, decision rules can be derived through data-driven optimization of a primary performance metric (e. g., F1) utilizing a dedicated configuration dataset. While one-dimensional cutoff scanning remains straightforward for binary classification tasks, extending this approach to multiclass scenarios introduces substantial computational and methodological complexities. Identifying optimal decision rules across multiple classes requires more sophisticated optimization strategies that can navigate the increased dimensionality and interdependencies of the classification space.

[4] Wynants et al. [424] have provocatively referred to the routine imposition of binary thresholds on continuous predictions as "dichotomania"

Argmax-based A common decision rule strategy is based on the argmax decision rule (see Def. 2.9). This approach is based on the fundamental assumption that the highest class score corresponds to the most likely true class. While rooted in Bayesian classification theory, which defines this as a 'Bayes classifier', the theoretical validity of the method depends on two critical conditions: equal severity of class confusions (FP2.5.2=False) and well-calibrated class probability scores. These nuanced requirements are often overlooked in practical implementations. Chap. 6 provides a comprehensive examination of the potential pitfalls and limitations inherent in this seemingly straightforward decision strategy[5].

Cost-benefit-based When probability scores are appropriately calibrated (see FP2.7) and task-specific confusion costs (see Def. 2.23) or risk thresholds (only for binary classification tasks, e. g., 'treat only patients with cancer risk >10%') are available, decision rules can be applied directly without additional data-driven optimization. In binary classification, cost-benefit cutoffs often reflect error asymmetry, such as tolerating no more than 10 false positives for every true positive (see Def. 2.57). More sophisticated approaches define explicit costs for both false positives and false negatives, allowing nuanced quantification of classification consequences (see **DG3.2** in Sec. 4.2.2). While traditionally conceived for binary contexts, these cost-based decision strategies can be systematically extended to multiclass classification scenarios [113], providing a flexible framework for integrating domain-specific risk considerations into predictive modeling.[6]

No decision rule applied An alternative methodological approach is to omit a predefined decision rule altogether. Instead, this strategy focuses on evaluating discriminative performance through multi-threshold metrics that systematically explore different classification cutoffs, coupled with proper scoring rules that simultaneously assess model calibration and discriminative capacity. This approach provides a characterization of a classifier's predictive capabilities beyond the constraints of a single decision threshold.

FP2.7: Calibration of predicted class scores

The choice of the calibration condition to be validated and the metric to be used depends on the domain interest. The methods to be validated in this context are either classification models, whose inherent calibration quality is to be assessed, or so-called *re-calibration methods*, i. e., transformations on the classifier outputs aimed at improving the calibration quality (see Def. 2.67). In the most common scenarios, the driving interest may be either a comparative performance evaluation (FP2.7.2), where methods are ranked

[5] More specifically, Sec. 6.2.4 shows the consequences of ignoring the necessary precondition of appropriate model calibration or considering performance measures that do not translate to 0-1-costs.

[6] We present such a cost-based decision rule in Def. 6.9 and analyze it in Sec. 6.2.4.

according to calibration quality, or an absolute performance evaluation (FP2.7.3), where an interpretable and communicable measure of calibration quality is desired. We have identified four main use cases (**U1-U4**) that our framework addresses (Fig. 4.6).

1. **FP2.7.2 Ranking methods to determine calibration quality:** The following use cases focus on the comparative assessment of the calibration quality of one or multiple classifiers.
   a) **Use case 1 (U1):** comparing the effect of one or more re-calibration methods on the same (fixed) classifier. The desired validation output is a ranking of re-calibration methods (possibly including the performance of 'no re-calibration') from which the best method can be selected.
   b) **Use case 2 (U2):** comparing the calibration quality of multiple classifiers on the same task. The desired output of the validation is a ranking of the classifiers according to their calibration quality. In practice, such a ranking should be accompanied by a ranking according to discrimination performance, as it is not recommended to base model selection on calibration performance alone.
   c) **Use case 3 (U3):** comparing the 'overall performance' of classifiers (optionally including potential re-calibration methods), i. e., a joint assessment of discrimination performance and calibration quality. The desired validation output is a single ranking that naturally weights both aspects.
2. **FP2.7.3 Interpreting model outputs:** Of complementary interest may be the analysis of the CE to assess the reliability of the predicted class scores of one or more classifiers.
   a) **Use case 4 (U4):** interest in understanding the reliability of predicted class scores for a given model as a basis for interpreting and communicating results. The desired validation output is a single score that provides insight into how well the model is calibrated. The reliability of model outputs is often considered crucial upon application, such as for clinical prediction models [101, 394, 432]. Importantly, **U4** can be used in addition to **U1**, **U2**, or **U3** as it is based on an orthogonal interest.

Because some decision rules assume calibrated model outputs, a further potential interest in calibration validation may be to determine the quality of a decision rule applied to predicted class scores (see FP2.6), i. e., to answer the question: 'How much better could the classifier's decisions under this rule have been if the predicted class scores had been calibrated?'. We give some answers to this question in Sec. 6.2.4.

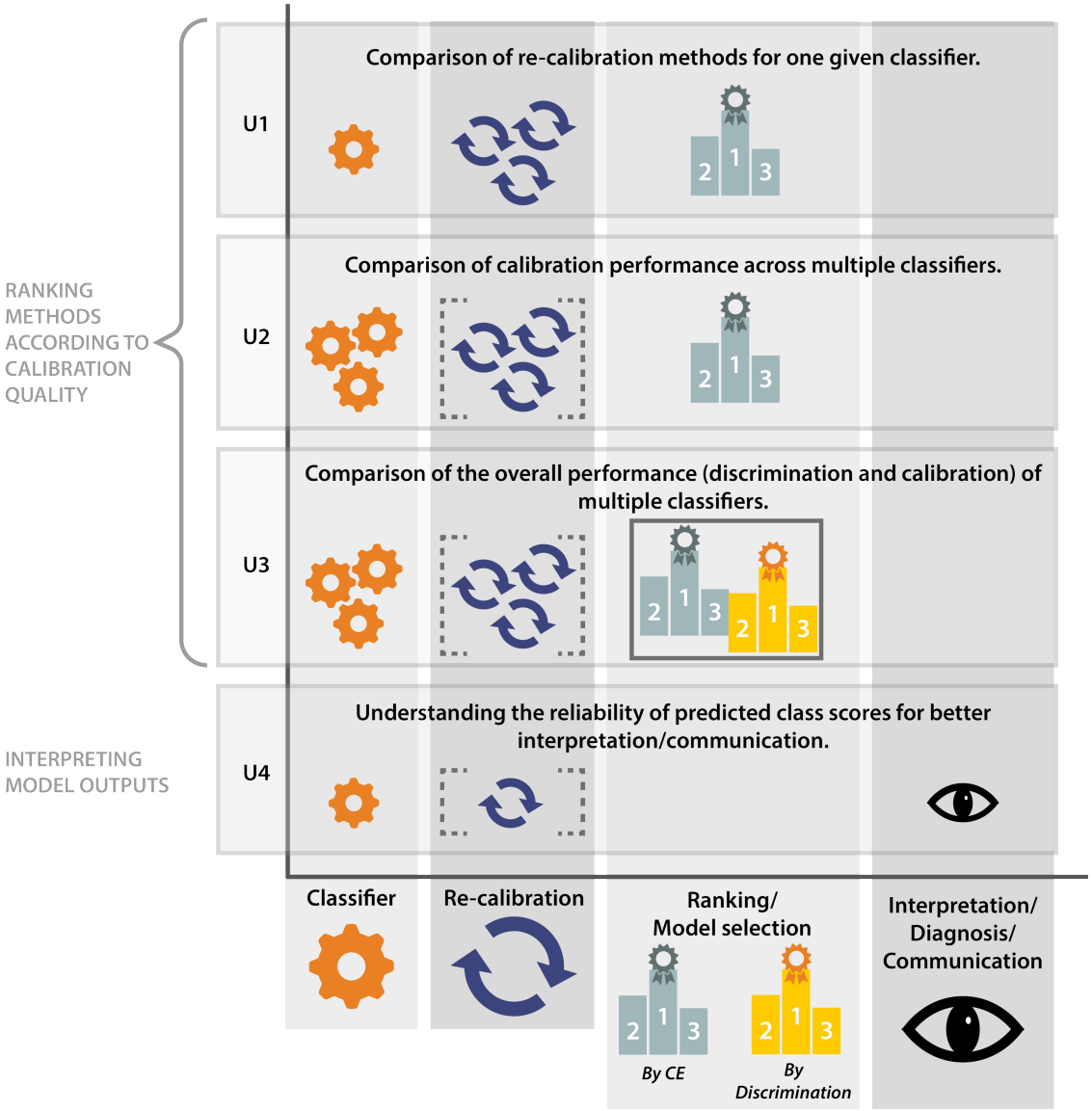


Figure 4.6: Underlying interest related to the assessment of calibration quality. The user is interested in the comparative calibration assessment (**U1-U3**) and/or obtaining a reliable estimate of the Calibration Error (CE) for interpreting and communicating the algorithm output (**U4**). The brackets around re-calibration methods denote that their application is optional in the corresponding use case. Originally published in Maier-Hein et al. [238].

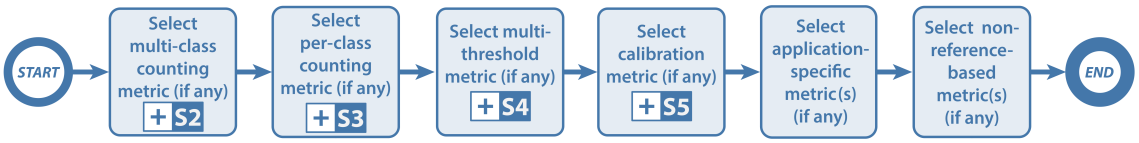


Figure 4.7: Selection process overview. The metric selection process is divided into subprocesses that each correspond to a different family of performance measures. Adapted from Maier-Hein et al. [238].

FP4.2 Class prevalences reflect the population of interest

Class prevalences and their differences between tasks are very important, although this aspect is often ignored in common validation practice. This is not a problem if the class prevalences of the provided test set reflect the population of interest, but can lead to problems otherwise (see Chap. 6). Therefore, this fingerprint should be set to true if either the validation interest is limited to the current dataset (no future comparison to datasets with different class prevalences is desired), or no variation in prevalences is expected in other cohorts and when applying the method.

4.2 Results

This section presents key findings from the *Metrics Reloaded initiative* previously introduced in Sec. 4.1.1. We first examine the metric selection process in Sec. 4.2.1. Following this, we provide comprehensive analysis of potential edge cases encountered during selection in Sec. 4.2.2. We will conclude with representative applications of the selection framework across specific clinical scenarios in Sec. 4.2.3.

4.2.1 Modelling the selection process

Process overview Fig. 4.7 provides a high-level overview of the metric selection process. Conceptually, it is divided into subprocesses that focus on different types of performance measures. The first four subprocesses (**S2-S5**) focus on *common reference-based metrics*. These performance measures all compare model outputs to labels. Next, the pool of standard metrics can be supplemented with *custom metrics* to address application-specific complementary properties. Finally, *non-reference-based metrics* can be added to the metric pool(s) to assess, for example, speed, memory consumption, or carbon footprint.

Subprocesses While the metrics from **S2** and **S3** rely on categorical decisions made by the model, the subprocesses **S4** and **S5** leverage potential probabilistic model outputs. Furthermore, metrics from **S3** and **S4** inspect individual classes (through the one-versus-the-rest mechanism described in Def. 2.15), while metrics in **S2** and **S5** assess classes

holistically. The multiclass metrics (see Tab. 4.1) from **S2** have the unique advantage that they capture the performance of an algorithm for all classes in a single score, without the need for custom class-aggregation schemes. On the other hand, they do now allow for detailed class-specific analyses. Therefore, we generally recommend performing an additional per-class validation with the metrics resulting from **S3** for all classes. To obtain a more comprehensive picture of a classifier's discriminatory performance, multi-threshold metrics (see Sec. 2.5) as given by **S4** work with a dynamic confusion matrix reflecting a range of possible thresholds applied to the predicted class scores. This compensates for the loss of information for decisions made by a single decision rule (see Def. 2.9). Finally, calibration metrics (see Sec. 2.6), such as those given by **S5**, operate without a decision rule and can evaluate either model calibration alone or joint calibration and discrimination by proper scoring rules (see Def. 2.71).

Phrasing of the biomedical task The recommendation framework has been designed in a way to support the metric selection and application process for one specific driving biomedical question. In practice, multiple questions are often addressed with one given data set, where a recommendation needs to be generated separately for each question. This specifically holds true for multi-label problems, in which multiple labels can simultaneously be assigned to the same image (e. g., 'multiple sclerosis' and 'brain tumor' both assigned to the same MRI image). In such a case, the problem should be converted to multiple binary tasks, for which the framework is traversed individually.

Process diagram symbols The notation for our recommendations has been based on Business Process Model and Notation (BPMN)[7]. The individual components used in the recommendation diagrams are explained in Fig. 4.8. Please note that we do not strictly follow BPMN to improve clarity of presentation.

Decision guides There are a couple of cases, when for a given problem fingerprint multiple metrics from the same family are feasible. Then a selection is driven by subtle nuances in the preferences from the domain experts. In our recommendation process such selection ambiguities are resolved via decision guides (**DG**) in Sec. 4.2.2, that help users make an educated decision when multiple options are possible.

S2: Select multiclass counting metric (if any) We recommend the selection of a multiclass counting metric based on **S2** (see Fig. 4.14) if a decision rule should be applied to the predicted class scores (FP2.6). In some use cases and especially in the presence of ordinal data, there may be an unequal severity of class confusions (FP2.5.2 = TRUE), implying that different costs to be applied to different errors reflected by the confusion matrix must be available (FP2.5.4 = TRUE). In this case, the only viable options are WCK

[7] `https://www.omg.org/spec/BPMN/`

and EC (see Tab. 4.1). While WCK is widely used, it comes with severe drawbacks (see **DG2.1**), such as high prevalence dependency and 'paradoxical results' [414] for the most common variant based on quadratic weights. For this reason, the consortium recommends EC as the default choice for the described scenario. In the case of equal costs, AC is the most widely used multiclass metric, but we recommend it in only one specific scenario: when the class prevalences in the data set reflect those in the target population (FP4.2) and potential class imbalances should not be compensated for. In the more general case, the decision boils down to either picking one of the prevalence-independent metrics EC or BA, which is specifically recommended if the class prevalences do *not* reflect the target population, or MCC, which has the important property that it requires not only the class-specific Sensitivities but also the corresponding predictive values to be high (see Cor. 2.52). Irrespective of the metric choice, we recommend additionally reporting the whole confusion matrix in the case of a reasonable number of classes.

S3: Select per-class counting metric (if any) As detailed class-specific analyses are not possible with multiclass counting metrics, which may potentially hide the poor performance of individual classes, we recommend an additional per-class validation with metrics selected according to **S3** (see Fig. 4.15). To this end, class-specific metric pools are generated. The choice of metric depends primarily on the decision rule applied to the predicted class scores (FP2.6; see Sec. 4.1.3). If a **target value-based** strategy is preferred, the decision rule applied to the predicted class scores is optimized such that a specific target value (e. g., Sensitivity = 0.95) is achieved. Complementary metrics, such as Specificity (see Fig. 4.2), can then be reported for this fixed value of the target metric (see **DG3.1**). In this case, the target metric is only reported for the specified target class. If a **cost-benefit-based** strategy is chosen (only recommended for binary classification tasks), we recommend selecting either NB (risk-centric view) or EC (cost-centric view) (see **DG3.2**). In contrast, in the case of **optimization-based** or **argmax-based** decision rules, the metric choice should be made between Sensitivity, LR+, and F-beta Score (see **DG3.3** and **DG3.4**).

S4: Select multi-threshold metric: To obtain a more comprehensive picture of the discrimination performance of a classifier, we always recommend the selection of a multi-threshold metric according to **S4** (see Fig. 4.16), irrespective of the decision rule. Multi-threshold metrics are again applied per class (see Sec. 2.5). A particular strength of AUROC is the fact that it is well-interpretable, as the value simply reflects the probability of a sample from the positive class being assigned a higher predicted class score compared to a sample from the negative class [111]. Furthermore, it is prevalence-independent and therefore well-suited for comparison of performance across different tasks. AP, on the other hand, is a prevalence-dependent metric, which comes with the advantage that predictive values are considered. This may be a crucial property in class-imbalanced scenarios where the focus is to be put on the rare class while AUROC scores are dominated

by the frequent class and may lead to overly optimistic interpretation.

S5: Select calibration metric (if any) If the calibration of a method should be assessed in addition to its discrimination capabilities (FP2.7.1), one or multiple calibration metrics should be chosen.

1: Select metric for comparative calibration assessment (if any): This step selects an adequate metric in case a comparative assessment of calibration methods is desired (FP2.7.2). The fingerprint FP2.7.2 covers the presented use cases **U1-U3** (see Fig. 4.6). For **U1** 'Comparison of re-calibration methods for the same fixed classifier', one option is to select a metric that assesses the canonical CE, such as KCE as an unbiased estimator of a canonical CE based on an alternative distance function, or ECE^{KDE} as a well-interpretable estimator of canonical calibration. Alternatively, an overall performance measure such as the BS can be used (see **DG5.2**), because the classifier is fixed in this scenario, the conflation of the CE with discrimination errors is no disturbing factor, and the true CE is exposed for relative comparison of scores. For **U2** 'Comparison of calibration quality across classifiers on the same task', we recommend reporting the CE per class by using an estimator of marginal CE, such as CWCE, if there is an unequal interest across classes (FP2.5.1). Otherwise the canonical CE should be assessed, e. g., using KCE or ECE^{KDE} (see **DG5.1**). For **U3** 'Comparison of overall performance across classifiers', we recommend reporting a Proper Scoring Rule (PSR) (i. e., BS or NLL, see **DG5.3**) as the joint assessment of calibration and discrimination is exactly what this category of metrics is designed for.

2: Select metric for assessing output interpretability (if any): This step selects an adequate metric for assessing the interpretability of the model output (FP2.7.3), which corresponds to **U4**. The first decision to be made in FP2.7.3 is whether to assess the calibration quality in isolation, as measured by CE estimates, or jointly with discrimination as measured by proper scoring rules. When deciding for calibration-only assessment, the core decision to be made is whether to measure top-label, marginal or canonical CE (see Def. 2.59 and **DG5.4**). If there is an unequal interest across classes (FP2.5.3), a well-interpretable estimator of the marginal CE, such as CWCE, is recommended. Otherwise, the default option is to select a well-interpretable estimator of the canonical CE (e. g., ECE^{KDE}) and a corresponding guaranteed upper bound (e. g., RBS), together with a per-class estimator of marginal CE (e. g., CWCE). Top-label calibration (as measured by ECE) is only recommended in rare cases (see **DG5.4**).

Note that the selection of the same metric (e. g., CWCE) in both steps is a potential outcome of the mapping. Crucially, metrics involving calibration assessment are generally prevalence-dependent. Thus, comparative studies as described in **U2** and **U3** are generally

restricted to one particular task and, if the prevalence of the data does not represent the population of interest (see FP4.2), the calibration quality of a classifier needs to be re-validated on each new study cohort (see Sec. 6.2.3).

Application of selected performance measures While metrics application may seem straightforward, numerous subtle pitfalls can compromise validation results. Tab. 4.3 presents our recommendations for avoiding these issues, organized into categories of implementation, aggregation, ranking, reporting, and interpretation – following the taxonomy established in parallel research [317]. Although certain aspects have been addressed in previous literature (e. g., Wiesenfarth et al. [420]), our work makes two significant novel contributions: the development of detailed 'Metric Cheat Sheets' that provide metric-specific guidance (accessible at [1]), and the implementation of all *Metrics Reloaded* metrics as part of the open-source MONAI framework [373] – making robust validation practices readily accessible to the broader research community.

Table 4.3: Recommendations for metric application addressing the pitfalls collected in Reinke et al. [317]. The first column comprises *all* sources of pitfalls captured by the published taxonomy that relate to the application of (already selected) metrics. The second column provides the *Metrics Reloaded* recommendation. The notation FPX.Y refers to a fingerprint item (see Sec. 4.1.3). Adapted from Maier-Hein et al. [238].

Source of Pitfall	Recommendation
	Metric implementation
Non-standardized metric definition and undefined corner cases	Use reference implementations in the open-source framework MONAI[373].
Discretization issues (e. g., in ECE or CWCE)	Use unbiased estimates of properties of interest if possible (e. g., RBS).
Metric-specific issues including sensitivity to hyperparameters	Read metric-specific recommendations in the cheat sheets [1].
	Aggregation
Hierarchical label/class structure	Address the potential correlation between classes when aggregating [190].
Multiclass problem	Complement validation with multiclass metrics (subprocess **S2**); perform weighted class aggregation if FP2.5.1 *Unequal interest across classes* holds.
Non-independence of test cases (FP4.5)	Respect the hierarchical data structure when aggregating metrics [227].
Risk of bias	Leverage metadata (e. g., on imaging device/protocol/center) to reveal potential algorithmic bias [22].
Possibility of invalid prediction (FP5.3)	set the corresponding metric value of 'undefined' predictions to the worst possible value, in cases of unbounded metrics (see Tab. 4.1) use the performance of a naive classifier (see Def. 2.26).
	Ranking
Metric relationships	Avoid combining closely related metrics (see Tab. 4.2) when choosing metrics to be used in algorithm ranking.
Ranking uncertainties	Provide information beyond plain tables that make possible uncertainties in rankings explicit [420].
	Reporting
Non-determinism of algorithms	Consider multiple test set runs to address the variability of results from non-determinism [374].
Uninformative visualization	Include a visualization of the raw metric values [420] and report the full confusion matrix unless FP2.6 = *no decision rule applied* holds.
	Interpretation
Low resolution	Read metric-related recommendations to obtain awareness of the pitfall [1].
Lack of lower/upper bounds	See Tab. 4.1 and read metric-related recommendations to obtain awareness of the pitfall [1].
Insufficient domain relevance of metric score differences	Report on the quality of the reference (e. g., intra-rater and inter-rater variability) [207]. Choose the number of decimal places such that they reflect both relevance and uncertainties of the reference. More than one decimal number is often not useful given the typically high inter-rater variability.

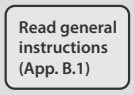
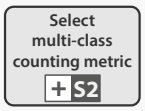
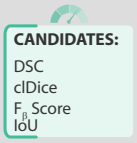
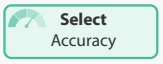
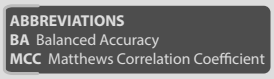
Symbol	Explanation
	Start of a process
	End of a process
	Task to be performed by the user
	Subprocess
	Task with reference to problem fingerprint
	Exclusion criterion
	Pool of options the user can choose from in the respective step
	A metric/criterion/strategy is selected
	Exclusive gateway: An exclusive gateway (or XOR-gateway) allows the user to make a decision. It can have multiple outgoing sequence flows. It is used when several conditions are mutually exclusive and only one selection is possible. An exclusive gateway is also used to join multiple incoming flows together and improve the readability of the diagram.
	Group
	Notes
	Further information
	The respective step needs to be repeated for each class

Figure 4.8: Overview of process diagram symbols. Our notation is based on the Business Process Model and Notation (BPMN) [418] graphical representation for specifying business processes. Originally published in Maier-Hein et al. [238].

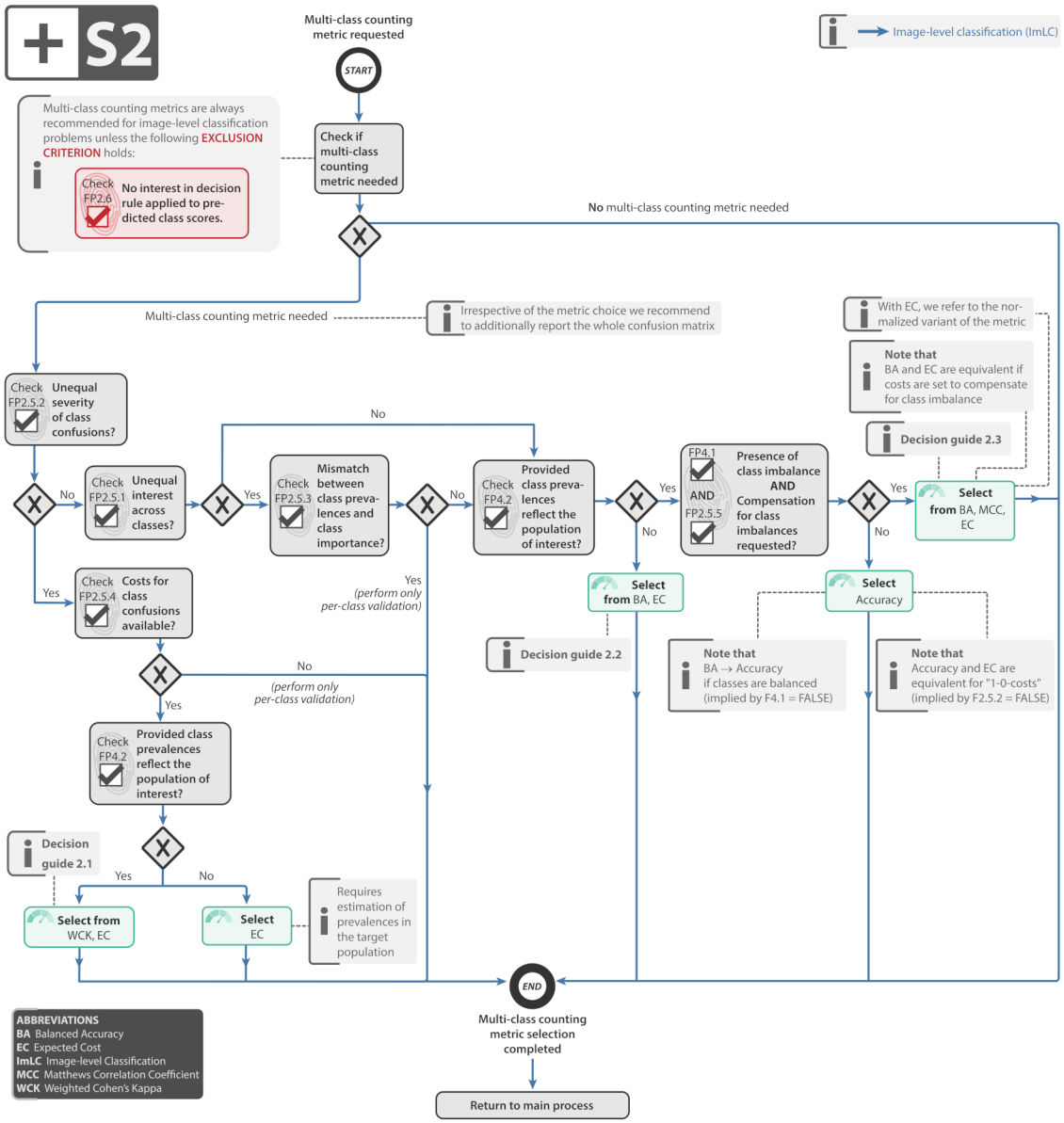


Figure 4.9: Subprocess S2 for selecting multiclass counting metrics. Decision guides are provided in Sec. 4.2.2. Originally published in Maier-Hein et al. [238].

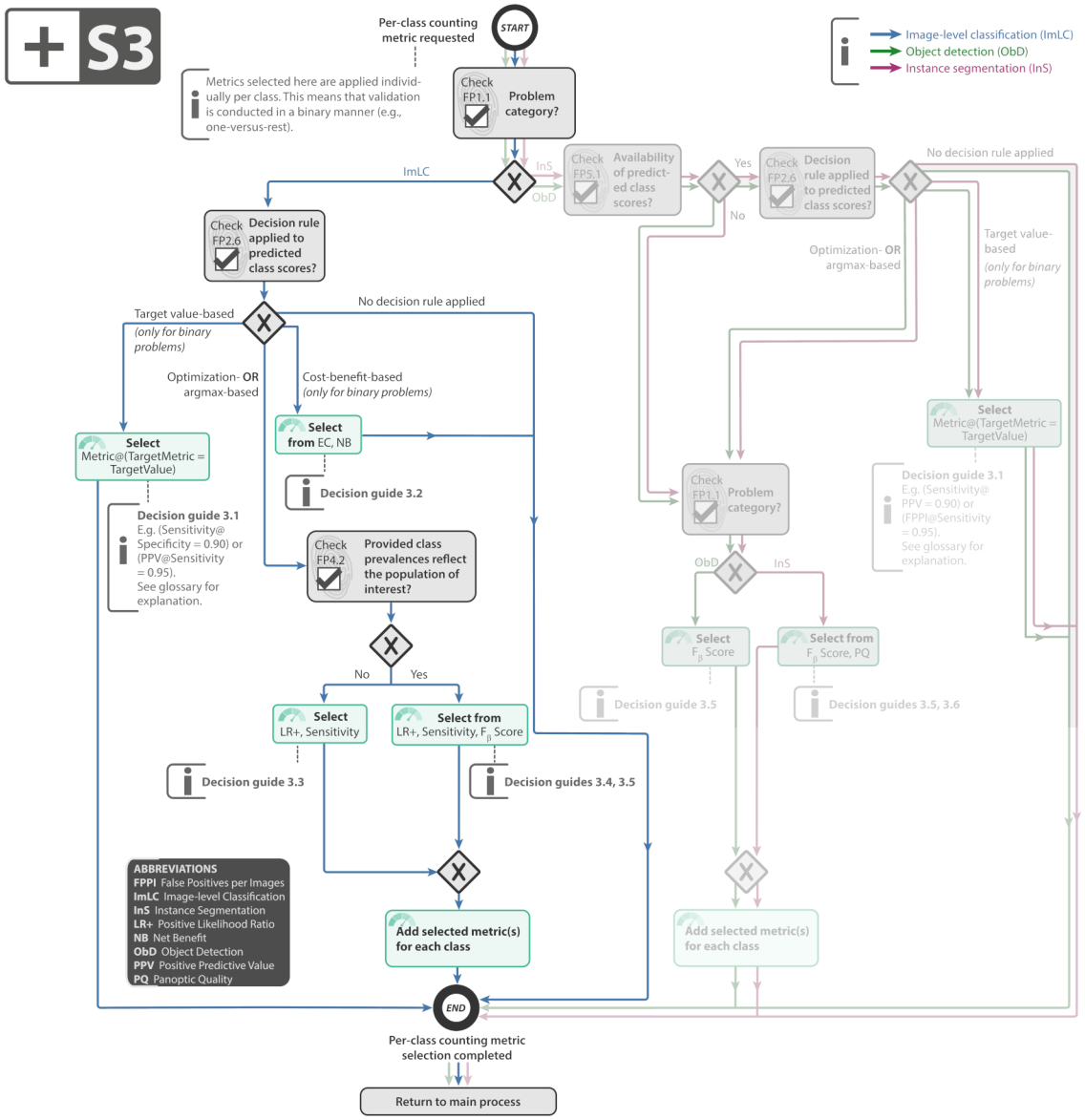


Figure 4.10: Subprocess S3 for selecting per-class counting metrics. Decision guides are provided in Sec. 4.2.2. Irrelevant paths for ObD and InS have been grayed out. Originally published in Maier-Hein et al. [238].

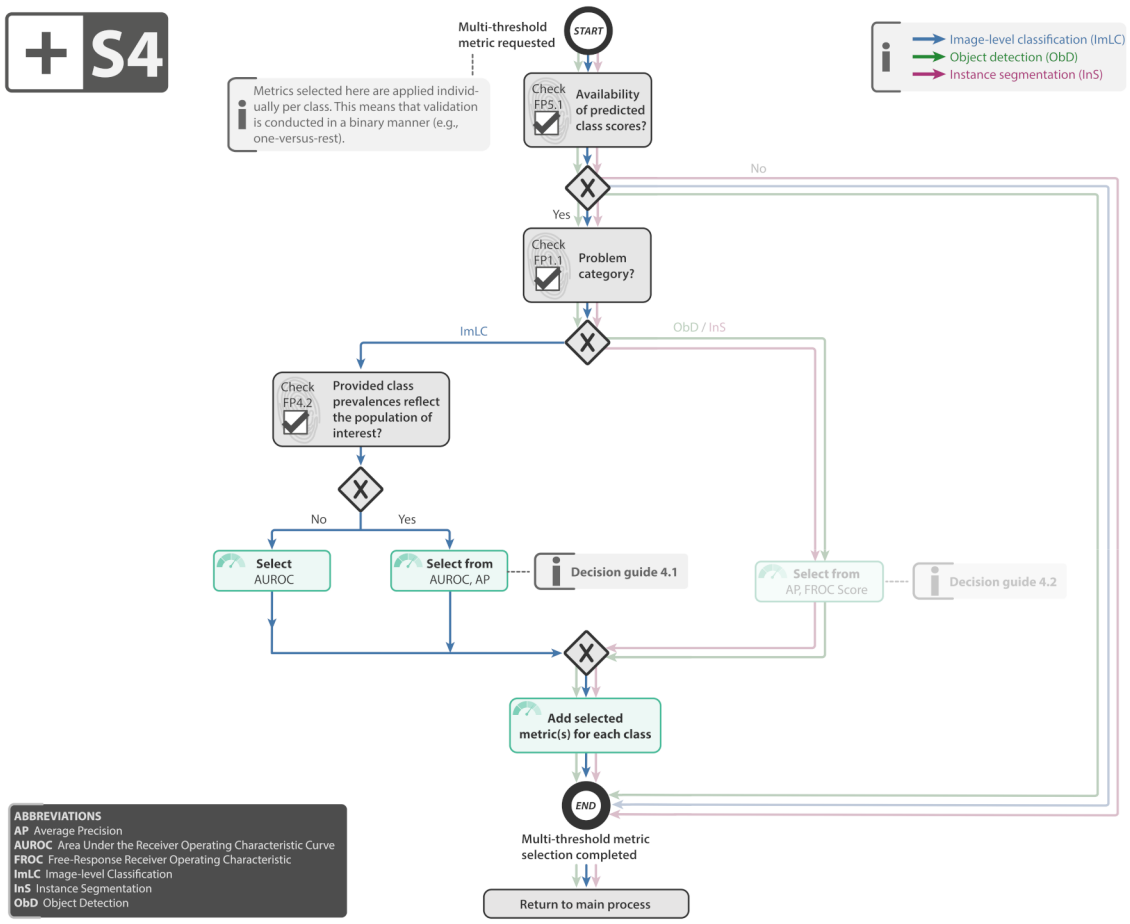


Figure 4.11: Subprocess S4 for selecting multi-threshold metrics. Decision guides are provided in Sec. 4.2.2. Irrelevant paths for ObD and InS have been grayed out. Originally published in Maier-Hein et al. [238].

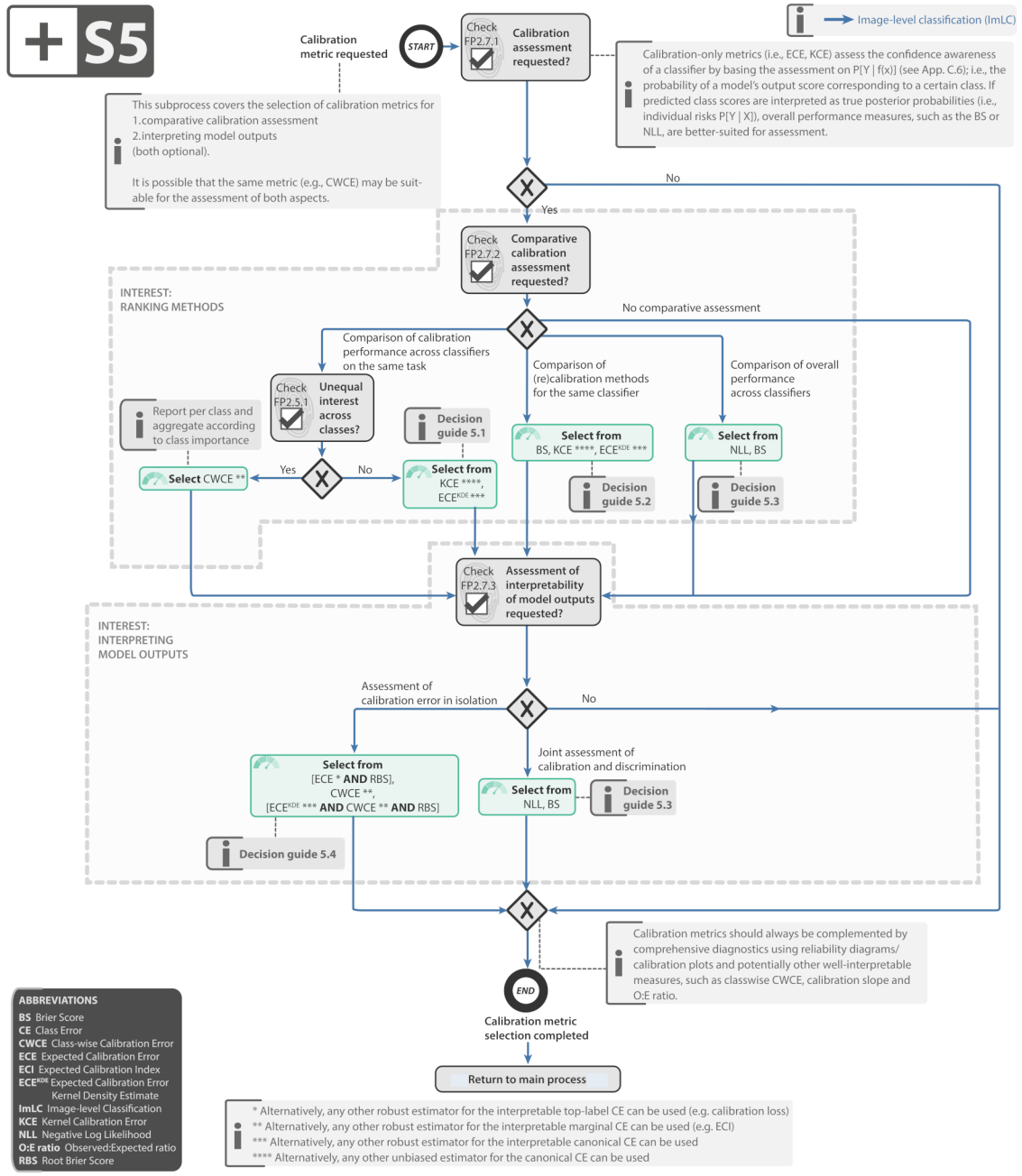


Figure 4.12: Subprocess S5 for selecting a calibration metric (if any). Decision guides are provided in Sec. 4.2.2. Further suggested calibration approaches include reliability diagrams [91], calibration slope [368], and O:E ratio [320]. Originally published in Maier-Hein et al. [238].

4.2.2 Resolving ambiguous cases

While the problem fingerprint helps exclude common metrics that are not suitable for the driving problem, the final choice in each subprocess may not be unambiguous. In these cases, decision guides support the domain experts in making an educated decision that best matches the domain interest.

DG2.1: Weighted Cohen's Kappa (WCK) versus Expected Cost (EC)

Table 4.4: Comparison of Weighted Cohen's Kappa (WCK) to Expected Cost (EC) in the context of decision guide DG2.1. Context: unequal severity of class confusions (FP2.5.2 = TRUE), costs for class confusions available (FP2.5.3 = TRUE), and provided class prevalences reflect the population of interest (FP4.2 = TRUE). Adapted from Maier-Hein et al. [238].

	WCK		EC
❕	Self-dual (for symmetric costs)	❕	Only Sensitivity perspective
✖	Limited interpretability	✔	Good interpretability with NEC
✔	Widely used	❕	Not widely used in biomedical image analysis
✖	Lack of framework to identify the optimal decision rule	✔	Optimal decision rule is straightforward
✖	Possibility of paradoxical results		

Both WCK (see Def. 2.30) and EC (see Def. 2.23) are metrics that allow for incorporating confusion costs. Common use cases for this property are tasks with ordinal classes or diagnostic decisions with errors of varying clinical severity. When deciding between the two metrics, the following properties are of relevance.

Symmetry Importantly, historically WCK was originally proposed as 'interrater agreement score', i. e., to compare the decisions of two raters, which is a symmetric problem by nature. In words of the notation introduced in Def. 2.37 this means, that given the confusion costs are symmetric then $\mathbf{WCK} = \mathbf{WCK}^T$, i. e., WCK is self-dual. Hence, unlike EC that relies solely on the Sensitivity perspective (see Prop. 2.25), WCK does not conceptually distinct reference from prediction.

Interpretability While both metrics can be interpreted as 'measures of (dis)agreement', the main difference is the fact that WCK provides this measure in reference to 'agreement by chance'. The equivalent concept for EC is its normalized variant NEC, where the disagreement measure is divided by a 'naive performance' measure (see Def. 2.27). Due to the conceptual similarity, it is more sensible to compare WCK to NEC. Both metrics

are prevalence-dependent due to relating model performance to a random performance reference. Their main difference is the definition of the 'random reference': In NEC this reference is straightforward to interpret as the 'best possible naive classification system' which always predicts the most dominant class (see Def. 2.26). The definition in WCK stems from its symmetric concept to compare the predictions of two raters. The random reference in this case is the probability of both raters agreeing by chance. Using this definition in classification tasks results in random reference systems that can be weaker than the naive system of NEC. Thus, the random reference in WCK is less intuitive and arguably also less useful in classification tasks [68, 92].

Undesired behavior in practice Using WCK with quadratic weights, often done for ordinal tasks, has been found to lead to 'paradoxical results' [414].

Popularity CK is widely used in the biomedical domain [240], and WCK in particular whenever customized penalties for class confusion are required. EC, on the other hand, is currently mostly found either in statistical textbooks [33, 157] or in non-related domains such as speech recognition [49, 100, 395], with few mentionings in the medical context [19, 206].

Decision rule EC comes with a comprehensive theoretical foundation based on Bayesian decision theory [113]. As a consequence, it is possible to analytically derive the optimal decision rule applied to the predicted class scores (see Def. 6.9). This is an important property in this context because the standard argmax-based decision rule is even for calibrated models unlikely to be optimal in scenarios with unequal costs of misclassifications.

Recommendation Due to the favorable properties, we generally recommend the usage of EC rather than WCK.

DG2.2: Balanced Accuracy (BA) versus Expected Cost (EC)

When deciding between BA and EC in the provided context, two primary scenarios should be distinguished:

Each class should contribute equally to the metric In this case, compensation for potential class imbalance is required in order to ensure equal contribution from each class. Here, we recommend BA as metric because it was designed for exactly this purpose. Although EC can be configured to be identical to one minus BA (see Tab. 4.1), we favor BA due to its widespread use.

Table 4.5: Comparison of Balanced Accuracy (BA) to Expected Cost (EC) in the context of decision guide DG2.2. Context: Equal severity of class confusions (FP2.5.2 = FALSE), either (i) unequal interest across classes (FP2.5.1 = TRUE) and no mismatch between class prevalences and class importance (FP2.5.3 = FALSE), or (ii) equal interest across classes (FP2.5.1 = FALSE) and provided class prevalences do not reflect the population of interest (FP4.2 = FALSE). Adapted from Maier-Hein et al. [238].

	BA		EC
○	Prevalence independence	✓	Possibility of reflecting expected prevalences in the target population
✓	Widely used	○	Not commonly known in biomedical image analysis

Classes should contribute according to prevalence in the *target* application Although the user may have an inherently equal interest in all classes (FP2.5.1 = FALSE), reporting a metric score to which all classes contribute equally may *not* necessarily be desired. Instead, the user may simply be interested in the overall performance on a given task and thus want classes to contribute according to their *prevalence* in the target application. This is not straightforward in the provided scenario because the data set at hand does not match the prevalences of the target population (FP4.2 = FALSE). In this case, we recommend EC, because it offers to explicit set (expected) class prevalences directly in the formula[8]. This strategy allows getting a glimpse of model performance on the target application while validating on the data at hand. Application of EC in this way, however, is only possible if the prevalences can be specified upfront. Depending on the use case one might argue that class priors being known upfront is quite a strong assumption.

DG2.3: Balanced Accuracy (BA) versus Matthews Correlation Coefficient (MCC) versus Normalized Expected Cost (NEC)

Three metrics are particularly attractive when class prevalences reflect the population of interest but compensation for class imbalance is desired (FP4.1 = TRUE and FP2.5.5 = TRUE). These are MCC, BA, and the normalized variant of EC, NEC. As described in Sec. 4.1.3 (FP2.5.5 *Compensation for class imbalance requested*), there are three effects of class imbalance that can be compensated for.

- **Effect 1:** Misleading metric values due to missing reference value for naive classifier
- **Effect 2:** Misleading metric values due to unequal contribution of classes

[8] This is exactly what we propose in step 3 of our workflow to mitigate prevalence shifts during model deployment (see Sec. 6.1.3).

Table 4.6: Comparison of Balanced Accuracy (BA) to Matthews Correlation Coefficient (MCC) to Normalized Expected Cost (NEC) in the context of decision guide DG2.3. Context: Equal severity of class confusions (FP2.5.2 = FALSE), either (1) unequal interest across classes (FP2.5.1 = TRUE) and no mismatch between class prevalences and class importance (FP2.5.3 = FALSE) or (2) equal interest across classes (FP2.5.1 = FALSE), provided class prevalences reflect the population of interest (FP4.2 = TRUE), presence of class imbalance (FP4.1 = TRUE) and compensation for class imbalances requested (FP2.5.5 = TRUE). Adapted from Maier-Hein et al. [238].

	BA		MCC		NEC
✓	Interpretable with respect to naive classifier	✓	Interpretable with respect to naive classifier	✓	Interpretable with respect to naive classifier
!	Implication of equal class contribution	!	Implication of equal class contribution	!	No setting of equal class contribution
!	Insensitive to predictive values	✓	High scores ensure high predictive values	!	Limited sensitivity to predictive values
✓	Optimal decision rule is straightforward	!	Lack of framework to identify the optimal decision rule	✓	Optimal decision rule is straightforward
✓	Good interpretability	!	Limited interpretability	✓	Good interpretability
✓	Widely used	✓	Fairly well-known	!	Not widely used in biomedical image analysis

- **Effect 3:** Misleading metric values due to missing consideration of predictive values

While the most common multiclass metric, AC, is subject to all three pitfalls when used in imbalanced settings, this decision guide discusses the three aforementioned alternatives (BA, MCC, and NEC) that compensate for one or more of these effects. The following aspects are relevant when deciding between the three:

Compensating for Effect 1 All three metrics establish a fixed score for the performance of a naive classifier, i.e., one that always predicts the most frequent class (see Def. 2.26) – which is a more realistic baseline in class imbalanced scenarios – compared to an entirely random system. The corresponding scores are 0 for MCC, 1 for NEC, and $1/C$ for BA, where C is the number of classes. However, the nature of the different compensation methods is fundamentally different.

Example 4.1. Consider the following confusion matrix of a binary classification system:

$$\boldsymbol{A} := \begin{bmatrix} 100 & 1 \\ 100 & 10000 \end{bmatrix}$$

The respective metric values are **BA** $= 0.99$, **MCC** $= 0.7$, and **NEC** $= 1$.
Although all metrics feature fixed values for a random classifier, the same system can be assessed differently, as it is being considered 'near-perfect' by BA (both TPR and TNR are high), 'fairly good' by MCC (high TPR, TNR and NPV, only low PPV), and 'random' by NEC (as predicting only class 2 would yield the same number of 101 wrong predictions).

Intuitively, the BA assessment in Ex. 4.2.2 seems overly optimistic, which can be attributed to the fact that BA does not compensate for Effect 3, as described in more detail below. On the other hand, the NEC assessment in Ex. 4.2.2 appears overly strict, which can be attributed to the fact that NEC does not compensate for Effect 2 as described in more detail below.

Compensating for Effect 2 In balanced scenarios, all classes are weighted equally by common discrimination metrics. In contrast, in imbalanced scenarios, common metrics such as AC are dominated by the frequent classes. Equal contribution of classes in this context would imply that each class can contribute equally to the final metric score, irrespective of prevalence. This is exactly what BA does by computing the average of individual class Sensitivities. An alternative way of thinking about this compensation is tweaking the costs of misclassification errors by assigning higher costs for errors in rare classes and vice versa. Hence, BA can be thought of as a cost instantiation of EC if the costs are set proportional to the inverse of class prevalences[9]. Importantly, the normalized variant of EC, NEC, does not generally compensate for Effect 2, but merely rescales metric scores in a way that the value of 1 corresponds to a naive classifier always predicting the most frequent class (see Effect 1). In other words, the rankings obtained for a set of test cases would be the same for EC and NEC. Analogously to EC, it is also possible to tweak the costs to compensate for Effect 2 in NEC, but the resulting metric would yield no advantages over BA. Importantly, the fact that NEC does not compensate for Effect 2 implies that if there is an unequal interest across classes (FP2.5.1 = TRUE), then NEC is the only correct choice. Analogously to BA, MCC establishes equal contribution of classes by assessing individual class sensitivities.

Compensating for Effect 3 The predictive values (PPV and NPV) are an important aspect of assessing the quality of a classification system. To showcase this importance, we give the following example.

[9] This is indicated in Tab. 4.1, for a full proof see Ferrer [113].

Example 4.2. Consider the following confusion matrix of a binary classification system:

$$\boldsymbol{A} := \begin{bmatrix} 10 & 1 \\ 100 & 10000 \end{bmatrix}$$

The respective metric values are **BA** $= 0.95$, **MCC** $= 0.29$, and **NEC** $= 9.2$. This system is assessed as 'near-perfect' by BA (both TPR and TNR are high), 'better than random, but not really useful' by MCC (high TPR, TNR and NPV, very low PPV), and 'much worse than random' by NEC (as predicting only class 2 would yield the much lower number of only 11 wrong predictions).

Ex. 4.2.2 shows that BA does not consider predictive values, thus yielding a near-perfect score despite a low PPV of 0.09. This assessment could be considered a pitfall in many scenarios, where the classification system would be fairly useless. Consider, for instance, a breast cancer screening program where, based on the provided system, **FDR** $= 100/(10 + 100) \approx 90\%$ of all biopsies would be unnecessary.

In contrast, the MCC score could be considered intuitive for many scenarios such as the screening example. This is due to MCC explicitly considering all four basic rates TPR, TNR, PPV, and NPV (see Cor. 2.52). Thus, MCC poses further requirements compared to BA, which focuses only on Sensitivities. NEC also ensures high predictive values by design. In practice, however, it is not always a good indicator for predictive values because of the sometimes overly strict penalization of errors, as seen in the above example. In theory the weights in NEC could be adjusted to simulate the behavior of predictive value-sensitive metrics like MCC, but this implies a trial-and-error tuning process on each new task.

Identifying the optimal decision rule The different strategies for identifying a decision rule applied to predicted class scores are described in Sec. 4.1.3 FP2.6. In the multiclass setting, argmax-based decisions are very common (see Sec. 6.2.1), but make arguably strong assumptions such as calibrated scores and equal penalization of all misclassifications (see Prop. 6.10). Noteworthy, some metrics can be viewed as instantiations of EC (see Tab. 4.2, in this case BA and NEC), which comes with a theoretical framework on how to choose the decision rule [113]. MCC, on the other hand, lacks such a framework.

Interpretability Arguably, BA features the most straightforward interpretation as the average over individual class Sensitivities, with bounded scores $[0, 1]$ and a fixed random reference at $1/C$. NEC scores are also fairly interpretable ('the EC of the system in relation to the EC of the naive system'), but scores are not bounded $[0, \infty)$. Furthermore, the random reference could be interpreted as 'too strict' for many scenarios such as in Ex. 4.2.2. As for MCC, a random reference value is provided at 0 and the scores

are bounded [-1,1], but all intermediate scores are arguably less intuitive. The general interpretation of MCC would be that it is a metric that depends on individual class Sensitivities and predictive values, i. e., a high MCC score guarantees all of these being high and a low MCC score indicates that at least one of them is low [64].

Popularity BA and MCC are fairly well-known (see Tab. 6.1). NEC is used prominently in the field of speaker verification but has not been introduced to the biomedical imaging or clinical community yet, although the statistical concepts it is based upon are long-standing in Bayesian decision theory.

DG3.1: Metric@(TargetMetric = TargetValue)

If a target value for a specific metric (typically TPR) is provided, the decision rule applied to the predicted class scores is optimized such that the specific target value is reached on a validation data set (see Sec. 4.1.3 FP2.6). Other metrics, depending on the target application, can then be reported for that specific threshold. In some cases, e. g., TNR@(TPR = 0.95) the corresponding value can directly been read from the ROC-curve. Possible candidates include TPR, TNR, PPV, and NPV.

DG3.2: Net Benefit (NB) versus Expected Cost (EC)

Table 4.7: Comparison of Net Benefit (NB) to Expected Cost (EC) in the context of decision guide DG3.2. Context: FP2.6 = cost-benefit-based decision rule applied to predicted class scores requested. Adapted from Maier-Hein et al. [238].

	NB		EC
!	Decisions can be defined directly based on predicted class scores, interpreted as risks	!	Decisions based on explicit definition of misclassification costs
!	Weighting of True Positive (TP) against False Positive (FP) in risk perspective	!	Weighting of False Positive (FP) against False Negative (FN) in cost perspective
!	Lack of framework to validate the decision rule applied to class scores	✓	Availability of framework to validate the decision rule applied to class scores
✓	Focus on reflectance of the (e. g., clinical) interest in the scores	✓	Inherent interpretability with respect to naive classifier
!	Popular metric in clinical studies but not common in image analysis	!	Not commonly known in biomedical image analysis

This decision guide is embedded in the framework in subprocess **S3**, which guides the selection of metrics that are reported separately for each class. In multiclass tasks this reporting amounts to a one-versus-rest validation scheme (see Def. 2.15). However, this scheme is not intuitively applicable to a cost-benefit analysis (what are the costs and benefits of the 'rest' class?), which is the concept behind decision rules of both metrics in this decision guide. Thus, for multiclass tasks we recommend to only proceed with the metrics selected in subprocess **S2** (e. g., EC or WCK) and not select any further metrics here to be reported in a one-versus-rest fashion, i. e., we recommend skipping the guide. Since the task is binary, we will work with the threshold operator and a cutoff τ as decision rule (see Def. 2.9).

Both NB and EC are linked to cost-benefit analysis [289] and are well-suited when a cost-benefit-based approach for determining an appropriate decision rule applied to the predicted class scores is desired (FP2.6 = cost-benefit-based). To this end, both require the knowledge of task-dependent trade-offs between benefits and costs, as detailed below. The following aspects are relevant when deciding between EC and NB:

Cost versus risk perspective *Cost perspective:* For EC, explicit costs for both basic misclassifications (FP, FN) need to be defined or estimated. The optimal threshold that minimizes these costs can be analytically determined without data-based optimization. *Risk perspective:* In contrast, NB does not require the costs to be defined explicitly. Instead, predicted class scores are interpreted as probabilities or 'risks' of certain model output scores belonging to the positive class and the cutoff on the scores is defined directly on this scale based on task interest (e. g., 'only treat patients with cancer risk >10%'). This can be interpreted as an implicit cost-benefit analysis resulting in a single intuitive risk score. However, it is also common for NB to make this cost-benefit analysis more explicit and define the risk as a relation of the benefit of TPs to the harms caused by FPs. A diagnostic test, for example, may lead to early identification and treatment of a disease, but typically the process will also cause some patients without disease being subjected to unnecessary further interventions. NB allows to consider such trade-offs by putting the benefits and harms of the test on the same scale so that they can be directly compared. A physician may, for example, state that 10 FPs, resulting in unnecessary biopsies, are acceptable to find one more cancer case (TP).

Decision curves In most scenarios it is not possible to precisely define the costs or risks associated with the task. For example, it is not straightforward to make an exact decision on how many FPs would be acceptable to obtain one more TP. To compensate for this uncertainty, it is common practice to plot NB over a 'reasonable range of risk thresholds' resulting in so-called decision curves (see Def. 2.57). This analysis allows assessing and comparing methods according to their NB scores without relying on a single cutoff. Although not common practice, one could also generate such curves for EC when expressing cost ratios as a risk score (i. e., switching from the cost to the risk

perspective).

Cutoff on predicted class scores In NB, the cutoff is determined directly from provided knowledge about the task and does not require data-based optimization. In contrast, EC allows to alternatively determine a data-based cutoff by minimizing EC on a dedicated data split, if available. A further difference between the two metrics is the way prevalence dependency is handled: EC isolates the class priors from the predicted class probabilities and defines them as a parameter of the cutoff itself, such that all application dependent parameters (costs and class priors) are part of the cutoff (see Def. 6.9). Upon deployment of a model on a new data set, the threshold can simply be updated analytically. Note that this process only works under the arguably strong assumption that the class priors of the new data set are known[10]. In contrast, NB considers risk scores that incorporate the class priors, implying that the threshold depends solely on the cost-benefit trade-off. As a consequence, when the class priors shift on a new data set, the risk-cutoff in NB requires predicted class probabilities to be re-calibrated[11]. The latter might be a harder requirement because it requires a labeled validation set for re-calibration as opposed to requiring merely the class priors of the new data set for a threshold update.

Interpretability EC allows reporting a normalized version (NEC), which makes the metric scores interpretable with regard to the performance of a random classifier. In contrast, in NB, the reference to a random classifier is typically done manually (by comparing the two scores), because NB itself allows for an interpretation as the 'proportion of net-TP', which would get lost by normalization.

Calibration Both metrics rely on the fact that predicted class probabilities are well-calibrated with regard to a chosen cutoff. EC allows assessing this requirement by calculating the extra cost entailed by miscalibration (or the potential for reducing cost by calibrating scores) [113]. The calibration error here is measured as the increase of EC with the analytical, i. e., task interest-based, cutoff compared to an empirical cutoff optimized on the data. Compared to related calibration errors (see Sec. 2.6), this technique assesses a weaker calibration condition, which is directly targeted to the decision process at hand. For instance, even when assessing the relatively weak top-label calibration condition by means of ECE with two bins and the border at the determined cutoff value, the distribution inside the bins would be considered, while EC only focuses on how many more cases would have been on the 'correct side of the cutoff' if scores were calibrated, without considering score distributions on either side of the cutoff.

[10] We will present methods to determine such priors dynamically in Sec. 6.1.3.

[11] See Fig. 6.9 (bottom left) for experimental validation that the EC optimized decision threshold can be transferred analytically without re-calibration.

Popularity Neither NB nor EC are widely used in the biomedical image analysis community. NB is a popular metric in clinical studies, while EC itself is currently not used, but many of its instantiations (see Tab. 4.2).

DG3.3: Positive Likelihood Ratio (LR+) versus TPR

Table 4.8: Comparison of Positive Likelihood Ratio (LR+) to True Positive Rate (TPR) in the context of decision guide DG3.3. Context: FP2.6 = optimization- or argmax-based decision rule applied to predicted class scores requested and provided class prevalences do not reflect the population of interest (FP4.2 = FALSE). Adapted from Maier-Hein et al. [238].

	LR+		TPR
✅	Straightforward application in the case of an optimization-based decision rule (FP2.6)	❌	Challenging application in the case of an optimization-based decision rule (FP2.6)
✅	Interpretation often reflecting interest in binary tasks	✅	Good interpretability

This decision guide helps deciding between LR+ and TPR in the context of per-class validation (subprocess **S3**) with an optimization- or argmax-based decision rule applied to predicted class scores (FP2.6).

Interpretability In the provided context of this decision guide, where metrics are reported individually per class, TPR and LR+ convey similar information and there is no 'incorrect' choice. Thus, the choice between the two can generally be made as the metric that is easier to interpret in the given task: In binary classification tasks, LR+ conveys TPRs of both classes in a single score. Due to its intuitive and meaningful interpretation ('How much more likely is the occurrence of a class 1 prediction for a class 1 sample compared to a class 2 sample?'), it is often reported in clinical studies. In multiclass settings (which, in this context, amount to a one-versus-rest validation scheme), TPRs are generally easier to interpret, while the interpretation of LR+ might still be helpful (class 2 encompasses the 'rest').

Decision rule In case the decision rule applied to predicted class probabilities is to be determined on the basis of optimization on the target class, one additional consideration is of importance (FP2.6 = optimization-based decision rule). When reporting TPR per class, the decision rule can not be optimized based solely on the single TPR at hand because this would always yield a cutoff value of 1. LR+ naturally overcomes this problem. Other possible workarounds include choosing a different decision rule (FP2.6) or optimizing a weighted average over TPRs for all classes instead. The latter option should only be

considered if meaningful weights across classes can be defined (e. g., based on class importance).

DG3.4: Positive Likelihood Ratio (LR+) versus TPR versus F-beta Score

Table 4.9: Comparison of Positive Likelihood Ratio (LR+) to True Positive Rate (TPR) to F-beta Score in the context of decision guide DG3.4. Context: FP2.6 = optimization- or argmax-based decision rule applied to predicted class scores requested and provided class prevalences reflecting the population of interest (FP4.2 = TRUE). Adapted from Maier-Hein et al. [238].

LR+		TPR		F-beta Score	
✓	Meaningful interpretation in binary tasks	✓	Generally good interpretability	!	Limited interpretability
✓	Interpretable with respect to naive classifier	!	Interpretable with respect to naive classifier only when averaging over classes	✗	No interpretability with respect to naive classifier
✗	Insensitive to PPV	✗	Insensitive to PPV	✓	High scores ensures high PPV

In the context of this decision guide, prevalence dependency is not an exclusion criterion (see FP4.2) and thus F-beta Score can be considered as an alternative to Sensitivity-based metrics (TPR and LR+). Details for the decision between the latter are provided in **DG3.3**; the present guide focuses on the pros and cons of opting for F-beta Score.

Per-class validation is commonly performed in a one-versus-rest fashion, naturally introducing class imbalance into the validation. Exceptions are binary scenarios with two balanced classes. For this exception, no compensation for class imbalance is needed (FP2.5.5) and the choice between F-beta Score and Sensitivity-based metrics becomes less relevant, i. e., there are no obvious incorrect choices. Thus, the decision can be made on the basis of which metric is easier to interpret in a given task. For all other cases, the decision should be based on whether compensation for class imbalance is required (FP2.5.5 = TRUE).

Compensation for class imbalance As described in Sec. 4.1.3 FP2.5.5, there are three aspects of compensation for class imbalance:

(i) **Establishing a reference value for random performance:** LR+ provides a fixed random reference value at $\mathbf{LR}+ = 1$, while for TPR the scores of individual classes

can vary and only their average is fixed at $1/C$ (equivalent to BA). F-beta Score does not provide a reference value for random performance.

(ii) **Establishing equal class contribution:** In the provided context (**S3**), the validation is performed per class, such that this aspect is irrelevant.

(iii) **Establishing consideration of predictive values:** This aspect is the main reason to opt for F-beta Score in this decision guide, because it is the only metric of the three where high scores ensure a high PPV. In contrast, LR+ and TPR are insensitive to PPV, which, depending on the task interest, can substantially diminish their utility. An exemplary pitfall related to this choice is the confusion matrix of a binary classification task, as given in Ex. 4.2.2. This classification system yields $\mathbf{TPR}_1 \approx 0.91$, $\mathbf{TPR}_2 \approx 0.99$, $\mathbf{LR}{+}_1 \approx 91.82$ and $\mathbf{LR}{+}_2 \approx 10.89$. While $\mathbf{F1}_2 \approx 1$, the low value $\mathbf{F1}_1 \approx 0.17$ indicates a low PPV on class 1. This pitfall may be of practical relevance in class-imbalanced tasks where FPs shall not be neglected. For example, in breast cancer screening, the provided classifier would not be useful, since $\mathbf{FDR} = 100/(10 + 100) \approx 90\%$ of all biopsies would be unnecessary.

Interpretability Out of the three, TPR is arguably the easiest-to-interpret metric (exceptions are binary tasks, where LR+ might be preferable as detailed in **DG3.3**). F-beta Score can be interpreted as the harmonic mean of TPR and PPV, which adds a layer of complexity to the interpretation compared to TPR. Thus, if the aspects discussed in 'compensation for class imbalance' are not relevant, F-beta Score might not be the metric of choice.

DG3.5: How to determine β in F-beta Score

Table 4.10: Determining the hyperparameter of the F-beta Score in the context of decision guide DG3.5. Context: FP2.6 = optimization- or argmax-based decision rule applied to predicted class probabilities requested and provided class prevalences reflecting the population of interest (FP4.2 = TRUE). Adapted from Maier-Hein et al. [238].

β < **1**	β = **1**	β > **1**
Higher weighting of FP penalties (i. e., PPV)	Harmonic mean of PPV and TPR	Higher weighting of FN penalties (i. e., TPR)

The most common choice is to set $\beta = 1$, resulting in equal weighting of FP and FN penalties (see Cor. 2.46). If unequal penalization of class confusions is desired (FP2.5.2), higher values of β result in higher weights on FN penalties compared to FP penalties and thus imply a focus on TPR compared to PPV [336].

DG4.1: Area under the Receiver Operating Characteristic Curve (AUROC) versus Average Precision (AP)

Table 4.11: Comparison of Area under the Receiver Operating Characteristic Curve (AUROC) to Average Precision (AP) in the context of decision guide DG4.1. Context: availability of predicted class scores (FP5.1 = TRUE) and provided class prevalences reflecting the population of interest. Adapted from Maier-Hein et al. [238].

AUROC		AP	
⚠	Insensitive to PPV under class imbalance	✓	High scores ensure high PPV including under class imbalance
✓	Interpretable with respect to naive classifier	⚠	Prevalence-dependent reference value for naive classifier
✓	Straightforward interpretability	⚠	Limited interpretability

The comparison between the two concepts behind AUROC and AP, i. e., the comparison between ROC curves and PR curves has been extensively studied [87]. In practice, the choice between the two metrics boils down to the following aspects (if no clear choice can be made, we recommend reporting both metrics):

Compensation for class imbalance effects Of relevance in the context of this decision guide is the third pitfall from FP2.5.5 (see Sec. 4.1.3): *Misleading metric values due to missing consideration of predictive values.* AUROC is based on the TPRs of the two classes and does not consider predictive values. In class-imbalanced scenarios, this may lead to near-perfect AUROC scores that hide the fact that a system might have limited to no predictive utility. AP assesses PPV and thus compensates for the undesired effects caused by class imbalance. A technical explanation is given by the fact that a high number of TNs dominates and suppresses the FPs in the calculation of the TNR, thus yielding high scores for AUROC. A practical example for this pitfall might be a breast screening program, where a high PPV is of great importance to prevent unnecessary biopsies (FPs). The focus of AP on a particular class further has the effect that the resulting scores differ depending on which of the two classes is being inspected. This is in contrast to AUROC, which yields the same scores irrespective of this perspective. The common approach for AP-based assessment in class-imbalanced scenarios is to define the rare class as the first class. The fact that AP focuses on this class reflects the task interest of not letting rare (important) events be dominated by frequent events in the metric score.

Interpretability AUROC is easy to interpret as it simply represents the probability of a randomly sampled positive case having a higher predicted class score than a randomly sampled negative case. It further comes with a fixed reference value for the performance

of a random classifier at 0.5. AP, on the other hand, is harder to interpret and features no fixed random reference value. Instead, the AP score of a random classifier is the prevalence of the positive class which varies on each data set.

Implementations The PR curve is more complex to interpolate compared to the ROC curve [87], which has led to the existence of various implementations of AP, whereas no such heterogeneity exist for AUROC.

Popularity Although AUROC is the common choice for multi-threshold metrics, AP is also widely known and used.

DG5.1: Kernel Calibration Error (KCE) versus Expected Calibration Error Kernel Density Estimate (ECE^{KDE})

Table 4.12: Comparison of Kernel Calibration Error (KCE) to Expected Calibration Error Kernel Density Estimate (ECE^{KDE}) in the context of decision guide DG5.1. Context: FP2.7.2 = **U2** - *comparison of calibration performance across classifiers on the same task requested* and no mismatch between class prevalences and class importance (F2.5.3 = FALSE). Adapted from Maier-Hein et al. [238].

KCE		ECE^{KDE}	
✅	Capture of isolated calibration quality	✅	Capture of isolated calibration quality
❗	Unbiased estimator of canonical calibration error based on an alternative distance function	❗	Potentially biased estimator of an ℓ_p canonical calibration error (bias might be rendered neglectable by future de-biasing schemes)
❌	Bad interpretability, also due to negative output values	✅	Straightforward interpretability of relative improvement
❗	Recent proposition, not widely used	❗	Recent proposition, not widely used
❌	Depends on nontrivial configuration choices of kernels and associated hyperparameters		

The context for this decision guide between KCE and ECE^{KDE} is use case **U2** in Fig. 4.6: 'comparing the calibration quality across multiple classifiers on the same task.'

General differences Both KCE and ECE^{KDE} are estimators of a canonical calibration error, but measure this error based on different divergences, i. e., distance functions: ECE^{KDE} is based on the L_p norm and thus straightforward to interpret and configure. In contrast, KCE is based on the 'maximum mean discrepancy' and thus not interpretable (it may even take on negative values) and requires nontrivial configuration of kernels as well as associated hyperparameters. On the other hand, L_p norm estimators such as ECE^{KDE} are inherently biased while KCE is an unbiased estimator. Arguably, in the context of this decision guide (**U2**), interpretability of the calibration error estimate is not required, since only a comparative, or relative assessment is requested rendering the unbiased KCE the intuitive choice. However, recent research on L_p norm estimators presents effective de-biasing schemes [302], which might render the resulting bias neglectable in the near future and thus make L_p estimators such as ECE^{KDE} a viable alternative for comparative calibration assessment.

Popularity Calibration error estimates KCE and ECE^{KDE} are both recently proposed measures that are not widely known in the biomedical community.

DG5.2: Brier Score (BS) versus Kernel Calibration Error (KCE) versus Expected Calibration Error Kernel Density Estimate (ECE^{KDE})

The context for this decision guide between BS, ECE^{KDE}, and KCEis use case **U1** in Fig. 4.6: 'comparing the effect of one or more re-calibration methods on the same (fixed) classifier.'

General differences BS can be decomposed into discrimination and calibration terms, where the calibration term exactly resembles the canonical calibration error (see Thm. 2.72). As the purpose of the metric in the provided context is to assess the performance of different re-calibration methods for the same classifier, a higher BS score also implies a better calibration in the case of *accuracy-preserving*[12] calibration methods. As a major difference to BS, KCE estimates the canonical calibration error directly. While this estimation is not biased, the resulting estimates are not interpretable, that is, they only allow for relative comparison on the same task (equivalently to BS). Further, KCE requires nontrivial configuration of kernels as well as associated hyperparameters. In contrast to KCE, current estimators of L_p calibration error are biased, but are highly interpretable and straightforward to configure. Moreover, recent developments in this line of research present effective de-biasing schemes [302], which might render the resulting bias neglectable in the near future and thus make L_p estimators such as ECE^{KDE} a viable alternative also for comparative calibration assessment.

[12] Although the common terminology is accuracy-preserving [387], the more appropriate formulation is 'ranking-preserving' in the sense that the re-calibration does not change the ranking (i. e., descending order) of logits. Because such a method would lead to the same decisions under the argmax operator it will ultimately also preserve the same confusion matrix, hence AC.

Table 4.13: Comparison of Brier Score (BS) to Kernel Calibration Error (KCE) to Expected Calibration Error Kernel Density Estimate (ECE^{KDE}) in the context of decision guide DG5.2. Context: **U1** - FP2.7.2 = comparison of re-calibration methods for the same classifier requested. Adapted from Maier-Hein et al. [238].

BS	KCE	ECE^{KDE}
Capture of effects of (re-) calibration methods on discrimination performance in addition to calibration quality	Capture of isolated calibration quality	Capture of isolated calibration quality
Unbiased measure of an L_2 norm canonical calibration error	Unbiased estimator of canonical calibration error based on an alternative distance function	Potentially biased estimator of an L_p canonical calibration error (bias might be rendered neglectable by future de-biasing schemes)
Straightforward interpretability of relative improvement	Bad interpretability, also due to negative output values	Straightforward interpretability of relative improvement
Established statistical concept with long history of applications in many fields of research	Recent proposition, not widely used	Recent proposition, not widely used
	Depends on non-trivial configuration choices of kernels and associated hyperparameters	

Applicability Generally, BS is attractive for ranking re-calibration methods that are guaranteed to be accuracy-preserving (such as the common temperature scaling (see Def. 2.67)). Otherwise, the metric must be applied with care, because altered discrimination performance will dilute the focus on calibration quality in the ranking. Note that it may also be desirable to capture the effect of (non-accuracy-preserving) re-calibration methods on the discrimination performance. In such cases of comprehensive assessment

of re-calibration methods, it is also appropriate to apply BS. In contrast to BS, calibration error estimators such as KCE and ECE^{KDE} are capable of comparing the calibration error of re-calibration while being agnostic to potential changes of discrimination performance caused by the transformations. For the provided use case, this property allows the ranking of non-accuracy-preserving transformations, such as recently proposed techniques employing spline interpolations [152] or Gaussian processes [416], purely according to their calibration error while ignoring their effects on the discrimination performance.

Interpretability Defined as the root mean square error between predictions and references, BS is bounded by $[0, 2]$ and therefore straightforward to interpret as an overall measure. However, as the calibration error is not isolated and scores are still conflated with the (same fixed) discrimination performance, only a relative comparison of calibration errors is possible. KCE is generally hard to interpret, also because it can yield negative values. ECE^{KDE} as an estimator of L_p calibration error is straightforward to interpret.

Popularity BS is a widely known metric for overall performance measures with a long history of usage. Calibration error estimates KCE and ECE^{KDE} are both recently proposed measures and not widely known in the biomedical community.

Reasons to not recommend NLL in this context NLL essentially assesses a weighted version of the canonical calibration error as the logarithm leads to heavy penalization of tail probabilities. As the implications of this weighting on calibration assessment (as opposed to the overall performance measure) are not intuitive, we generally do not recommend NLL in this use case.

DG5.3: Brier Score (BS) versus Negative Log Likelihood (NLL)

The context for this decision guide between BS and NLL is use case **U3** in Fig. 4.6: 'overall performance measure requested.' Both BS and NLL are overall performance measures, which capture discrimination and canonical calibration in a single score.

Penalization of errors Like AC, BS penalizes errors of all events equivalently irrespective of the class prevalence. This implies that scores may drastically change when the prevalence changes and thus renders BS a highly prevalence-dependent metric. For instance, in imbalanced scenarios, a naive system that simply predicts the dominant class can receive a low BS, similarly to a high AC or low EC. One strategy to cope with this is to divide the BS by the BS achieved with a naive system, resulting in the normalized variant BSS. Equivalently to NEC, this transformation is a rescaling of scores to establish a 'naive' baseline and enhance interpretability, but errors are still penalized equivalently irrespective of class prevalence. In other words, equal importance of classes (FP2.5.1) is

Table 4.14: Comparison of Brier Score (BS) to Negative Log Likelihood (NLL) in the context of decision guide DG5.3. Context: FP2.7.2 = **U3** - *comparison of overall performance across classifiers requested.* Adapted from Maier-Hein et al. [238].

	BS		NLL
✗	Bounded penalization of errors leads to preference of naive systems in imbalanced settings	!	Heavy penalization of extreme scores (close to 0 or 1), thus ability to capture missing rare events. General preference of conservative models
✓	Straightforward interpretability as the mean squared error	!	Difficult interpretability due to lack of upper bound
✓	Established statistical concept with long history of applications in many fields of research	✓	Established statistical concept with long history of applications in many fields of research

not reflected in the metric, and missing a frequent event is still as heavily penalized as missing a rare event although missing a rare event has a greater effect on the respective class TPR. This results in a strict interpretation where the total amount of errors has to be lower than the number of events in the rare class in order for a system to be considered 'better than random'.

Compared to squared error penalization in BS, the logarithm introduces a stronger penalization of tail probabilities [307]. In consequence, overconfident predictions (probabilities close to one) lead to higher losses. For example, predicting probability 0.999 rather than 0.99 on an incorrect class increases BS by $\approx 2\%$ and NLL by $\approx 230\%$ (for this single entry). A practical effect of this penalty is a naturally higher penalization of naive systems in class imbalance scenarios, addressing the pitfall of BS above. NLL is thus of potential interest in scenarios with high class imbalance, where missing rare events would be heavily penalized, compared to BS which is prone to favoring naive systems. Generally, the penalization effect can also be described as NLL favoring more conservative models that avoid predictions of extreme class scores.

Interpretability BS is relatively straightforward to interpret as the mean squared error between predictions and the reference. The resulting scores are bounded ($[0, 2]$). NLL is arguably harder to interpret featuring logarithmic penalization of errors and thus no upper bound of the resulting score (bounds: $[0, \infty]$).

Popularity Both metrics are common statistical concepts and come with a long history of usage in many fields of research.

DG5.4: Expected Calibration Error (ECE)/ Root Brier Score (RBS) versus Class-wise Calibration Error (CWCE) versus Expected Calibration Error Kernel Density Estimate (ECE^{KDE})/ Class-wise Calibration Error (CWCE)/ Root Brier Score (RBS)

The decision between the sets of metrics boils down to determining whether predicted class probabilities should be tested for top-label calibration (as measured by ECE), marginal calibration (as measured by CWCE), or canonical calibration (as measured by ECE^{KDE}) – see Def. 2.59. If there is an unequal interest across classes (FP2.5.1), CWCE is the natural choice. In this case we recommend both per-class and weighted reporting (by class importance). Note that only aggregated reporting comes with the pitfall of unstable results, specifically in the case of few samples or many classes. In the case of equal interest across classes, the key question is whether the task interest is limited to the predicted probabilities that lead to the classification decision (top-label) or whether there are reasons to request all predicted probabilities to be calibrated.

Notably, in binary classification tasks, the two conditions are equivalent [393].

Reasons for and against focusing on top-label calibration (ECE) In case the underlying biomedical research question has a dedicated focus on the decision process, top-label error might be the right choice, because it directly reflects this focus. Conflating the calibration of decisions with other probabilities might be interpreted as washing out the task focus in this case. Although it is common practice to assess calibration quality with ECE, this approach comes with various pitfalls. Importantly, it is often ignored that top-label calibration implies an argmax decision rule based on the predicted class scores, which is often not an optimal decision rule (see Fig. 6.8). Caution should also be exercised if there is a mismatch between class prevalences and class importance (FP2.5.3) as the top-label calibration is highly biased towards the high-prevalence classes. Furthermore, ECE commonly relies on binning of class scores, which introduces a dependency of the resulting metric score on the specific binning scheme. The number of bins is a configuration parameter that should by no means be optimized on the final validation data. Note in this context that binning has been shown to result in a more biased estimation compared to density estimation methods [302].

Reasons to extend the focus to all predicted scores (ECE^{KDE} and CWCE) A common perception is that the canonical calibration condition, which is the strongest condition considering all predicted class scores, is the appropriate one in many application scenarios [113, 150, 302]. One reason lies in the limitations of top-label calibration and associated binning estimators described above. Another reason could be a broad task interest in all predictions beyond the classification decision. In the clinical context, for instance, the risk for all potential outcomes might be relevant for further treatment or shall be communicated to the patient. In such scenarios, calibration of all probabilities might be of interest. Consider, for instance, a multiclass classification of tumor categories,

where one category is more aggressive than others. Even though the final prediction of the system is 'benign lesion', it might be of clinical interest to know (and communicate to the patient) whether the probability for this outcome was 5% or 20%. While the primary calibration metric for such scenarios should be ECE^{KDE} as an estimate of the canonical calibration, it might be of interest to additionally report marginal calibration (as measured by CWCE) separately for each class. Notably, for these scenarios, alternatively splitting the problem into individual domain questions that result in separate traversals for each class of interest should be considered.

Additional reporting of RBS as a guaranteed upper bound on the calibration error In top-label and canonical calibration, we recommend the additional reporting of RBS as a guaranteed upper bound on the calibration error. As popular methods to assess calibration quality such as ECE or ECE^{KDE} are known to over- or underestimate the error [150], this guarantee provides additional information, especially in safety-critical applications where the calibration error must not be underestimated.

4.2.3 Instantiations

We instantiated the framework for several biological and medical image analysis use cases that were identified by the *biomedical expert group*. The resulting metric recommendations are summarized in Fig. 4.13, while Figs. 4.14-4.17 provide a detailed overview of the paths the use cases traversed in the recommendation subprocesses **S2-S5**. Noteworthy, and similar to the task pool presented in Tab. 2.1, the tasks cover a variety of imaging modalities (microscopy, dermoscopy, sonography, MRI, and X-ray). The selected scenarios are:

ImLC-1 Frame-based sperm motility classification from microscopy time-lapse video of human spermatozoa [160]

ImLC-2 Disease classification in dermoscopic images [72, 391]

ImLC-3 Classification of the overall autophagy stage for a collection of cells [266, 446]

ImLC-4 Diagnostic standard plane classification in ultrasound images [24]

ImLC-5 Identification of new lesions in brain multi-modal MRI images of patients with multiple sclerosis [78, 204]

ImLC-6 Breast cancer classification in mammography images [218]

ImLC-7 Multiclass cardiac disease classification in MRI images [30]

The choice of use cases was influenced by two particular goals. First, **ImLC-1** and **ImLC-2** were chosen such that two scenarios with very different modality and application

ultimately lead to the same metric recommendations. This was to demonstrate that our recommendation process abstract the *problem fingerprints* enough to be broadly applicable. Second, the remaining scenarios **ImLC-3 - ImLC-7** were chosen to cover a diverse set of applications and modalities, and follow different branches in our recommendation process. This was to demonstrate the sufficient coverage of use case by our recommendation process.

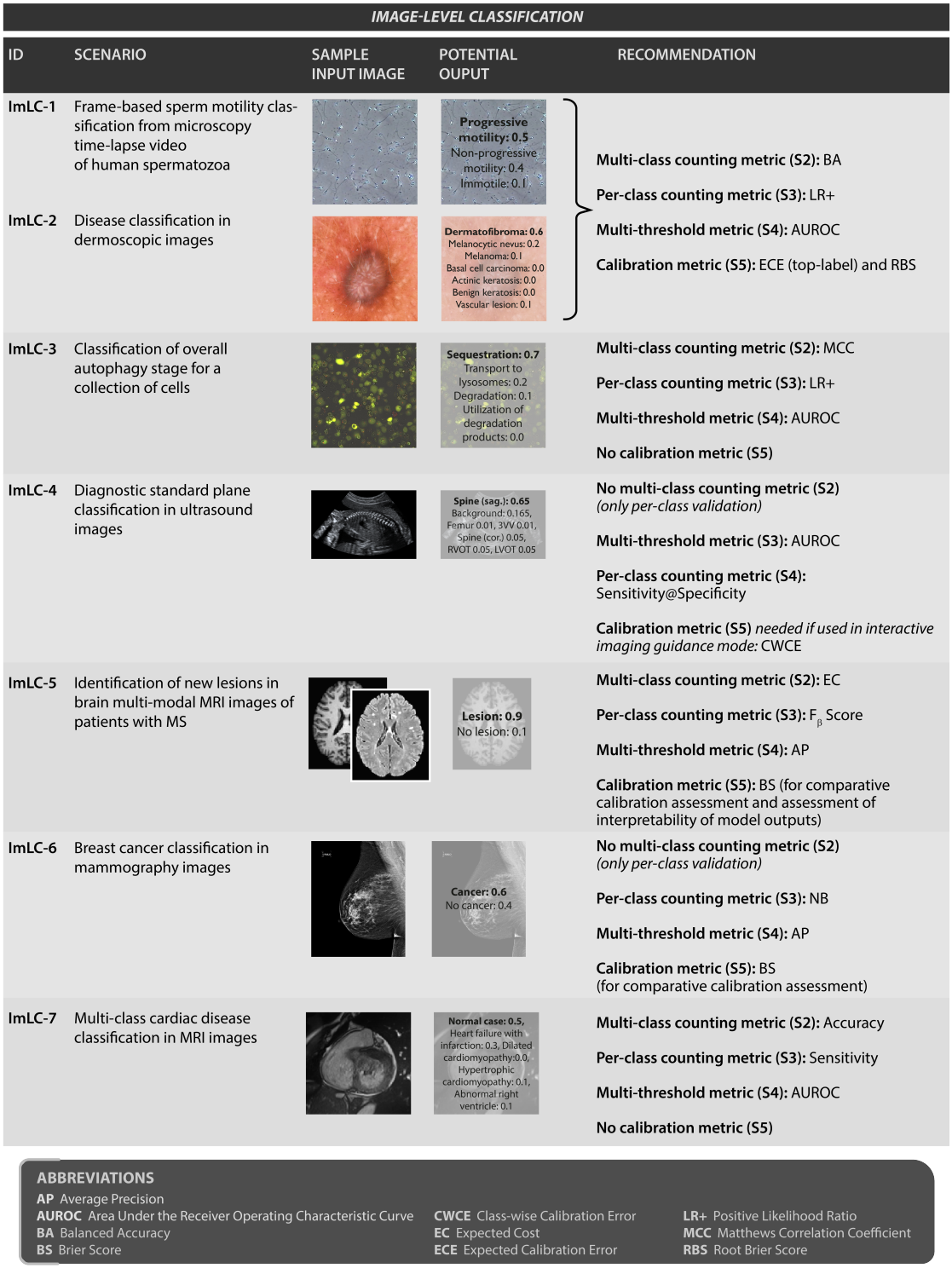


Figure 4.13: Metric recommendations for seven concrete biomedical ImLC problems. The seven use cases **(ImLC-1)** - **(ImLC-7)** with example images, example labels and the resulting metric recommendations after following our proposed process. Originally published in Maier-Hein et al. [238].

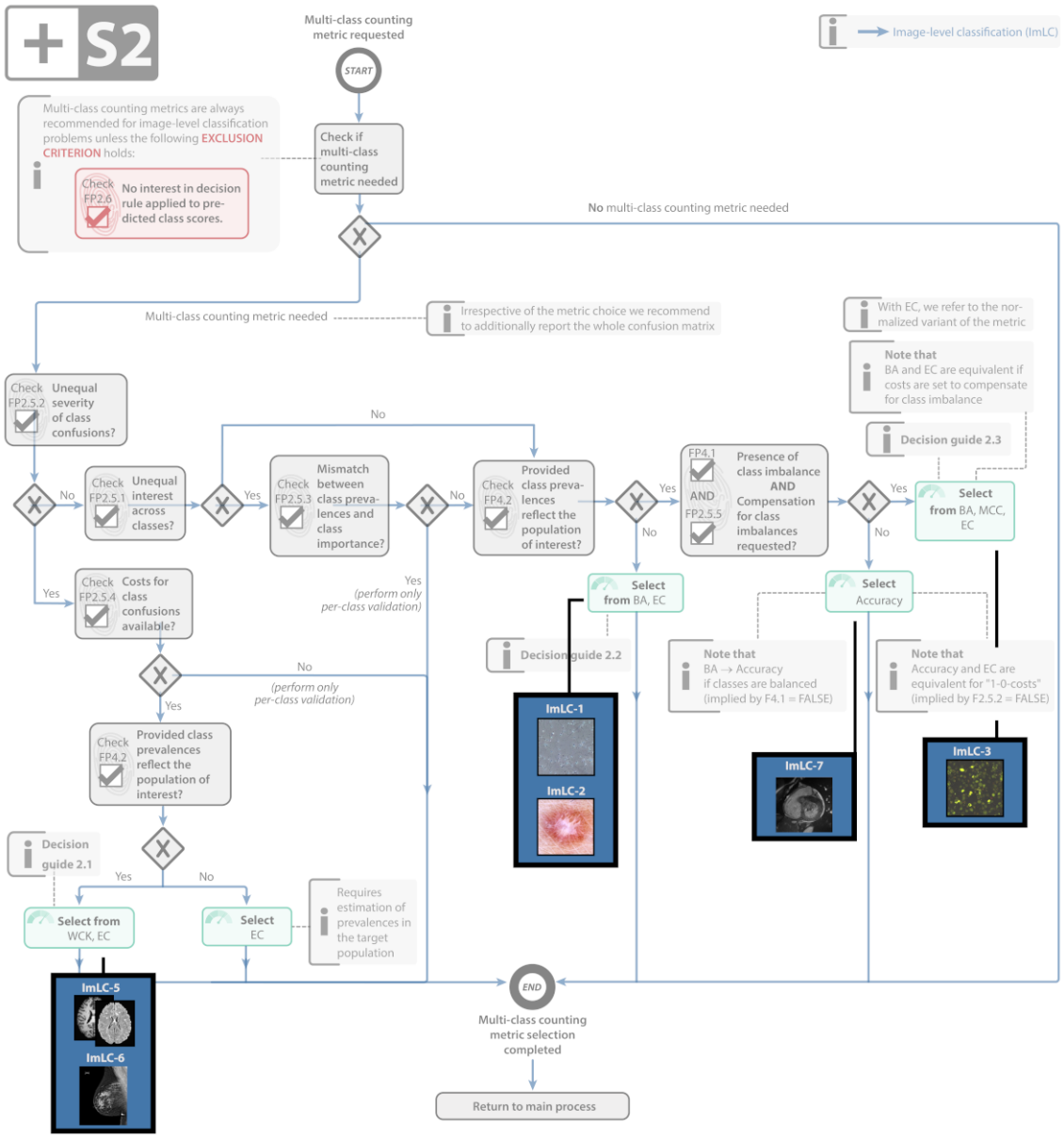


Figure 4.14: Traversal paths through of subprocess S2 to select multiclass counting metrics for seven concrete biomedical problems. The seven use cases **(ImLC-1)** - **(ImLC-7)** cover most branches of the recommendation process. Originally published in Maier-Hein et al. [238].

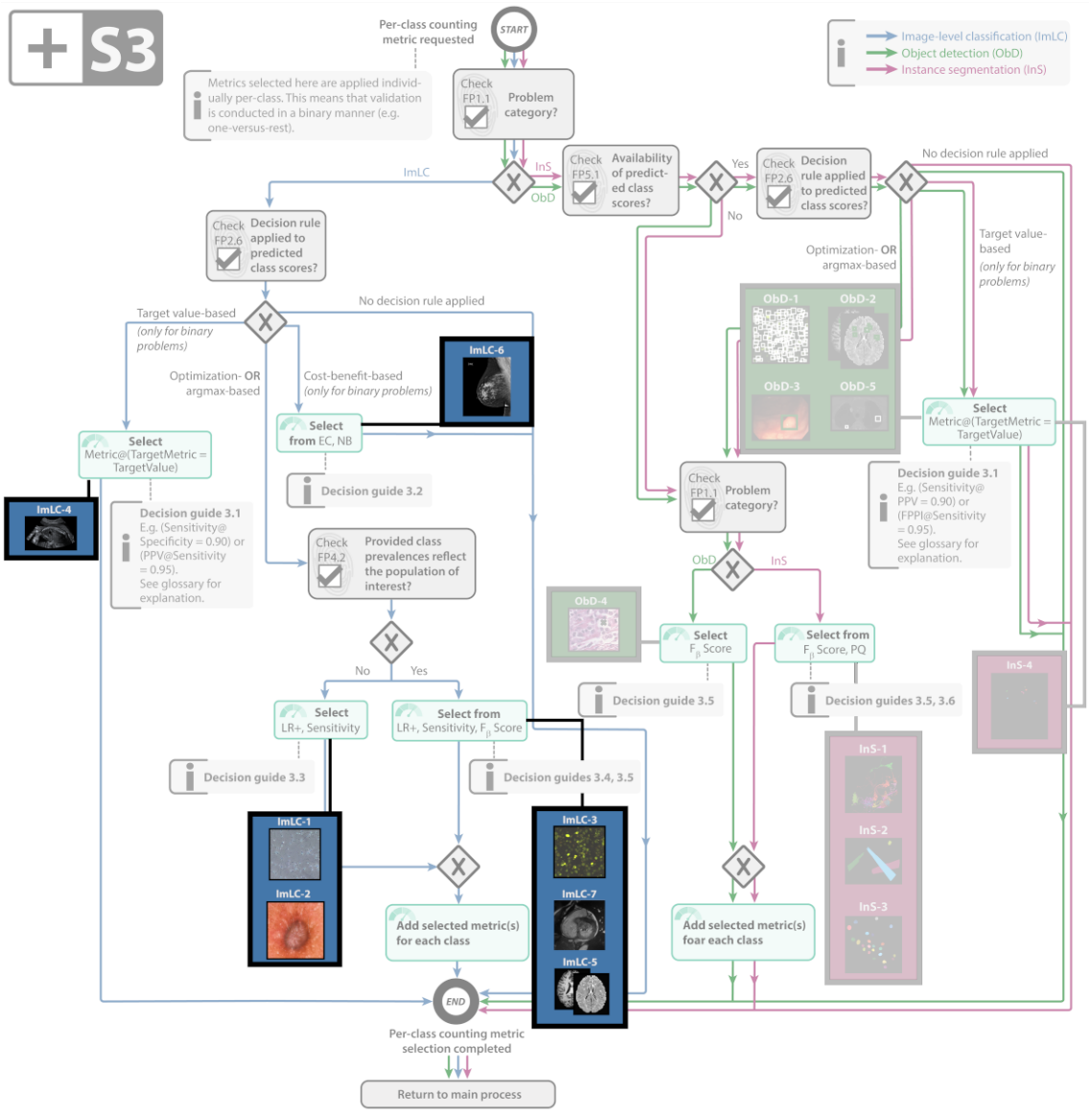


Figure 4.15: Traversal paths through of subprocess S2 to select per-class counting metrics for seven concrete biomedical problems. The seven use cases **(ImLC-1)** - **(ImLC-7)** cover all branches of the recommendation process. Irrelevant use cases for ObD and InS have been grayed out. Originally published in Maier-Hein et al. [238].

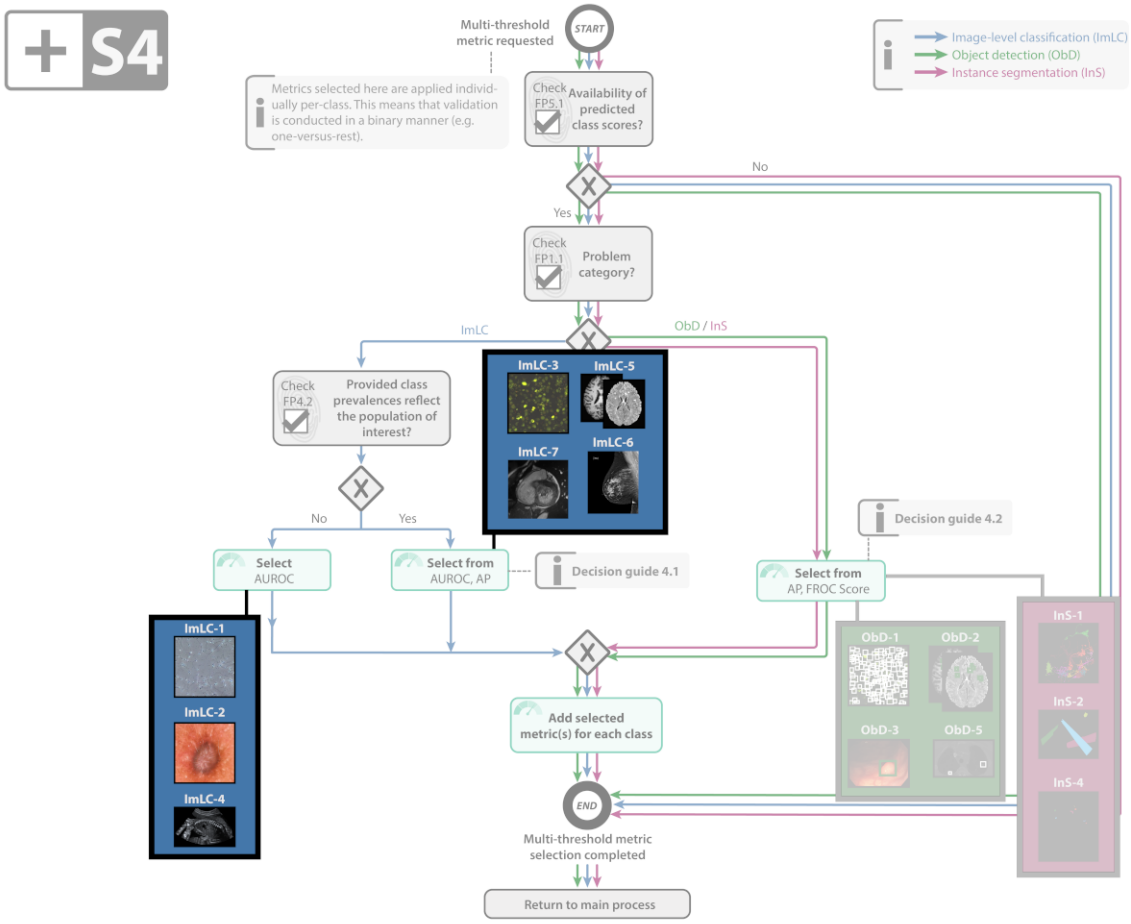


Figure 4.16: Instantiation of subprocess S4 for the selection of multi-threshold metrics with recommendations for concrete biomedical problems. The seven use cases **(ImLC-1)** - **(ImLC-7)** cover both branches of the recommendation process. Irrelevant use cases for ObD and InS have been grayed out. Originally published in Maier-Hein et al. [238].

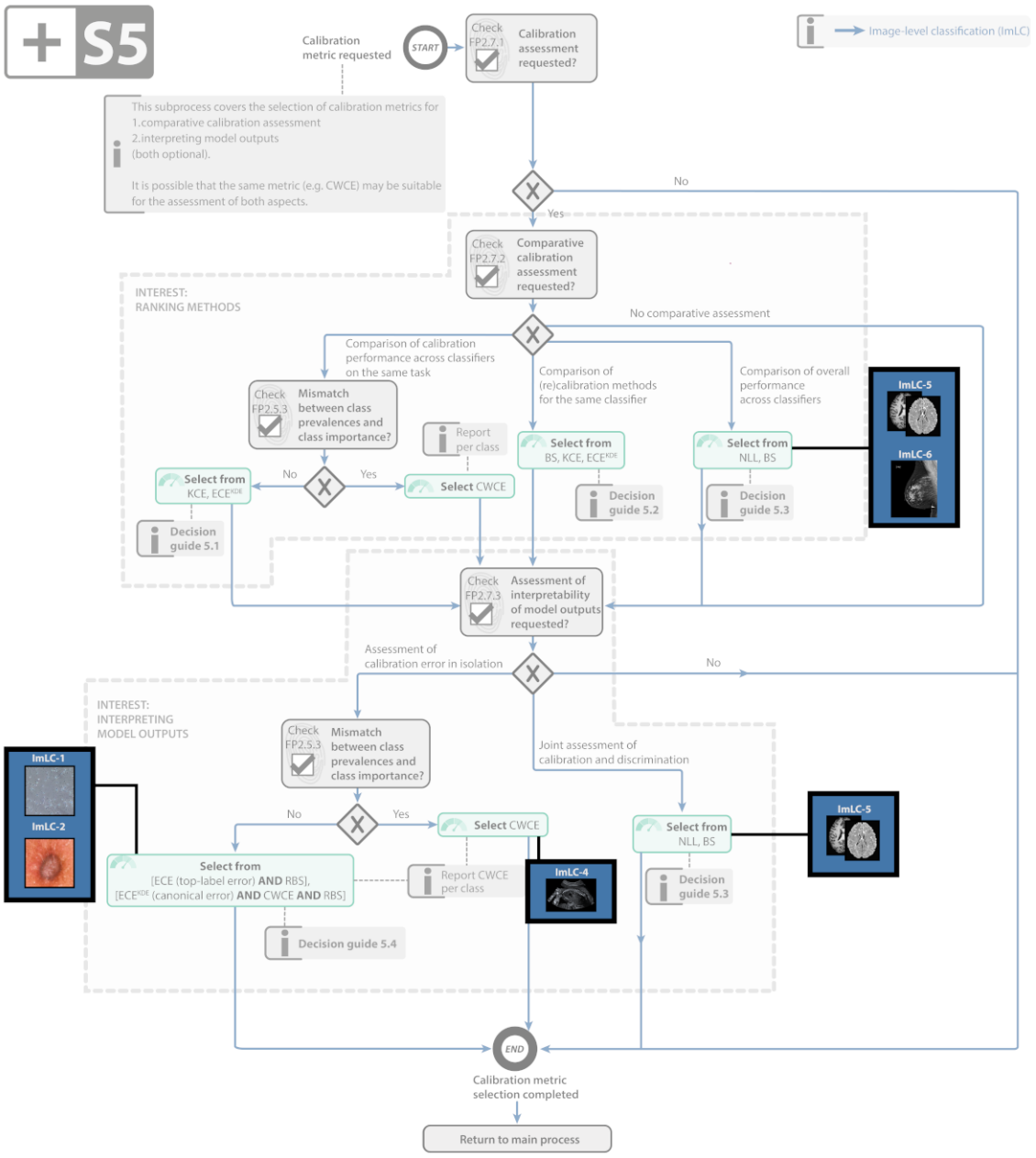


Figure 4.17: Instantiation of subprocess S5 for the selection of calibration metrics with recommendations for concrete biomedical problems. The seven use cases **(ImLC-1)** - **(ImLC-7)** cover multiple branches of the recommendation process. Originally published in Maier-Hein et al. [238].

4.3 Discussion

Our international Delphi process to answer **(RQ1)** has yielded several significant results, most notably the consortium agreed upon:

(i) A systematic workflow that interrogates a few fundamental properties of the use case (the *problem fingerprint*) to determine a set of appropriate performance measures (see Sec. 4.2.1),

(ii) decision guides to resolve some remaining ambiguous situations (see Sec. 4.2.2),

(iii) other general recommendations for the use of performance measures (see Tab. 4.3).

The broad applicability has been demonstrated with seven diverse tasks in Sec. 4.2.3. The entire interview process has been implemented as a web-based tool that guides the domain experts through the entire recommendation process. It restricts the interview questions to the relevant information required in each specific step and is freely available [1]. Reference implementations for all metrics in the metric pool (see Tab. 4.1) are available within the MONAI open source framework [373].

Interpretation

Metrics Reloaded emerged from a rigorous 2.5-year process involving five workshops, nine surveys, and numerous expert discussions, balancing established (potentially flawed) and new (not yet stress-tested) metrics. We addressed the fundamental challenge of validation – the absence of definitive 'best metrics' – through a three-part strategy: leveraging established consensus-building methods used in guidelines such as CONSORT [344], TRIPOD [260], and STARD [39, 40], gathering feedback through social media campaigns, and testing in diverse biomedical use cases (see Sec. 4.2.3). This approach achieved 93% median agreement across the subprocesses, with debates primarily centered on calibration metrics. For example, some members questioned the value of stand-alone calibration metrics altogether. The reason for this view is the critical misconception that the predicted class scores of a well-calibrated model express the true posterior probability of an input belonging to a particular class [295].

The significance of our framework extends well beyond academic exercise, forming a critical bridge between scientific innovation and patient benefit in biomedical imaging. Recent advances in technology readiness assessment and regulatory science emphasize the pivotal role of metrics that address end-user needs and real-world applicability [217, 223]. By streamlining metric selection, *Metrics Reloaded* enhances the quality of biomedical image analysis research while potentially accelerating AI translation into clinical practice. For Lifelong Learning systems operating in healthcare environments, this framework provides essential guidance on dynamically selecting appropriate evaluation metrics as tasks and contexts evolve. This capability is crucial when the system must autonomously

evaluate its own performance across changing clinical scenarios without constant human supervision.

Research context

Metrics Reloaded primarily provides guidance for selecting metrics that measure some notion of the 'correctness' of an algorithm's predictions on a set of test cases. However, holistic assessment of algorithm performance encompasses additional critical dimensions that are actively being researched in the biomedical imaging community and beyond.

Robustness represents a fundamental research direction that is particularly relevant in medical imaging, where continuous changes in the data distribution can be expected due to manifestation, acquisition, and prevalence shifts, all of which directly affect the characteristics of the imaging data (see Sec. 6.1). Current approaches to assessing robustness include stress testing, where performance is monitored under simulated but realistic perturbations of image characteristics [105]. Notably, recent work shows that the robustness challenge is prevalent on a large scale beyond medical imaging [85].

Reliability, defined as the ability of an algorithm to communicate its confidence and raise a flag when uncertainty is high, is another active area of research [343]. For calibrated models, this can be achieved via predicted class scores, although other methods based on dedicated model outputs trained to express confidence or density estimation techniques are also popular [125]. The research community is increasingly recognizing that algorithms with reliable uncertainty estimates or increased robustness to distributional shifts may not always have the best predictive performance [184], suggesting that safe deployment of classification systems requires careful balancing of robustness and reliability against accuracy.

Bias detection and mitigation constitute critical components of the research landscape. Learning-based algorithms rely on historical datasets for training, creating a risk that existing biases may be replicated or exacerbated – a phenomenon called 'Shortcut Learning' [126]. This concern is particularly acute in healthcare, given the documented scarcity of representative data from underserved populations and higher error rates in diagnostic labels for certain subgroups [6, 178, 274]. Current research emphasizes the need for relevant meta-information, such as patient demographics, to be accessible for test sets to detect potentially disparate performance across subgroups [245].

While *Metrics Reloaded* focuses on technical validation, the research community recognizes that clinical translation requires further validation steps that compare algorithm performance to conventional care according to patient-related outcome measures, such as 'overall survival' [285]. Moreover, there are broader efforts toward responsible research frameworks that encompass environmental, ethical, economic, social, and societal aspects of digital technologies [186, 259, 383].

Limitations

While we believe that our framework covers the vast majority of biomedical image analysis use cases, suggesting a comprehensive set of metrics for every possible biomedical problem may be beyond its scope. The focus of our framework is to correct poor practices related to the selection of common metrics by incorporating use case-specific knowledge into the decision process. However, for some use cases, our pool of common performance measures (see Tab. 4.1) may be inappropriate. In fact, in some cases *application-specific metrics* may be required. To make our framework applicable to such specific use cases, we have integrated the step of selecting application-specific metrics into the main workflow (see Fig. 4.7). Examples of such application-specific metrics can be found in related work [79, 107].

Note that while *Metrics Reloaded* focuses on the *selection* of metrics, proper *application* is also important. Detailed failure case analysis [327] and performance evaluation on relevant subgroups have been highlighted as critical components for better understanding when and where an algorithm may fail [57, 273].

The generation and handling of fuzzy reference data (e. g., from multiple observers) is a topic in its own right [377] and was deemed to be beyond the scope of this work. This limitation should be addressed in future work, especially as fuzzy references are common in clinical settings where inter-observer variability is significant.

Future work and broader impact

As is common in the development of scientific guidelines and recommendations, it will be necessary to regularly update our framework to reflect current developments in the field, such as the inclusion of new metrics or biomedical use cases. It is already planned to extend the scope of the framework to other problem categories, such as regression and reconstruction. In order to accommodate future developments in a fast and efficient manner, we envision our consortium building consensus through accelerated Delphi rounds organized by the *Metric Reloaded* core team. Once consensus is reached, changes will be implemented in both the framework and the online tool and will be highlighted so that users can easily identify changes from the previous version, ensuring full transparency and comparability of results. In this way, we envision the *Metrics Reloaded* framework and online tool as a dynamic resource that will reliably reflect the state of the art at any given point in the future, for years to come.

Of note, while the recommendations provided originate from the biomedical image analysis community, many aspects are generalizable to imaging research as a whole. In particular, the recommendations derived for individual fingerprints (e. g., implications of class imbalance) hold across domains, although it is possible that for different domains the existing fingerprints would need to be complemented by additional features that are not known to this community.

So far, *Metrics Reloaded* focuses on common *reference-based* methods that compare

model outputs to corresponding reference annotations. We made this design choice based on our hypothesis that reference-based metrics can be selected in a modality- and application-agnostic manner using the concept of problem fingerprinting. As indicated by the step of selecting potential *non-reference-based* metrics (see Fig. 4.7), validation and evaluation of algorithms should go far beyond purely technical performance. Recent efforts specifically devoted to estimating the energy consumption and greenhouse gas emissions of ML algorithms highlight one such dimension [288, 371]. While tracking this particular aspect is feasible from the perspective of a Lifelong Learning system [215], most of the remaining aspects of responsible research – ethical, economical, social, and societal implications – remain open problems.

For future iterations, incorporating these broader responsibility metrics will be essential to ensure that autonomous AI systems are not only optimized for clinical performance, but also consider their broader societal and environmental impacts over their operational lifetime.

Conclusion

This chapter demonstrates that a systematic approach to metric selection can significantly enhance the validation of biomedical image analysis algorithms and achieved high consensus within the international Delphi consortium. The *Metrics Reloaded* framework provides a structured methodology for selecting appropriate metrics based on problem fingerprints, addressing a critical gap in the validation process.

These findings contribute significantly to the first metacognitive loop of our Lifelong Learning system: the alignment of model validation during the **Design** phase. By autonomously interviewing domain experts, we enable AI systems to learn about the necessary healthcare contexts, aligning the task-specific goals of the Lifelong Learning system with guided minimal human interaction. With the capacity for self-directed metric selection, we lay the groundwork for AI systems that can continue to learn, adapt, and validate their own performance throughout their operational lifecycle in healthcare environments.

5 Knowledge Transfer for Training Image Classification Algorithms in Sparse Data Settings

Disclosure

Parts of the results of this chapter have been submitted for publication to *Nature Communications Medicine*. Preliminary results have been published at the *Medical Imaging Meets NeurIPS workshop* [342] and the *Medical Image Computing and Computer Assisted Interventions (MICCAI) conference* [136]. See App. A for full disclosure.

This chapter addresses the second research question of designing a **pipeline-learning loop** as part of the **Develop** phase in the AI lifecycle (see Fig. 5.1):

Research Question 2

How to enable effective knowledge transfer across biomedical image analysis tasks?

In Sec. 2.1 we elaborated on *data scarcity* as a leading roadblock for translational success of AI models in medical imaging. The essence of a Lifelong Learning system is to leverage *experience* from other, previously encountered tasks to improve on each individual one (see Sec. 2.8). In this chapter we want to present our approach for such *knowledge transfer* between tasks and how it could be instantiated in the challenging medical environment. Sec. 5.1 will formalize the problem more precisely and derive possible solutions. It will also describe multiple scenarios to leverage knowledge and how to properly evaluate the success of knowledge transfer. Afterwards, Sec. 5.2 will report on the experiments we conducted with respect to the described methodology. Finally, Sec. 5.3 closes the chapter with a discussion of our results.

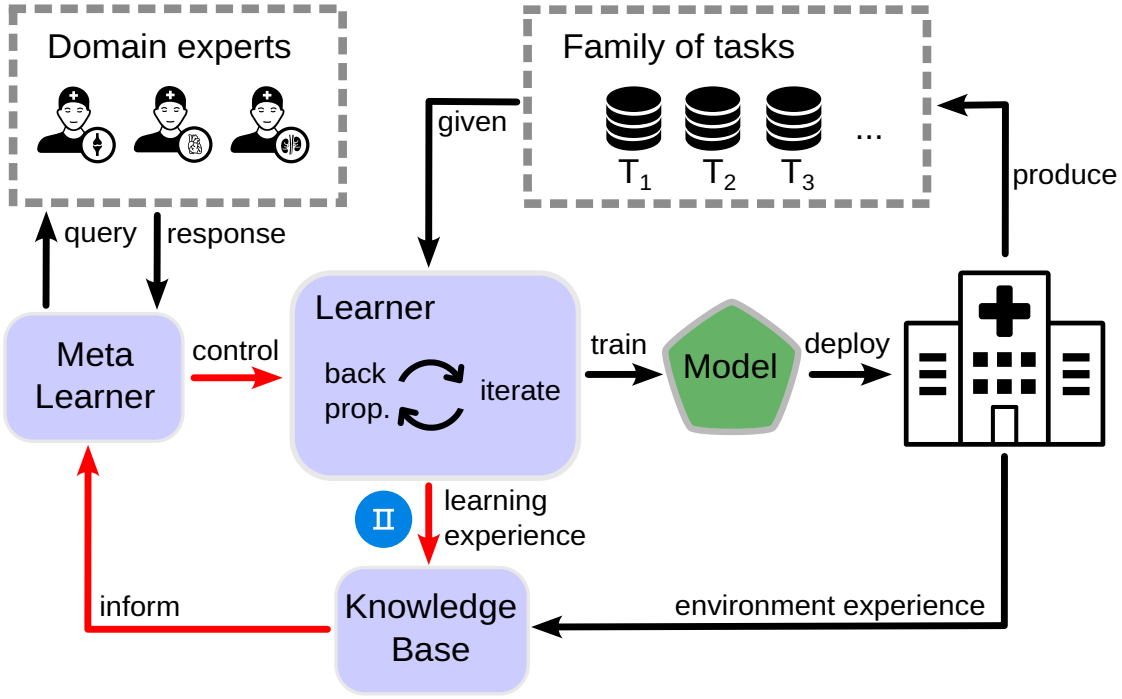


Figure 5.1: Pipeline-learning loop. Anchoring of this chapter in the overall Lifelong Learning system (see Fig. 1.1). Given a task, the *Meta Learner* leverages previous training experience from the *Knowledge Base* to determine the optimal training pipeline for the *Learner* to train a model. Loop is highlighted in red.

5.1 Methods

From Fig. 5.1 it is clear that we intend to store and update experience in the *Knowledge Base*. In this section we describe our main methodological contributions as well as the evaluation concept we have developed to assess the value of our approach. With respect to privacy concerns in the medical domain, we strive to find a solution that allows knowledge to be 'exchanged' across institutions for more efficient use of existing experience, better generalizability of solutions, and reduced overall resource consumption. The second crucial step is to identify important knowledge and derive a concrete instruction for the *Meta Learner* on how to influence the learning process of the model. Importantly, *hyperparameters* (see Def. 2.84) are decoupled from the training data, whose retransmission may be limited by regulatory requirements, and they also constitute concrete instructions for the *Meta Learner*. While such hyperparameters, e. g., on successful model trainings are thus easy to store, they are difficult to *match* to a new task without any information on the corresponding source task. For this, we need an associated 'identifier' that compresses – while preserving privacy – the task information corresponding to the hyperparameter. On the other hand, for some datasets that have been encountered before, the samples *are available* – for example, previous in-house tasks or publicly available

datasets from medical or non-medical computer vision research. But even for these tasks, any Transfer Learning approach (see Def. 2.86) requires task matching.

5.1.1 Task fingerprinting

Our proposed approach to fulfill these roles can be summarized by the following definition.

Definition 5.1. A **task fingerprinting method** is a tuple (f, d) comprising a **fingerprinting function** f that maps any task $\mathcal{T}$ onto a set of real valued vectors (called the **fingerprint** or **task embedding**) and a **distance measure** d (not necessarily symmetric or positive) that maps two fingerprints onto a real value. By convention, we will interpret such values as 'the lower, the higher the similarity'.

The challenge of a task fingerprinting method is to efficiently extract relevant task information into a fingerprint, while managing the trade-off between sufficient detail (for task matching) and removal of private information (for sharing). We chose the codomain *set of real-valued vectors* (an element of the powerset for some $\mathbb{R}^n$) as the most generic to capture a variety of existing and newly proposed methods. A diagram of the concept of task fingerprinting within a network of contributors to the shared *Knowledge Base* is given in Fig. 5.2. It is worthwhile to explicitly state the desirable properties of a 'good' task fingerprinting method.

(i) **encryption**: a fingerprint should not reveal any information about a particular sample to ensure that no sensitive patient data is shared.

(ii) **efficiency**: a complex computation of a fingerprint may be acceptable since it is likely to be a one-time event; however, the computation of pair-wise distances must be highly efficient since as the number of tasks in the *Knowledge Base* increases, the number of pairs grows quadratically and is a repetitive operation.

(iii) **quality**: task matching based on the distance measure between task fingerprints must result in a beneficial knowledge transfer.

An immediate simple idea for fingerprinting is to use a generically pretrained DNN as a *feature extractor*. Typically, the classification head of an ImageNet [93] pretrained DNN is 'removed' and the features from the penultimate layer are used [444].

Example 5.2. Let $\varphi : \mathcal{X} \to \mathbb{R}^m$ be a feature extractor, $p \in \mathbb{N}$ and $\|\cdot\|_p$ the usual L_p norm of vectors. For some integer n let s be a **sampling function** (see Def. 2.77) that chooses n samples from a task $\mathcal{T}$. The **mean pairwise distance** d_p between two sets of vectors is given by computing L_p for any difference of combinations of

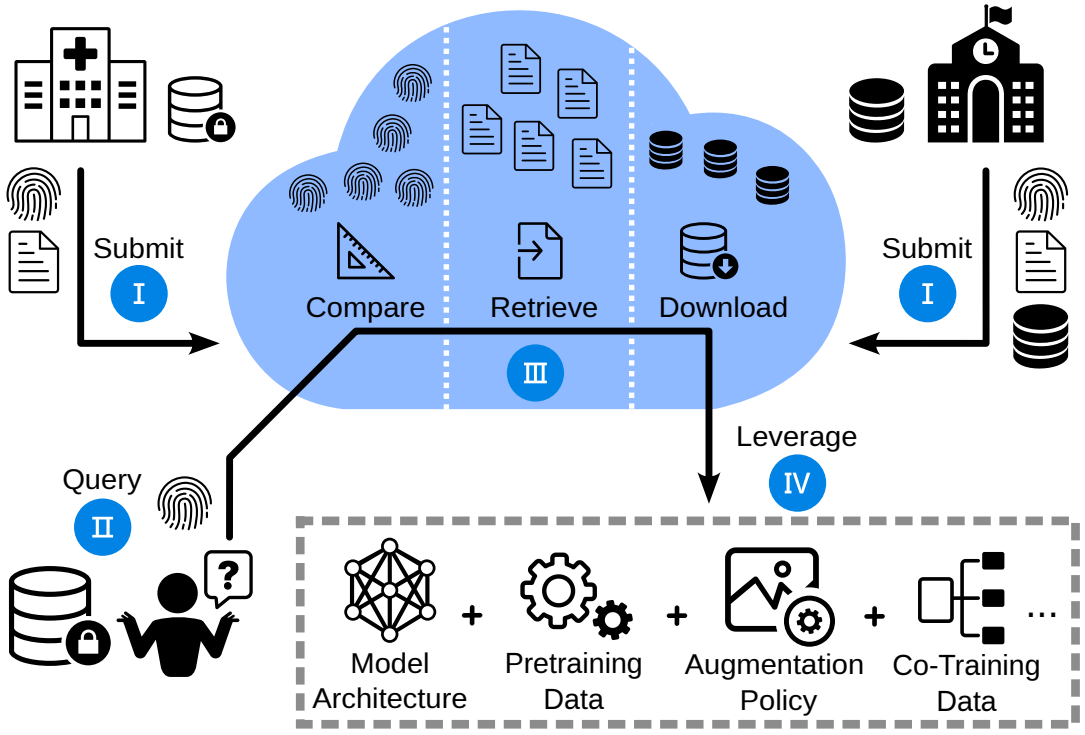


Figure 5.2: Collective knowledge acquisition through task fingerprinting. (I) A network of contributors may transmit experience on model training to the joint *Knowledge Base* by submitting a shareable task representation ('fingerprint'), meta information about their training strategies, and, optionally, their data. (II) To query the *Knowledge Base* for existing experience, the fingerprint for the current task is generated. (III) Based on the most relevant tasks in the pool according to fingerprint matching, relevant training strategies and data can be retrieved. (IV) The retrieved meta information and data are used to compile a training pipeline with different components of transferred knowledge. In this study, we investigate four scenarios of knowledge transfer, namely (a) model architecture, (b) pretraining data, (c) augmentation policy, and (d) co-training data. Published as a preprint in Godau et al. [137].

vectors from the set and averaging the results:

$$d_p(\{u_i\}_{i\leq n}, \{v_i\}_{i\leq n}) := n^{-2} \sum_{i,j\leq n} \|u_i - v_j\|_p.$$

Then $(\varphi \circ s, d_p)$ is a task fingerprinting method, called **naive fingerprinting**.

There are several problems with naive fingerprinting. First, the *encryption* is inadequate because deep features of individual samples can be identified from the fingerprint, which has been shown to potentially reveal sensitive information [153]. Second, the pairwise computation of vector norms scales quadratically with the number of extracted samples n and can lead to *inefficient* fingerprint comparison. Finally, the Euclidean (or any other L_p norm based) distance in feature space may not be a strong indicator of conceptual similarity due to the high dimensionality [7]. Thus, it may not be surprising that the naive fingerprinting method does not meet our requirements. But some of its core ideas have made it into a class of fingerprinting methods in the literature, which we want to collect and discuss. Since most of them are at least partially based on a 'sampling plus feature extraction' technique, we will fix φ, $n \in \mathbb{N}$ and s from ex. 5.2 throughout this section.

Definition 5.3. The **Maximum Mean Discrepancy (MMD)** [148] is the largest difference in expectations over functions in the unit ball of a RKHS (see Def. 2.69). It is given by

$$d_{\text{MMD}}(\{u_i\}_{i\leq n}, \{v_i\}_{i\leq n}) = n^{-1} \cdot (n-1)^{-1} \sum_{i,j\leq n, i\neq j} k(u_i, u_j) + k(v_i, v_j) - 2n^{-2} \sum_{i,j\leq n} k(u_i, v_j),$$

for some **kernel** k, e. g., the **Cauchy kernel** $k(u, v) = (1 + \|u - v\|^2 \nu^{-2})^{-1}$ with hyperparameter ν (called **bandwidth**)[a]. Then $(\varphi \circ s, d_{\text{MMD}})$ is a task fingerprinting method, called **MMD fingerprinting**.

[a] Often σ is used in literature for the bandwidth (e. g., in Song et al. [365]), but we wanted to avoid confusion with the softmax function (see Def. 2.8).

MMD fingerprinting only partially solves the problems of naive fingerprinting, since it relies on the same *encryption*, and MMD computation also scales poorly with the number of extracted samples n, i. e., it does not qualify our *efficiency* requirement. Nevertheless, MMD recently showed some success for distribution comparison in the training of GANs [144] and Invertible Neural Networks (INNs) [16, 104].

Another common distance function of feature distributions in ML is named after Leonid

Vaserstein.

Definition 5.4. Let $m \in \mathbb{N}$, $p \in [1, +\infty)$, d be a metric (in the mathematical sense) on $\mathbb{R}^m$, and q, q' be two probability measures on $\mathbb{R}^m$, the **p-Wasserstein distance** [401] is defined as:

$$W_p(q, q') := \inf_{\gamma \in \Gamma(q,q')} (\mathbb{E}_{(x,y)\sim\gamma} d(x, y)^p)^{1/p}, \tag{5.1}$$

where $\Gamma(q, q')$ is the set of all **couplings** between q and q', i. e., the set of probability measures on $\mathbb{R}^m \times \mathbb{R}^m$ whose marginals are q and q' on the first and second factors respectively. If $p = 1$ we call W_1 the **Earth Mover's distance** [35]. In case $p = 2$ we call W_2 the **Fréchet distance** [119].

The Wasserstein distance quantifies the minimum effort required to transform one probability distribution into another. It can be conceptualized through the classical mass transport problem: if we consider each distribution as a mass of unit weight distributed over a metric space, the Wasserstein distance represents the minimum work required for this transformation, calculated as the product of the mass transported and the distance traveled. The simplest example to demonstrate this is to compare two point masses in $\mathbb{R}^n$.

Example 5.5. Let $m \in \mathbb{N}$, $p \in [1, +\infty]$, d be a metric on $\mathbb{R}^m$, $\hat{x}, \hat{y} \in \mathbb{R}^m$ and χ_x be the **characteristic function** for $x \in R^m$, i. e., for any $y \in R^m$

$$\chi_x(y) := \begin{cases} 1 & \text{, if } x = y \\ 0 & \text{, else.} \end{cases}$$

Then the point-probability measure $\chi_{\hat{x}}$ and $\chi_{\hat{y}}$ have p-Wasserstein distance

$$W_p(\chi_{\hat{x}}, \chi_{\hat{y}}) = d(x, y), \tag{5.2}$$

as the only element in $\Gamma(\chi_{\hat{x}}, \chi_{\hat{y}})$ is $\chi_{(\hat{x},\hat{y})}$.

Ex. 5.5 shows an interesting difference between W_p and the previously introduced Kullback-Leibler Divergence (KLD) (see Def. 2.73) as an alternative dissimilarity measure for probability distributions: It compares the distances in the *domain* of the probability measures, as opposed to caring only about the probability differences at each point of the domain (see Fig. 5.3). This emphasizes the aforementioned perspective of W_p to quantify the minimal energy required to move the mass of one distribution to form another. This transport based interpretation, originally formulated by Monge in 1781 [258], has led to its alternative designation in the computer science literature as the *earth mover's*

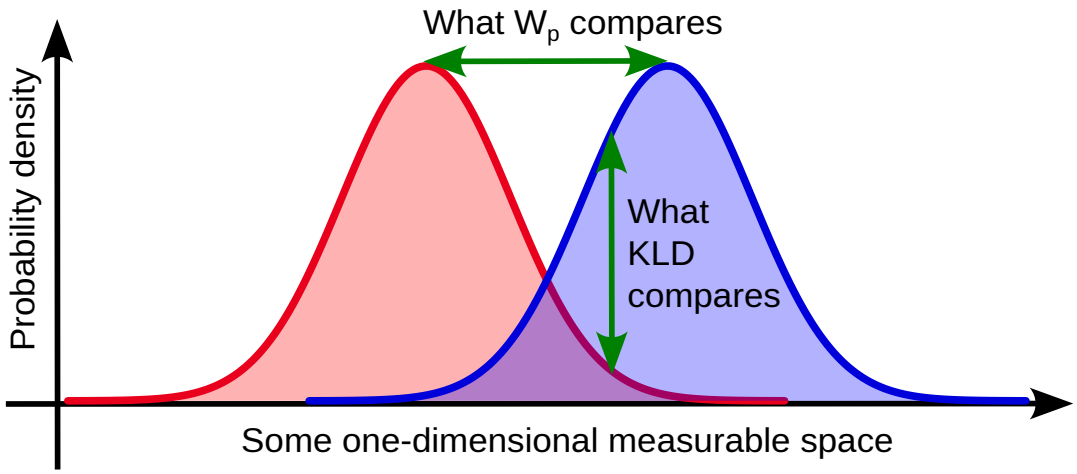


Figure 5.3: Fundamental difference between p-Wasserstein distance (W_p) and Kullback-Leibler Divergence (KLD) to compare probability distributions. While the Wasserstein distance compares distances along the probability domain direction, KLD only conducts pointwise comparison of probability density differences.

distance. The metric has found widespread applications in machine learning, particularly in generative models [17] and optimal transport theory [405], due to its ability to capture geometric properties of the underlying space.

For our purposes – comparing distributions of extracted features – the probability measures q and q' will be 'empirical distributions' (assigning equal weight to each sample), and computing W_p is equivalent to solving a linear program [279]. For this general case, the best algorithms scale in worst-case complexity as $O(n^3 \log n)$ given n samples each for q and q' [279]. This unfortunate scaling hurts our desired property of *efficiency* for comparing task distances. Fortunately, there are two special cases where the computation is drastically simplified. The first is the one-dimensional case, i. e., $m = 1$, where W_p can be computed in one pass using the quantiles of the distributions [312]. To do this, we need to examine each feature dimension by itself when comparing empirical distributions.

Definition 5.6. We define

$$d_{\mathrm{EMD}}(\{u_i\}_{i\leq n}, \{v_i\}_{i\leq n}) := m^{-1} \sum_{j\leq m} W_1(\sum_{i\leq n} \chi_{u_i(j)}, \sum_{i\leq n} \chi_{v_i(j)}),$$

then $(\varphi \circ s, d_{\mathrm{EMD}})$ is a task fingerprinting method, called **Earth Mover's Distance (EMD) fingerprinting** [136].

While the EMD fingerprinting improves the *efficiency* problem compared to the multivariate W_p, the *encryption* in the above version does not differ from the naive fingerprint-

ing. However, since the empirical distributions of each feature dimension are examined separately, it is possible to permute entries of each feature dimension individually. Thus, for large n and m, we can decompose features sufficiently to ensure *encryption*. Both *efficiency* and *encryption* can be further improved by a 'binning' strategy.

Definition 5.7. Let $B \in N$ and $b_0 < b_1 ... < b_B \in R$, called the edges of B bins. For a set of $n \in \mathbb{N}$ scalars $\{x_i\}_{i \leq n}$, and $j \leq B$ we define the j-th **binning** function

$$\text{bin}_j(\{x_i\}_{i \leq n}) := \sum_{i \leq n} \begin{cases} 1 & \text{, if } x_i(j) \in [b_{j-1}, b_j) \\ 0 & \text{, else.} \end{cases}$$

Now let $b_0 = -\infty$ and $b_B = +\infty$, then the **normalized histogram** function hist maps the set of n scalars $\{x_i\}_{i \leq n}$ onto $p \in \Delta_{B-1}$, where $p_j := n^{-1}\text{bin}_j(\{x_i\}_{i \leq n})$ is the relative prevalence of elements in the set falling into the j-th bin. Applied to a set of feature vectors $U = \{u_i\}_{i \leq n}$ with $u_i \in \mathbb{R}^m$, we let further $\text{hist}(U) := \{\text{hist}(\{u_i(j)\}_{i \leq n})\}_{j \leq m}$ be the set of normalized feature histograms.

Binning trades the 'resolution' of feature values for a simpler data structure with only B quantiles to compute for EMD. For our later experiments, we will fix b_1 and b_{B-1} per feature dimension and optimize the resolution hyperparameter B. We will continue to refer to the combination of normalized histograms and EMD fingerprinting as **VDNA** because it was first introduced by Ramtoula et al. [313].

We now want to return to the second case of simplified W_p computations, which is independent of the dimension m, but fixes $p = 2$ and makes the strong assumption that q and q' are both Gaussian distributions.

Definition 5.8. For a set of $n \in \mathbb{N}$ feature vectors $U = \{u_i\}_{i \leq n}$, with $u_i \in \mathbb{R}^m$ for some $m \in \mathbb{N}$, we define the **empirical Gaussian measure** [279] $\hat{\mu}_U = \mathcal{N}(\hat{m}, \hat{\Sigma})$, with mean $\hat{m}$ and covariance matrix $\hat{\Sigma}$ given by

$$\hat{m} := n^{-1} \sum_{i \leq n} u_i$$
$$\hat{\Sigma} := (n-1)^{-1} \sum_{i \leq n} (u_i - \hat{m})(u_i - \hat{m})^T.$$

The optimization of W_2 for two Gaussian measures has a closed-form solution, allowing efficient computation [102] and motivating our next fingerprinting method.

Definition 5.9. For formal precision we let $\hat{\mu}$ be a mapping from extracted features $U = \{u_i\}_{i \leq n}$ to the corresponding tuple $(\hat{m}, \hat{\Sigma})$ of empirical mean and covariance[a]. We define

$$d_{\text{FID}}((\hat{m}_1, \hat{\Sigma}_1), (\hat{m}_2, \hat{\Sigma}_2)) := W_2(\mathcal{N}(\hat{m}_1, \hat{\Sigma}_1), \mathcal{N}(\hat{m}_2, \hat{\Sigma}_2)),$$

then $(\hat{\mu} \circ \varphi \circ s, d_{\text{FID}})$ is a task fingerprinting method, called **Fréchet Inception Distance (FID) fingerprinting** [97, 166].

[a] These still can be represented as an ordered set of $m + 1$ vectors from $\mathbb{R}^m$.

FID elegantly solves both the *encryption* and the *efficiency* problem: Instead of image-level features, only the means and covariance matrices of the features need to be revealed. Also, the closed-form solution speeds up the computation[1].

Before we dive into our own proposed fingerprinting method, we want to quickly note that task embeddings do not necessarily have to rely on extracted features of task images.

Definition 5.10. Let $\mathcal{T}$ be a task and φ_w a probabilistic model for $\mathcal{T}$ with some weights $w \in \mathbb{R}^k$ (see Def. 2.78). The **Fisher Information Matrix** [337] $F \in \mathbb{R}^k \times \mathbb{R}^k$ is defined as

$$\boldsymbol{F} := \mathbb{E}_{x,y \sim p_{\mathcal{T}}(X), \varphi_w(x)}[\nabla_w \log \varphi_w(x)_y (\nabla_w \varphi_w(x)_y)^T].$$

For a fixed backbone model φ_w we define the mapping $\mathbf{e}$ from the class of all tasks onto $\mathbb{R}^k$, with $\mathbf{e}(\mathcal{T})_i := \boldsymbol{F}_{i,i}$ called the **Fisher embedding** of $\mathfrak{T}$. Here the *Fisher Information Matrix* $\boldsymbol{F}$ corresponds to a fine-tuned model $\hat{\varphi}$ to task $\mathfrak{T}$ (see Def. 2.87), where we restrict the weights of the SGD to a single attached linear layer (see Def. 2.79). On the contrary we restrict the weights w in the computation of $\boldsymbol{F}$ to those of the backbone φ_w.

The i-th entry in the Fisher embedding of a task indicates the 'importance' of the backbone model parameter w_i for predictions of φ on $\mathcal{T}$ [4]. Intuitively, a convolutional kernel (see Def. 2.80) that is relevant to the task, e. g., one that 'detects' certain shapes in an image [310] and thus receives a high value in the Fisher embedding, is highly indicative of the 'nature' of the task. Unlike the previous feature extraction approach, Fisher embeddings also contain information about the *labels* of $\mathcal{T}$. This property is particularly useful for distinguishing between tasks that share the same images (e. g., the Cholec80 tasks and some CheXpert tasks from Sec. 2.2)[2]. Note that although the Fisher

[1] Of all actually computed task fingerprinting methods along the experiments for this chapter, FID was, next to MMD, still among the slowest.

[2] We would like to briefly point out that there are also a number of approaches that combine feature extraction with label information [12, 84].

embedding consists of a single vector, its dimension k is much larger than the dimension m of a feature vector (assuming the same backbone model). While m refers to the neuron activations in the last layer of φ, k corresponds to the Fisher information for all neurons of φ[3]. In the setting of our experiments (see Sec. 5.1.3), k and $m \cdot n$ are in the same order of magnitude. To use the Fisher embedding for task fingerprinting, we need to choose a distance method.

Definition 5.11. The **cosine similarity** [357] for vectors $u, v \in \mathbb{R}^k$ is defined as

$$\operatorname{sim}(u, v) := \frac{\sum_{i \leq k} u_i v_i}{\sqrt{\sum_{i \leq k} u_i^2} \cdot \sqrt{\sum_{i \leq k} v_i^2}}.$$

As the cosine similarity is positively oriented we further introduce $d_{\text{FED}}(u, v) := 1 - \operatorname{sim}(u, v)$ and may call the task fingerprinting method $(\mathbf{e}, d_{\text{FED}})$ the **Fisher Embedding Distance (FED)** [4, 136].

Since cosine similarity is easy to compute, FED satisfies our *efficiency* criterion, although the (one-time) cost of generating the fingerprints is higher than for feature extraction approaches. With respect to *encryption* FED obfuscates all samples from the computation of the Fisher embedding jointly per weight. For a sufficiently large number of samples used in the computation of the *Fisher Information Matrix*, this can be expected to preserve individual patient privacy. Note, however, that FED also needs samples from $\mathcal{T}$ for the fine-tuning step of the backbone model φ_w. We therefore postulate – without rigorous proof – that for a fixed total number of samples used in the embedding process, the *encryption* quality can be considered worse than for EMD fingerprinting (with the mentioned random permutations).

Turning our attention back to feature extraction-based approaches, we present a precursor to our soon-to-be-proposed task fingerprinting method. As shown in Fig. 5.3 one can also use the KLD to compare feature distributions[4]. Moreover, our definition of KLD (see Def. 2.73) extends beyond probability measures on the particular label space to any finite measurable space – and can even be extended to continuous probability distributions [33]. A simple way to use KLD is on a highly aggregated feature fingerprint.

Definition 5.12. We define an **aggregation function** a on a set of feature vectors

[3] In our implementation we made some tweaks to the weights used for $\boldsymbol{F}$. More specifically, we partially summarize the parameters of a convolutional kernel by taking the mean and ignore the bias parameters as well as some early network layers.

[4] Noteworthy Tan et al. [381] propose a fingerprinting method that combines p-Wasserstein and cross entropy (i. e., KLD based).

$U = \{u_i\}_{i \leq n}$, with $u_i \in \mathbb{R}^m$ onto $\mathbb{R}^m$ for $j \leq m$ as follows:

$$a(U)_j = \sum_{i \leq n} \frac{u_i(j)}{\|u_i\|_1}$$

Note that $\|a(U)\|_1 = 1$, i. e., $a(U) \in \Delta_{m-1}$. We call task fingerprinting method $(a \circ \varphi \circ s, \mathbf{KLD})$ **P2L fingerprinting** [32] (also **KLD fingerprinting** [136]).

P2L fingerprints are by far the most *encrypted* fingerprints we have presented so far. The m dimensional[5] fingerprints encapsulate all n samples used for feature extraction, where usually $m < n$. The KLD computations are also quite *efficient*, making P2L a top fingerprinting method from these perspectives. On the other hand, our preliminary experiments showed that P2L performs *qualitatively* worse than FED or EMD fingerprinting [136].

We therefore propose the following new fingerprinting method, which combines the binning idea from VDNA fingerprinting (see Def. 5.7) with the distance measure concept from KLD fingerprinting (see Def. 5.12). Note, however, that if there is an empty bin in the construction of the normalized histogram p, there may be some $i \leq B$ such that $p(i) = 0$. If the comparison is with a task whose corresponding bin is not empty, i. e., $q(i) > 0$, then $\mathbf{KLD}(p, q)$ is not defined (or set by convention to $+\infty$ [82]). We therefore add the softmax σ to avoid such scenarios. We also add a weighting scheme over the feature dimension.

Definition 5.13. Let $\{p_j\}_{j \leq m}$ and $\{q_j\}_{j \leq m}$ two sets of normalized feature histograms (see Def. 5.7) and $w \in \Delta_{m-1}$ be a **weighting scheme**. Now let

$$d_{\text{bKLD}}(\{p_j\}_{j \leq m}, \{q_j\}_{j \leq m}) = \sum_{j \leq m} w_j \mathbf{KLD}(\sigma(p_j), q_j)$$

Then $(\text{hist} \circ \varphi \circ s, d_{\text{bKLD}})$ is task fingerprinting method, called **binned Kullback-Leibler Divergence (bKLD) fingerprinting**.

The weighting scheme in bKLD is – besides the number of bins B – a hyperparameter that allows to shift the focus to some feature dimensions that are particularly relevant. Based on our exploratory experiments on the development tasks (see Sec. 5.1.3), we identified three favorable settings of hyperparameters: **bKLD(small,target)** a 'small' fingerprint with $B = 100$ bins and weighting by the softened feature average of the target task $w = \sigma(\sum_{i \leq n} v_i / n)$, **bKLD(large,source)** with $B = 1000$ bins and weighting by the normalized feature average of the source task $w = a(\{u_i\}_{i \leq n})$ as well as **bKLD(large,un-**

[5] From an information content perspective, the fingerprints are only $m - 1$ dimensional, as we know they live on the simplex Δ_{m-1}.

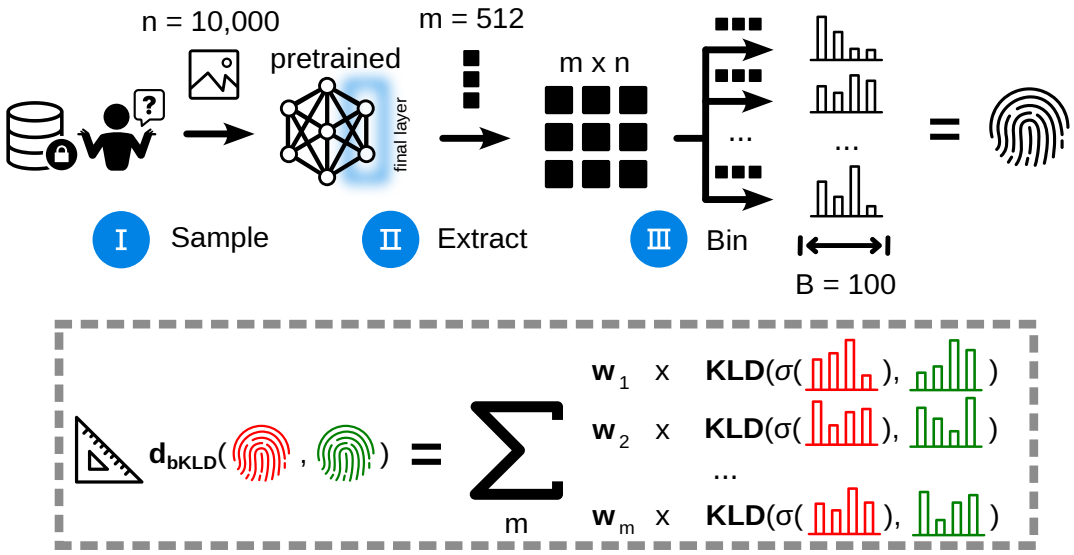


Figure 5.4: Concrete processing pipeline for our proposed binned Kullback-Leibler Divergence (bKLD) fingerprinting method. To compute a bKLD fingerprint for some task $\mathcal{T}$, we first sample $n = 10000$ images from $\mathcal{T}$ (I) and extract deep features through an ImageNet [93] pretrained ResNet34 [161] backbone, generating $m = 512$ features per image (II). We compute normalized histograms with $b = 100$ bins for each of the m features over the n samples (III), which comprise the fingerprint. To compare two fingerprints, a weighted sum of the KLD across all histograms is computed. The softmax operator σ is applied to source task histograms to avoid empty bins. Adapted from Godau et al. [137].

weighted) with $B = 1000$ and uniform weighting, i. e., $w_j = m^{-1}$ for all $j \leq m$. An overview of all computational steps for bKLD(small,target) is given in Fig. 5.4.

With these choices of hyperparameters B, the resulting fingerprints, i. e., normalized feature histograms $\{p_j\}_{j \leq m}$ with $p_j \in \Delta_{B-1}$ contain $B \cdot m$ scalar values and thus about one to two orders of magnitude fewer entries than fingerprinting methods that rely on the full $n \cdot m$ features, such as naive fingerprinting (see Ex. 5.2), MMD fingerprinting (see Def. 5.3), or EMD fingerprinting (see Def. 5.6). Binning automatically decouples the features per sample (which could be achieved at least for EMD fingerprinting by the proposed random permutation strategy) and further complicates any reverse reconstruction by the information loss due to bin-width resolution. From these considerations, and with our choice of hyperparameters $n = 10000$, $m = 512$, the *encryption* capabilities of the presented fingerprinting methods could thus be roughly ordered (from least favorable to most favorable) as follows

$$\textbf{naive} = \textbf{MMD} < \textbf{EMD} < \textbf{FED} < \textbf{VDNA} = \textbf{bKLD} < \textbf{FID} < \textbf{KLD}$$

Of course, a much more comprehensive information-theoretic analysis would be required to validate sufficient privacy, which is beyond the scope of this work. As a final

remark on this topic, we would like to point out that for a very small number of samples, obfuscation via bin resolution becomes more relevant than obfuscation via decoupling (or aggregation), as the extreme case of $n = 1$ shows.

Regarding *efficiency*, we already mentioned the inferiority of FID and MMD in our computations. We also noted that binning speeds up the computation of the 1-Wasserstein distance, i. e., VDNA is faster than EMD fingerprinting. The computations of the remaining distances are mostly linear in the size of the fingerprints.

Finally, we consider the *quality* of the fingerprinting methods, which is the primary focus of this work.

5.1.2 Validation of task fingerprinting

Recall from Def. 2.84 that a trainer $\mathfrak{T}$ takes a task $\mathcal{T}$ and some hyperparameters ω to compute a model $\varphi = \mathfrak{T}(\mathcal{T}, \omega)$. In our Lifelong Learning system the *Meta Learner* should provide the *Learner*, (i. e., the trainer), with favorable hyperparameters ω *particularly* for a given target task $\mathcal{T}$. Our proposed strategy is to query the *Knowledge Base* with a fingerprinting method (f, d), by computing $f(\mathcal{T})$ and comparing this fingerprint with all potential source tasks $\mathcal{S}$ in the *Knowledge Base* via $d(f(\mathcal{S}), f(\mathcal{T}))$ – similar to a retrieval problem. Suppose the *Knowledge Base* contains experience from a pool of n source tasks $\{\mathcal{S}_i\}_{i\leq n}$. We will denote $\hat{\mathcal{S}} = \operatorname{argmin}_{\{\mathcal{S}_i\}} d(f(\mathcal{S}_i), f(\mathcal{T}))$ for the best matching source task for (f, d). For each potential source task $\mathcal{S}$ we distinguish two cases:

The data of $\mathcal{S}$ is not accessible This allows only to use hyperparameters ω for $\mathcal{S}$ that are stored in the *Knowledge Base*.

The data of $\mathcal{S}$ is accessible This also allows to integrate $\mathcal{S}$ directly into the training process. Examples are sequential training, i. e., fine-tuning (see Def. 2.87) or parallel training, i. e., Multitask Learning (see Def. 2.85).

For simplicity, we will only consider a single 'best' existing combination of hyperparameters ω for each source task $\mathcal{S}$ in the *Knowledge Base*[6] The *Meta Learner* can set up the training pipeline according to the availability of data from any source task $\mathcal{S}$ by adopting some hyperparameters for model training. We will skip the exact steps of this setup for now, but in general we will refer to such a $\mathcal{S}$-informed trainer (with potential access to $\mathcal{S}$ data) as $\mathfrak{T}_{\mathcal{S}}$ and the updated hyperparameters as $\omega_{\mathcal{S}}$. To evaluate this process, we will consider a variety of **meta metrics**.

[6]Technically, each submitted hyperparameter ω on a source task $\mathcal{S}$ must be associated with the value of some performance measure, for example, some validation loss. This would allow a kind of 'distribution of good hyperparameters' to be derived, and subsequently the *Meta Learner* could propose a family of hyperparameters drawn from this distribution. However, due to computational constraints, we stick to the simple procedure of choosing only the best hyperparameters for $\mathcal{S}$.

Improvement

The first metric tries to quantify the benefit of using an informed trainer compared to an uninformed one. For this, we assume that the *Meta Learner* chooses rather well generalizable hyperparameters ω and possibly a small hyperparameter search strategy on the target task $\mathcal{T}$. Zamir et al. [438] used a meta metric called **gain**, which takes into account (the binary information) whether the informed trainer performs better than the uninformed one. We extend this metric to be more granular about the performance difference.

Definition 5.14. Let $\hat{\mathcal{S}}, \mathcal{T}$ be tasks and μ a performance measure for $\mathcal{T}$. Given some strategies from the *Meta Learner* to determine informed $\omega_{\hat{\mathcal{S}}}$ and uninformed ω hyperparameters, with corresponding trainers $\mathfrak{T}_{\hat{\mathcal{S}}}$ and $\mathfrak{T}$, we define the **improvement** as

$$\mu(\mathfrak{T}_{\hat{\mathcal{S}}}(\mathcal{T}, \omega_{\hat{\mathcal{S}}})) - \mu(\mathfrak{T}(\mathcal{T}, \omega)).$$

Improvement inherits the orientation of the underlying performance measure μ. The value range also shifts from the assumed previous bounds $[a, b]$ to $[a - b, b - a]$ with the new property that a value of zero indicates no benefit from the experience of $\hat{\mathcal{S}}$. While *improvement* compares the outcome of the $\hat{\mathcal{S}}$-informed training process with an uninformed training process, it does not take into account the other potential choices of source tasks available in the *Knowledge Base*. Following the pitfall considerations in Chap. 4, imagine a scenario where the majority of source tasks are beneficial, so that a 'random' source task selection process would – in expectation – lead to a positive *improvement*, hiding the fact that no strategy was necessary to achieve it.

Percentile

The *quality* of a task fingerprinting method should not only relate performance on a selected source task $\hat{\mathcal{S}}$ to the uninformed training process – as *improvement* does – but also to all other potential knowledge transfer sources. We will do this in several ways. First, we examine the relative ranking of the selected source task among all potential source tasks.

Definition 5.15. Let $\{\mathcal{S}_i\}_{i \leq n}$ be a pool of source tasks, $\mathcal{T}$ a target task, μ a positively oriented performance measure for $\mathcal{T}$ and $\hat{\mathcal{S}} \in \{\mathcal{S}_i\}_{i \leq n}$ the selected source task. The **percentile** for this choice is given by the fraction of source tasks that perform worse or equal compared to $\hat{\mathcal{S}}$, or more precisely

$$n^{-1} \cdot |\{\mathcal{S}_i | i \leq n, \mu(\mathfrak{T}_{\mathcal{S}_i}(\mathcal{T}, \omega_{\mathcal{S}_i})) \leq \mu(\mathfrak{T}_{\hat{\mathcal{S}}}(\mathcal{T}, \omega_{\hat{\mathcal{S}}}))\}|.$$

For negatively oriented μ the '$\leq$' can be replaced by a '$\geq$'. *Percentile* is positively

oriented and takes values within $[0, 1]$. Since a random selection process would naturally – in expectation – yield a *percentile* value of 0.5, the metric is overall well interpretable. On the downside, it requires access to *all* models that can be trained with each of the potential source tasks, which is computationally expensive. We also present a pitfall for this meta metric: Consider a scenario where the pool of source tasks consists of a number of clusters in the sense that the scores $\mu(\mathfrak{T}_{\mathcal{S}_i}(\mathcal{T}, \omega_{\mathcal{S}_i}))$ within each cluster are very close to each other, but the clusters themselves are rather separated. In such a case, the *percentile* values for the source tasks selected from within a cluster may be much more spread out than the actual performances indicate. And the 'performance jumps' between clusters would be reflected only by marginal increases in *percentile*. This is due to the fact that *percentile* only considers the relative ranking of the source tasks. Therefore, we now introduce a metric that includes actual performance differences, but unlike *improvement*, takes into account a potentially better selection of source tasks.

Regret

Definition 5.16. Let $\{\mathcal{S}_i\}_{i\leq n}$ be a pool of source tasks, $\mathcal{T}$ a target task, μ a positively oriented performance measure for $\mathcal{T}$ with finite upper bound $b \in \mathbb{R}$, and $\hat{\mathcal{S}} \in \{\mathcal{S}_i\}_{i\leq n}$ the selected source task. We call the performance resulting from the best source task $o := \max\{\mu(\mathfrak{T}_{\mathcal{S}_i}(\mathcal{T}, \omega_{\mathcal{S}_i}))|i \leq n\}$ the **oracle performance**. Then the **regret** [318] for the choice of $\hat{\mathcal{S}}$ is given by

$$\frac{o - \mu(\mathfrak{T}_{\hat{\mathcal{S}}}(\mathcal{T}, \omega_{\hat{\mathcal{S}}}))}{b - \mu(\mathfrak{T}_{\hat{\mathcal{S}}}(\mathcal{T}, \omega_{\hat{\mathcal{S}}}))}.$$

For the special case $o = b = \mu(\mathfrak{T}_{\hat{\mathcal{S}}}(\mathcal{T}, \omega_{\hat{\mathcal{S}}}))$ we set the *regret* equal zero.

Regret is negatively oriented with a value range of $[0, 1]$ and measures how much of the remaining performance gap (from the performance determined by the chosen source to the best possible performance value, i. e., the upper bound) could be overcome by making the optimal choice from the task pool. Here *regret*, unlike *percentile*, does not care about the distribution of the remaining source task performances. This means, in pitfall terms, that for a fixed selected source task and a fixed optimal source task corresponding to the oracle performance, we can add an infinite number of either near-optimal source tasks or source tasks that result in worse performance than the selected one without changing *regret*. However, in the former case, a random source selection strategy would be much better than the evaluated one, and in the latter case, it would be much worse. This makes *regret* difficult to interpret.

Computational budget

The meta metrics presented so far only consider a single source task choice $\hat{\mathcal{S}}$. In a more realistic scenario and given sufficient computational resources, the *Meta Learner* would be able to test multiple source tasks. Therefore, we will investigate **multi-shot** evaluation schemes specifically for *improvement* and *percentile*. For some computational budget $k \in \mathbb{N}$, a pool of tasks $\{\mathcal{S}_i\}_{i \leq n}$ (where $k < n$)[7] and a fingerprinting method (f, d), the k source tasks with the smallest distance to the target task are selected, and individual *improvement* or *percentile* values are computed for each of the k selected source tasks. Then some aggregation of the values is considered to reflect the result of the optimization routine. The aggregation by max reflects the 'optimistic' result of the final selection of this best source task during the optimization process. On the contrary, an average aggregation puts more emphasis on the holistic evaluation of the top-k proposals by (f, d). We will perform both types of aggregation in our experimental evaluation.

Weightedtau

The *multi-shot* evaluation approach allows to extend the fingerprint evaluation beyond a single selection. Our final meta metric will extend this concept and evaluate the full ranking of the source tasks. The idea is to use statistical correlation measures and to assess the computed task distances with the actual outcomes. For this purpose, a variety of correlation measures have been used in the literature, e. g., the *Pearson correlation coefficient* [233, 294, 381] and *Spearman's rank correlation coefficient* [103, 158][8]. Advocated by You et al. [433], however, recent analyses [3, 9, 282] have turned their attention to a different rank correlation measure.

Definition 5.17. Let $\{\mathcal{S}_i\}_{i \leq n}$ be a pool of source tasks, $\mathcal{T}$ a target task, μ a negatively oriented[a] performance measure for $\mathcal{T}$, and (f, d) a fingerprinting method. We call the performance resulting from the sources tasks $o_i := \mu(\mathfrak{T}_{\mathcal{S}_i}(\mathcal{T}, \omega_{\mathcal{S}_i}))$ for $i \leq n$ the **outcomes** and shortly note $d_i := d(f(\mathcal{S}_i), f(\mathcal{T}))$ for the **estimated distances**. We define a weighted inner product on $\mathbb{R}^n$, by

$$\langle o, d \rangle_w := \sum_{i<j \leq n} \operatorname{sgn}(o_i - o_j)\operatorname{sgn}(d_i - d_j) w_{i,j},$$

where $w \in \mathbb{R}^{n \times n}$ is a symmetric and non-negative weight matrix and

$$\operatorname{sgn}(x) := \begin{cases} 1 & , \text{if } x > 0 \\ 0 & , \text{if } x = 0 \\ -1 & , \text{if } x < 0 \end{cases}$$

[7] Taking the *multi-shot* evaluation approach to the (unfeasible) extreme and setting $k = n$ would yield the 'brute force' method.

[8] See Schober et al. [341] for a detailed comparison of these two correlation coefficients.

is the usual sign function. Now let

$$\tau_w(o, d) := \frac{\langle o, d\rangle_w}{\sqrt{\langle o, o\rangle_w} \cdot \sqrt{\langle d, d\rangle_w}}.$$

The (unweighted) **Kendall's tau** [194] for some $o, d \in \mathbb{R}^n$ is then given by $\tau_w(o, d)$, where all $w_{i,j} = 1$.
For our use cases we define the following weighting: Given some permutation r on $\{0, ..., n-1\}$, we let the **hyperbolic ranking weights** be given by $w(r)_{i,j} := (r(i) + 1)^{-1} + (r(j) + 1)^{-1}$. Now let r_o be the 0-indexed decreasing ranking function of elements in o, i. e., if $r_o(i) = j$ there are j elements in o that are larger than o_i (ties are resolved by ranking in decreasing order according to the respective entries in d). In other words, $r_o(i)$ is the index of element (o_i, d_i) in the lexicographically ordering of $\{(o_j, d_j)\}_{j \leq n}$. Vice versa we define r_d. The corresponding hyperbolic weight matrices will be denoted by $w(r_o)$ and $w(r_d)$. Finally, the **weightedtau** [404] is then given by

$$\frac{\tau_{w(r_o)}(o, d) + \tau_{w(r_d)}(o, d)}{2}$$

[a] Positively oriented performance measures may simply be inverted by multiplication with -1.

Originally, *Kendall's tau* [194] was proposed without weights. An intuitive interpretation for the unweighted variant is that for a value of τ, the probability that the fingerprinting method will rank two random source tasks according to their transfer success is $(\tau + 1)/2$ [433]. A weighted version was later given [351], and recently improved to break ties [404]. The value range of *weightedtau* is $[-1, 1]$, where 1 indicates perfect rank correlation and -1 indicates inverse rank correlation. The ranking scheme and hyperbolic weights ensure that the 'mixup' in the order of the source tasks becomes more severe if either the source task has a good outcome or a small distance. It can be seen as a smooth extension of the *multi-shot* approach to focus both on the top suggestions by a fingerprinting method and the best possible source tasks. While rank-based correlation statistics are less sensitive to outliers than those that measure linear correlation (e. g., *Pearson correlation coefficient*), *weightedtau* still suffers from similar pitfalls as *percentile*, since it only examines the ranking of outcomes and task distances, ignoring their concrete distributions.

Considerable baselines

We have already mentioned two important baselines to consider when evaluating task fingerprinting methods. First, the 'uninformed' training approach described along *improvement*. Second, the 'random' selection described along *percentile*. By design, careful

evaluation of these two metrics covers these baselines. In addition, we strive to include a baseline that captures current research practice: Eisenmann et al. [106] surveyed the strategies of participants in international biomedical imaging competitions.The predominant approach to model development for a new task was to manually inspect existing related literature and modify such existing work. Most of the time was spent selecting existing architectures that fit the task and configuring data augmentation[9]. Since we assume that such literature reviews are performed by queries to search engines, the 'relatedness' of tasks is primarily defined by a high-level semantic description. To simulate this behavior, we defined the following task fingerprinting method.

Definition 5.18. Let $\mathcal{K}$ be a finite set of **keywords** with $m := |\mathcal{K}|$. Let f be an embedding function that maps a task $\mathcal{T}$ to a subset of $\mathcal{K}$, in the sense that $f(\mathcal{T}) \in \{0,1\}^m$ and $f(\mathcal{T})_i = 1$ iff the i-th keyword in $\mathcal{K}$ is assigned to $\mathcal{T}$. Further we define the distance function

$$d_{\text{MAN}}(u,v) := 1 - \frac{\sum_{i \leq m} u_i \cdot v_i}{2m - \sum_{i \leq m} u_i \cdot v_i}.$$

We call the task fingerprinting method (f, d_{MAN}) the **Manual fingerprinting**[a].

[a] The chosen distance function may be interpreted as inverse IOU (see Def. 2.47).

For the tasks described in Sec. 2.2, we extracted keywords from the semantic description of the tasks, e. g., imaging modality, anatomical regions, and entities of interest. Since the fingerprinting often resulted in many equally close source tasks, we used task size as a tiebreaker, favoring larger source tasks – as it would likely be done by many researchers.

Aggregation matters

As we can see, *percentile*, *weightedtau*, and (most of the time) *regret* require training *all* possible source-informed models from the pool of source tasks. Since SGD (see Def. 2.77) is a randomized process, it is also necessary to repeat each model training several times for a solid assessment. Furthermore, for a robust evaluation of a fingerprinting method, the *Meta Learner* should be able to adapt different potential strategies to inform the trainer. Assuming that each selected source task results in a different choice of hyperparameters for the trainer given a strategy from the *Meta Learner*, the total necessary model trainings (per target task) amount to

$$\text{number of source tasks} \times \text{number of strategies} \times \text{number of repetitions}.$$

The evaluation of each such model can be done with multiple performance measures. Finally, using a meta metric, we are able to assign a scalar to each fingerprinting method

[9] We will cover both aspects as *scenarios* one and three in our experiments (see Sec. 5.1.3).

and experiment. In other words, for any combination of target task $\mathcal{T}$, random seed r, *Meta Learner* strategy s, performance measure μ, and metric ν, we will be able to rank a set of fingerprinting methods. However, previous studies have shown that such rankings are rather unstable and small deviations in the setup can lead to different results [3, 9, 58]. Agostinelli et al. [9] came up with a quantification for this instability with respect to each factor in the experimental setup.

Definition 5.19. Let $\{(f_i, d_i)\}_{i \leq n}$ be a set of fingerprinting methods. Consider the weighted symmetric graph $G := (V, E)$, where the vertices correspond to the conducted and evaluated experimental setups $(\mathcal{T}, r, s, \mu, \nu)$ and two nodes are connected via an edge in case *all but one* of the describing elements $\mathcal{T}, r, s, \mu, \nu$ are equal. The weight of an edge $e = (v_1, v_2)$ is given by the (unweighted) *Kendall's tau* (see Def. 5.17) between meta metrics scores for all $\{(f_i, d_i)\}_{i \leq n}$ evaluated on the setup v_1 and the corresponding scores evaluated with setup v_2.
For each component of a setup $(\mathcal{T}, r, s, \mu, \nu)$, the corresponding **setup stability** [9] is given by the average weight over the edges in G that correspond to variations in this component.

For each component, a *setup stability* of one corresponds to a consistent, i. e., stable ranking of fingerprinting methods when varying this component. In particular the setup stability of the randomness factor r, measures the intrinsic uncertainty of ranking fingerprinting methods, just by the nondeterminism during model training. Note that it is possible to circumvent the randomness component by replacing all repetitions of $\mu(\mathfrak{T}_{\hat{\mathcal{S}}}(\mathcal{T}, \omega_{\hat{\mathcal{S}}}))$ with the respective mean value.

Agostinelli et al. [9] also propose an overall score per fingerprinting method to summarize all outcomes.

Definition 5.20. For a subgraph $G' \subset G$ of the graph described in Def. 5.19 and a set of fingerprinting methods $\{(f_i, d_i)\}_{i \leq n}$ we define the **win rate** [9] for method $(f, d) \in \{(f_i, d_i)\}_{i \leq n}$ as the fraction of nodes $v \in G'$ where the ranking of fingerprinting methods according to v results in a (possibly shared) first position of (f, d).

Note that the sum of the win rates for all fingerprinting methods for a fixed subgraph G' may be greater than one due to shared first ranks. Therefore, and because the *win rate* considers only the first entry of each setup ranking, we will rather rely on an aggregation scheme proposed by Wiesenfarth et al. [420] in the context of compiling uncertainty-aware rankings for international biomedical imaging competitions.

Definition 5.21. Once more consider the graph G described in Def. 5.19 and a set of fingerprinting methods $\{(f_i, d_i)\}_{i \leq n}$. We describe two approaches to compute an overall ranking score of fingerprinting methods $\{(f_i, d_i)\}_{i \leq n}$ on the conducted experiments G. For each meta metric ν we will inspect the corresponding subgraph $G' \subset G$ that uses this meta metric.
The **rank then mean** [420] approach first ranks all fingerprinting methods $\{(f_i, d_i)\}_{i \leq n}$ on each node of G'. The achieved ranks for each fingerprinting method are afterwards averaged to compute the final score.
Conversely, **mean then rank** [420] first averages all meta metric scores individually for each fingerprinting method achieved on each node of G'. Afterwards the means for each fingerprinting method are ranked.
For both approaches, applying **bootstrapping** corresponds to the repeated evaluation while randomly re-combining the subgraph G' via drawing $|G'|$ many elements from G' with replacement.

5.1.3 Experimental design

Next, we describe the details for our experimental assessment of task fingerprinting and our proposed bKLD method.

Task pools

To ensure a broad applicability of bKLD, we chose a very heterogeneous task pool from publicly available datasets (see Sec. 2.2). Fig. 5.5 gives an overview of all used tasks, which amount to 71 in total. Inclusion criteria for tasks were (i) public availability, (ii) use of 2D images, and (iii) provided classification labels. An emphasis was placed on medical datasets (62 tasks) and a wide variety of imaging modalities (see Tab. 2.1).

Each task was divided into train $\mathcal{T}_{\text{train}}$ and test $\mathcal{T}_{\text{test}}$ sets, ensuring proportions of 80:20 and equal class distributions between splits. All images were preprocessed to a resolution of 256 × 256 pixels. Grayscale images were further converted to RGB. For tasks that exceeded 1000 samples, we subsampled a shrunken version $\mathcal{T}^{(s)}_{\text{train}}$ of 800 train samples that served as the target task variant (as previously done by Renggli et al. [318]). Meanwhile, $\mathcal{T}_{\text{test}}$ was left untouched. This was done to (i) reduce computational resources, (ii) increase comparability across target tasks, and (iii) allow knowledge transfer to have a significant impact, as e. g., pretraining is most beneficial in scenarios where the target task $\mathcal{T}$ is small [310]. This implies that all task distances are estimated from a 'complete' source task $\mathcal{S}^{(s)}_{\text{train}}$ to a 'shrunken' target task $\mathcal{T}^{(s)}_{\text{train}}$.

We divided the set of all tasks into a development set $\mathbf{P}_{\text{dev}}$ (T01-T28) and a validation set $\mathbf{P}_{\text{val}}$ (T29-T71). We made sure that all tasks related to the tasks we used in our previous study [136] were selected as developmental tasks, so that the final evaluation is performed on an 'unseen' set of validation tasks $\mathbf{P}_{\text{val}}$. Note that there is a strong (imaging)

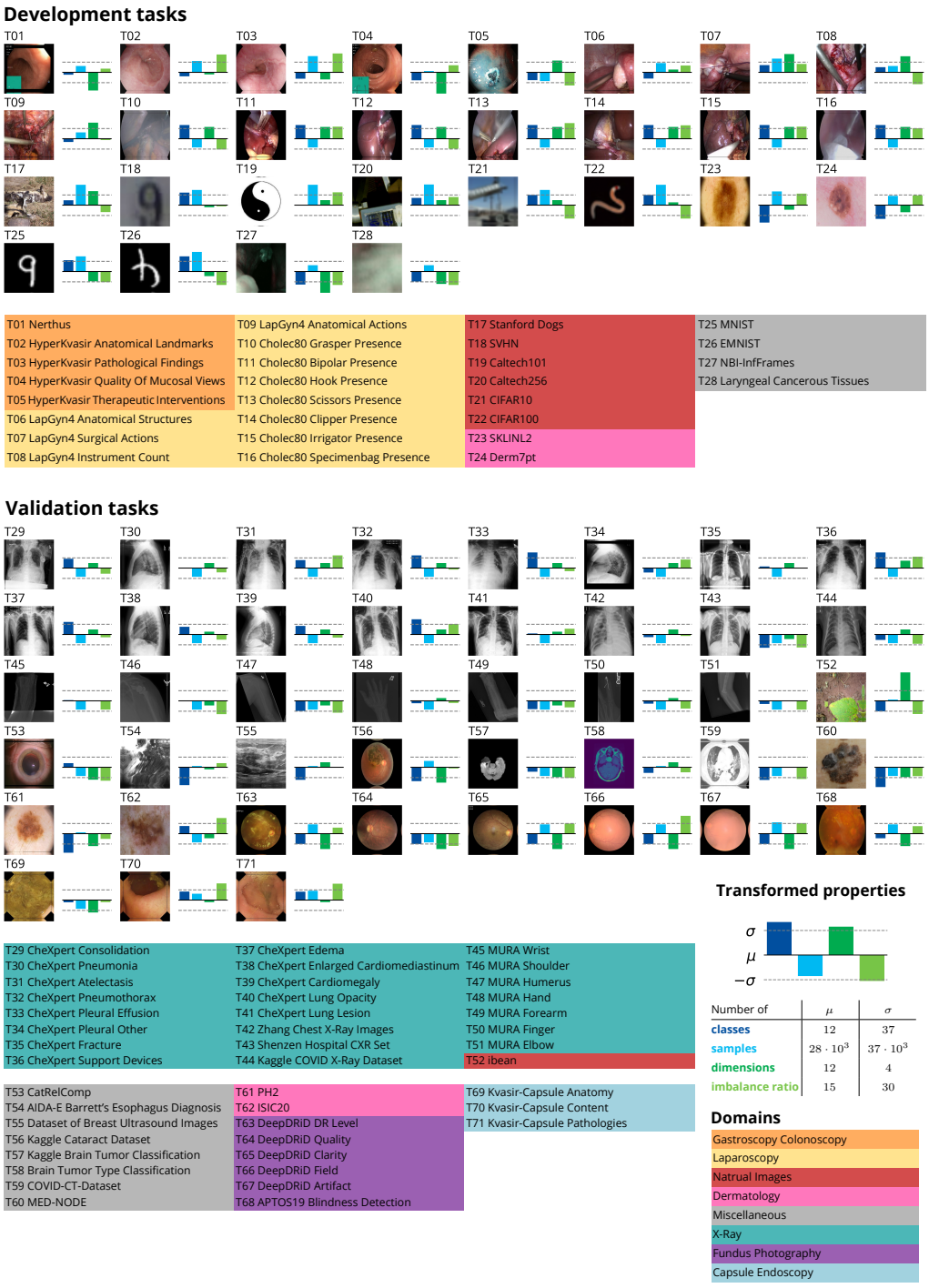


Figure 5.5: A set of 71 heterogeneous imaging tasks is the basis for our assessment. For each of the 28 tasks of the development split as well as 43 tasks of the validation split, we show a sample image next to the distribution of the following Box-Cox and Z-score transformed properties: number of classes, number of samples, intrinsic data dimension [301], and Imbalance Ratio (IR). The imaging domain is encoded as the background color of the dataset name. Published as a preprint in Godau et al. [137].

domain shift from development tasks $\mathbf{P}_{\text{dev}}$ to validation tasks $\mathbf{P}_{\text{val}}$, which challenges the generalizability of our task fingerprinting method along with the high variability in task size, number of classes, and IR. To the best of our knowledge, no task distance measure has been evaluated this extensively (see Sec. 3.2).

The tasks from the validation split were further masked for the hyperparameter selection process for bKLD (see Sec. 5.2.1), where we fixed $\mathbf{P}_{\text{dev}}$ as the pool of both source and target tasks. For the final evaluation, we set $\mathbf{P}_{\text{val}}$ as the pool of target tasks and allowed both development and validation tasks as the pool of available source tasks, i. e., $\mathbf{P}_{\text{dev}} \cup \mathbf{P}_{\text{val}}$. In every case of image overlap between two tasks $\mathcal{S}$ and $\mathcal{T}$, we excluded $\mathcal{S}$ from the pool of potential source tasks for $\mathcal{T}$ (e. g., the Cholec80 [390] (T10-T16), CheXpert [181] (T29-41), and DeepDRiD [232] (T63-T67) tasks). This also ensured that the target task itself could not be selected as source task, i. e., we assume the target task has not been encountered before by the Lifelong Learning system.

Model training

The purpose of our experiments was to demonstrate the utility of task fingerprinting for a variety of *Meta Learner* strategies. We therefore designed four realistic scenarios, covering both cases where only source task hyperparameters are available (the scenarios '1: Model Architecture' and '3: Augmentation Policy') and cases where source task data is available (the scenarios '2: Pretraining Data' and '4: Co-Training Data'). For the former, we created separate *Knowledge Bases* and conducted initial experiments as described below. For better disentanglement, the *Meta Learner* strategies vary only one aspect of the training pipeline per scenario.

As a baseline, we also trained models on all (shrunken) target tasks $\mathcal{T}_{\text{train}}^{(s)}$ in isolation, i. e., as an uninformed trainer would process them. Coming up with such a fair and meaningful training scheme for each task is a major challenge, given the large heterogeneity between tasks. Too much individual optimization per task is costly and introduces another potential source of bias; on the other hand, applying a uniform training scheme to all tasks might lead to unrealistic performance. We aimed to strike a balance with the following strategy: For faster convergence and more stable training, we use ImageNet [93] pretrained models throughout [310]. Based on the recommendations of Wightman et al. [422], we made slight adjustments to the training pipeline to ensure SGD convergence, avoid overfitting, and generally improve performance on $\mathbf{P}_{\text{dev}}$. At the individual task level, we used the automatic learning rate tuning [359] implemented in *Pytorch Lightning* [109]. Once we found a solid configuration, we kept it for all subsequent experiments. This pipeline served as a baseline for calculating *gain* and *improvement* (see Def. 5.14).

Performance measures

The related literature we identified (see Sec. 3.2) consistently used a single performance measure μ in their evaluation[10]. The vast majority chose AC as such a metric, while we also noticed the use of BA for a study focused on medical imaging [58]. Following our own recommendations from Chap. 4, we measured model performance with two complementary metrics (BA and AUROC) for the following reasons:

Since cross-target-task aggregation would suffer from interpretability if we chose per-task performance measures, we followed the most generally applicable path through subprocesses **S2-S4** (see Sec. 4.2.1), excluding **S5** (see Fig. 4.12), as we were primarily interested in discriminative performance. For **S2** (see Fig. 4.9), we assumed no unequal severity of class confusions (F2.5.2), equal interest across classes (FP2.5.1), and a mismatch of class prevalences to the population of interest (FP4.2). Following DG2.2 (see Tab: 4.5), we finally chose BA. In **S3** (see Fig. 4.10) we ended up with TPR after choosing argmax as decision rule (FP2.6) because we wanted to save computational time and also because we have tasks with very few samples, which complicates any optimization. After considering DG3.3 (see Tab. 4.8), recognizing the fact that BA is equal to the mean of all TPRs (see Def. 2.21), and the requirement to combine values into a summarizing performance measure for all classes, we omitted adding another performance measure in **S3**. In **S4**, the AUROC results from our previous decision to FP4.2. We use the average of AUROC over all classes in the one-versus-rest setup (see Def. 2.15) (also called 'macro-average').

Scenario 1: Model Architecture

The first strategy of the *Meta Learner* concerns the transfer of the neural architecture (including its pretrained weights), which has already been proposed in the literature [96, 136]. Given a selected source task $\hat{\mathcal{S}}$, the entries in the *Knowledge Base* corresponding to $\hat{\mathcal{S}}$ are searched for the best associated performance. From this specific entry, the architecture (and initialization) is transferred to the hyperparameters $\omega_{\hat{\mathcal{S}}}$ used by the trainer $\mathfrak{T}_{\hat{\mathcal{S}}}$.

Note that this scenario is closely related to 'source-free model transferability estimation' [98], but differs in some key assumptions. In both cases, the goal is to infer the optimal model for a given target task, but in our case we are given the fingerprints $f(\mathcal{S})$ from the source tasks, while in the corresponding setup we are only given access to the respective models $\{\varphi_i\}_{i\leq n}$. Therefore, 'source-free model transferability estimation' usually requires passing (some) samples from the target task $\mathcal{T}$ through each model φ_i – a process we expect to be less scalable than most of our task fingerprinting methods. In both cases, we assume that the neural architectures and initialization weights are freely available. Apart from the availability of these descriptions, 'source-free model transferability estimation' does not rely on (collaborative) knowledge acquisition. While

[10] An exception is the study by Dias et al. [96], which used AC, TPR, PPV, and F1 in the case of homogeneous balanced binary tasks. Apparently a set of not necessarily complementary metrics (see Tab. 4.2).

Table 5.1: Overview on neural architectures used for scenario 1. ImageNet Accuracy has been provided by the *timm* library [421]. Train parameters only comprise the shared backbone. Published as a preprint in Godau et al. [137].

Architecture	*timm* [421] reference	ImageNet AC (%)	train params (mio.)
EfficientNet B2 noisy student [426]	tf_efficientnet_b2_ns	82.38	7.7
ResNet50 SWSL [428]	swsl_resnet50	81.166	23.5
ResNeSt50 [441]	resnest50d	80.974	25.4
ECA ResNet50 [412]	ecaresnet50d	80.592	23.5
ResNeXt50 SSL [428]	ssl_resnext50_32x4d	80.318	23.0
EfficientNet B2 AdvProp [425]	tf_efficientnet_b2_ap	80.3	7.7
EfficientNet B2 [379]	tf_efficientnet_b2	80.086	7.7
CSP DarkNet53 [34]	cspdarknet53	80.058	26.6
CSPResNeXt50 [409]	cspresnext50	80.04	18.5
CSPResNet50 [409]	cspresnet50	79.574	20.6
VoVNet [220]	ese_vovnet39b	79.32	23.5
MixNet-L [380]	mixnet_l	78.976	5.8
RegNetY [308]	regnety_032	78.886	17.9
RegNetX [308]	regnetx_032	78.172	14.3
Res2Net50 [123]	res2net50_26w_4s	77.964	23.7
RexNet100 [154]	rexnet_100	77.858	3.5
EfficientNet B0 CondCov [429]	tf_efficientnet_cc_b0_4e	77.306	12.0
SK ResNet34 [226]	skresnet34	76.912	21.8
HRNetV2-W18 [375]	hrnet_w18	76.758	19.3
MobileNetV3-Large [175]	mobilenetv3_large_100	75.766	4.2
ResNet34 [161]	resnet34	75.11	21.3

both approaches attempt to optimize $\mathfrak{T}(\mathcal{T}, \omega)$ with an architecture induced in ω, task fingerprinting uses $d(f(\mathcal{S}), f(\mathcal{T}))$ as a proxy for the outcome. In contrast, 'source-free model transferability estimation' often compares (and evaluates) the generated predictions $\{\varphi_i(\mathcal{T})\}_{i \leq n}$.

To generate the necessary meta information about the best architecture for each potential source task $\mathcal{S}$, we established a set of 20 candidate architectures based on the following criteria: (i) availability within the *PyTorch Image Models (timm)* library [421], (ii) a reported ImageNet [93] AC of over 75% (according to *timm*), (iii) while using a maximum of 30 million trainable parameters to allow for (iv) applicability of the architecture with our compiled training pipeline within the constraints of our hardware (max. GPU VRAM of 24GB). Finally, we reduced the list to (v) architectures published since 2019 to capture the current state of the art. For architectures available in different scales, we chose the single largest variant that met our hardware requirements. The list of resulting 20 neural architectures is presented in Tab. 5.1. Note that all neural architectures have at least one source task on which they perform best, demonstrating their competitiveness in contrast to architecture choices in previous studies [96].

Scenario 2: Pretraining Data

The second strategy of the *Meta Learner* involves the commonly used concept of fine-tuning (see Def. 2.87). Given a selected source task $\hat{\mathcal{S}}$, the trainer $\mathfrak{T}_{\hat{\mathcal{S}}}$ first performs a pretraining phase, during which a model $\varphi_{\hat{\mathcal{S}}}$ is trained on the task $\hat{\mathcal{S}}$, followed by fine-tuning $\varphi_{\hat{\mathcal{S}}}$ (with a newly added final classification layer) on the target task $\mathcal{T}$. For both phases, we keep the hyperparameters ω developed for the uninformed trainer. This setup has been widely used in previous work [9, 32, 438].

Scenario 3: Augmentation Policy

The third strategy of the *Meta Learner* centers around the data augmentation pipeline (see Def. 2.82). Optimal data augmentation is an important part of training state-of-the-art models [422], but remains highly task dependent [353, 430]. Automating this part of the training pipeline offers potential benefits, especially in the low-data regime when augmentation is necessary to avoid model overfitting [353]. Nevertheless, the process of automatically generating augmentation policies is very resource intensive [83], even with proposed speed improvements [159]. Therefore, the transfer and reuse of learned policies has been proposed [83]. In our experimental setup, this is modeled as follows: For each potential source task $\mathcal{S}$, we compute a task-specific augmentation policy using the *Albumentations* [50] implementation of 'FasterAutoAugment' [159], and attach this hyperparameter to the source task in the *Knowledge Base*. The *Meta Learner* composes the hyperparameters $\omega_{\hat{\mathcal{S}}}$ from the selected source task $\hat{\mathcal{S}}$ by replacing the augmentation policy in the default parameters ω with the policy derived from $\hat{\mathcal{S}}$, which can then be used by the trainer $\mathfrak{T}_{\hat{\mathcal{S}}}$. We are not aware of any comparable task transferability evaluation

for augmentation policies.

Scenario 4: Co-Training Data

The fourth strategy of the *Meta Learner* focuses on the immediate interplay of source and target tasks during the co-training of neural networks. Research on Multitask Learning (see Def. 2.85) has provided interesting insights into this interplay: Optimal source tasks for pretraining may differ from optimal learning partners in Multitask Learning [367]. Thus, we explicitly include this setup as a complementary scenario that has received only moderate attention in the task similarity literature [115, 448]. Given the selected source task $\hat{\mathcal{S}}$ and target task $\mathcal{T}$, the *Meta Learner* instructs the trainer $\mathfrak{T}_{\hat{\mathcal{S}}}$ to attach two separate classifier heads to the shared backbone of the neural network (so-called 'hard parameter sharing', see Fig. 2.17). Training samples are taken with equal probability from each of the two tasks $\hat{\mathcal{S}}$ and $\mathcal{T}$, while only the corresponding samples are used to compute the respective losses for each of the heads. Both losses are equally weighted before backpropagation. For the evaluation of $\mathfrak{T}_{\hat{\mathcal{S}}}(\mathcal{T}, \omega_{\hat{\mathcal{S}}})$ only the target task $\mathcal{T}$ (more precisely $\mathcal{T}_{\text{test}}$) is considered.

5.2 Results

This section presents the results of the experiments described in Sec. 5.1.3. We start with our search for a new fingerprinting method on the development task pool $\mathbf{P}_{\text{dev}}$ in Sec. 5.2.1. Then, we thoroughly evaluate the *quality* of our identified bKLD fingerprinting approach(es) in Sec. 5.2.2. To do so, we build on the metrics and aggregation strategies described in Sec. 5.1.2.

5.2.1 The search for an appropriate task fingerprinting method

During the development phase, we started with previously proposed task fingerprinting methods as described in Sec. 5.1.1. Focusing primarily on feature extraction based approaches (see Ex. 5.2), we varied the following components within a fingerprinting method:

The embedding function f: In addition to the full $m \times n$ features, we explored the binning strategy (see Def. 5.7) from coarse $B = 5$, over medium $B = 100$ to fine $B = 1000$ granularity, and the average feature aggregation (see Def. 5.12).

The distance function d: For all embedding functions, we explored cosine similarity (see Def. 5.11), KLD (see Def. 5.12), L_p norm (for $1 \leq p \leq 4$), Jensen-Shannon divergence (a symmetric derivative of KLD), EMD (see Def. 5.6). We explored several smoothing functions (i. e., normalization, softmax, and symmetric uniform smoothing) for both source and target embeddings, as well as weighting schemes. We also explored some variants involving transformations in log and exp spaces.

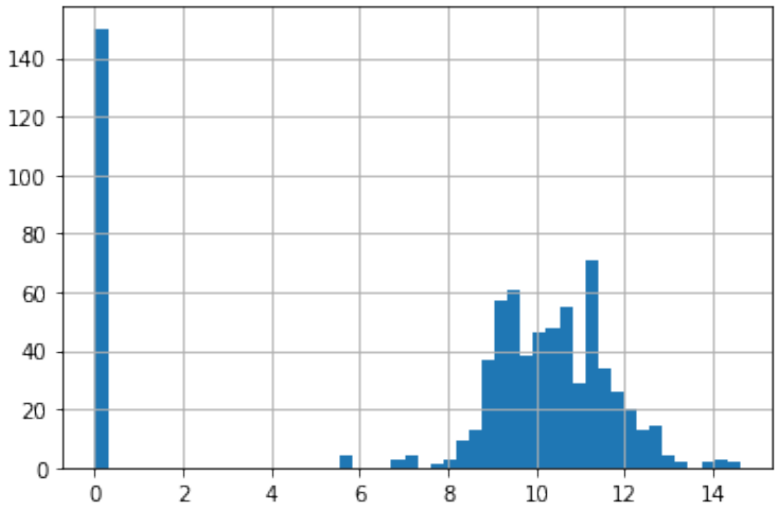


Figure 5.6: Histogram of summarized scores for all fingerprinting methods during development. The 50-bin histogram shows the distribution of all inspected fingerprinting variants during method development. The found methods show substantial lead over the majority of alternative candidates.

In total, we computed a pool of 749 candidates. Some of these were considered 'degenerated', as they concentrated their preferred source tasks into a very small set, regardless of varying target tasks. We filtered out all variants that selected less than one fourth of the development task pool $\mathbf{P}_{\text{dev}}$ (for comparison, manual fingerprinting (see Def. 5.18) selects 37% of $\mathbf{P}_{\text{dev}}$ as source tasks). After filtering, 599 variants remained. We evaluated these task fingerprinting methods on the 28 target tasks with the four described scenarios and two chosen performance measures. For robustness, we aggregated the results of the performance measures over three repetitions before applying the meta metrics. For each of the meta metrics *gain*, *improvement*, *percentile*, *regret*, and *weightedtau* (see Sec. 5.1.2) and each of the four scenarios, we first computed the mean of each variant and then the relative position of that mean by rescaling to the achieved best and worst means. That is, given the best achieved mean value o for a variant in this combination of scenario and meta metric, the corresponding worst value w, and the actually achieved mean value x, we assign a relative score of

$$\frac{x - w}{o - w}$$

The 20 resulting scores are then summed and ranked. The top three variants are the ones described next to Def. 5.13. Out of an optimal score of 20, **bKLD(small,target)** scored 14.6, **bKLD(large,unweighted)** scored 14.3, and **bKLD(large,source)** scored 14.2. The full distribution of the scores is shown in Fig. 5.6.

Table 5.2: Knowledge transfer scenarios require different task distances Mean and standard deviation of weightedtau on pairwise transfer scenario outcomes over 43 tasks, three repetitions and two performance measures for the four knowledge transfer scenarios we investigate: Model Architecture (M. A.), Pretraining Data (P. D.), Augmentation Policy (A. P.) and Co-Training Data (C. D.). Published as a preprint in Godau et al. [137].

	Model Architecture	Pretraining Data	Augmentation Policy	Co-Training Data
M. A.	1.000 ± 0.000	0.084 ± 0.147	0.078 ± 0.141	0.052 ± 0.137
P. D.	0.084 ± 0.147	1.000 ± 0.000	0.037 ± 0.164	0.038 ± 0.136
A. P.	0.078 ± 0.141	0.037 ± 0.164	1.000 ± 0.000	-0.002 ± 0.140
C. D.	0.052 ± 0.137	0.038 ± 0.136	-0.002 ± 0.140	1.000 ± 0.000

5.2.2 Evaluation of binned Kullback Leibler Divergence

We continue with the full evaluation of bKLD in our experiments. Since we repeated each possible knowledge transfer $\mathfrak{T}_{\mathcal{S}}(\mathcal{T}, \omega_{\mathcal{S}})$ three times to compensate for non-determinism during model training, we trained more than 30 000 DNNs just for the evaluation part, resulting in about 10 000 GPU hours of training[11]. Before diving into the individual results, we note that the *weightedtau* between the optimal (i. e., ground truth measured results) source task rankings for the individual knowledge transfer scenarios differ significantly, as reported in Tab. 5.2. This underscores the challenging nature of the evaluation and justifies our presentation of multiple variants of bKLD, as a single (non-parametric) task fingerprinting method may not sufficiently predict knowledge transfer across all scenarios.

Absolute assessment

Gain Architecture selection recommended by the *Meta Learner* improved model performance for 67% of validation tasks compared to our uninformed baseline approach. Similarly, adopting augmentation strategies from the best-matching source task benefited 58% of tasks on first attempt. Knowledge transfer through pretraining and co-training using image data from the best-matching source task outperformed the baseline in 41% and 57% of validation tasks, respectively. When expanding computational resources and considering multiple knowledge source candidates (*multi shot*), performance improvements reached up to 90% of validation tasks, as illustrated in Fig. 5.7 (top). Importantly, transferring model architecture (90%) and augmentation policy (83%) required only sharing the fingerprint identifier, without exchanging actual data samples between institutions. When task data was available for Transfer Learning, the co-training scenario showed

[11] Although the development pool of tasks is much smaller, we had numerous redesigns of the training pipelines, so we estimate the resources expended to be in the same order of magnitude.

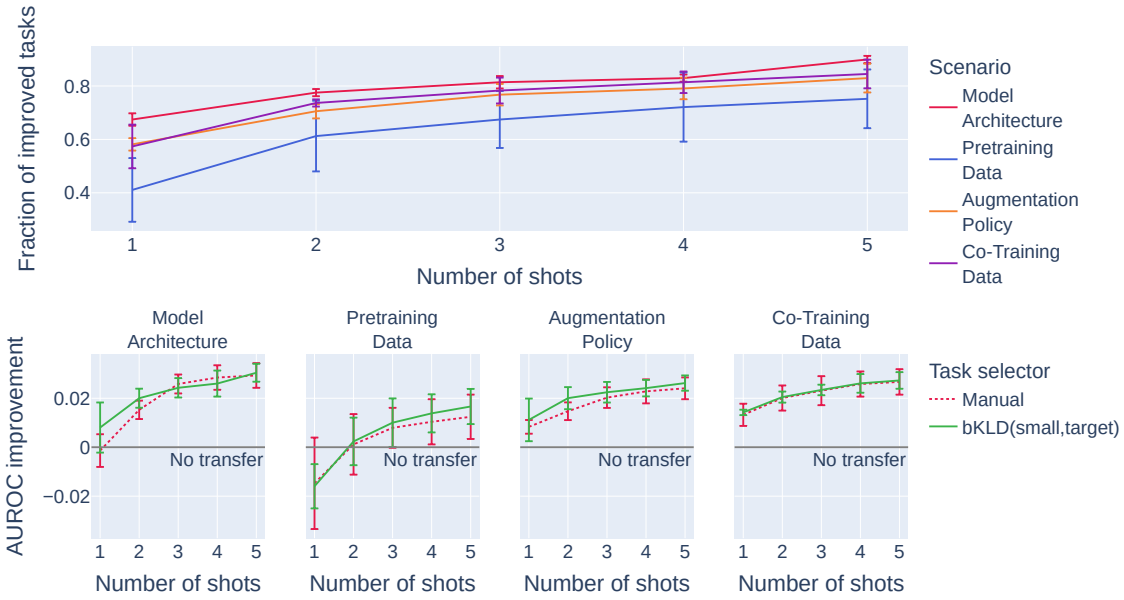


Figure 5.7: Task fingerprinting benefits training pipeline configuration and beats manual knowledge transfer. Top: Fraction of n=43 validation tasks that improve Balanced Accuracy (BA) through knowledge transfer (*gain* [438]) in four scenarios. **Bottom**: Average *improvement* in Area under the Receiver Operating Characteristic Curve (AUROC) across n=43 validation tasks. In all subplots X-axis shows the number of shots, translating to the best of top k suggestions of our framework. Error bars correspond to standard deviation over three repetitions of all model trainings. Our proposed binned Kullback-Leibler Divergence (bKLD) fingerprint (here: the small variant) improves training for up to 90% of validation tasks. Published as a preprint in Godau et al. [137].

substantial improvement rates (84%) compared to our baseline. The pretraining scenario demonstrated comparatively lower improvement rates (75%), with higher variability across repeated experiments.

Improvement The results for the more granular comparison with our uninformed baseline are shown in Fig. 5.7 (bottom), which also displays the average performance increase per task for the manual fingerprinting approach (see Def. 5.18). The relative performance increase by bKLD fingerprinting compared to the simulated manual approach was 12%, 57%, 15%, and 2% larger for the four scenarios respectively (when averaging across all target tasks, repetitions, and the *multi shot* range from 1 to 5).

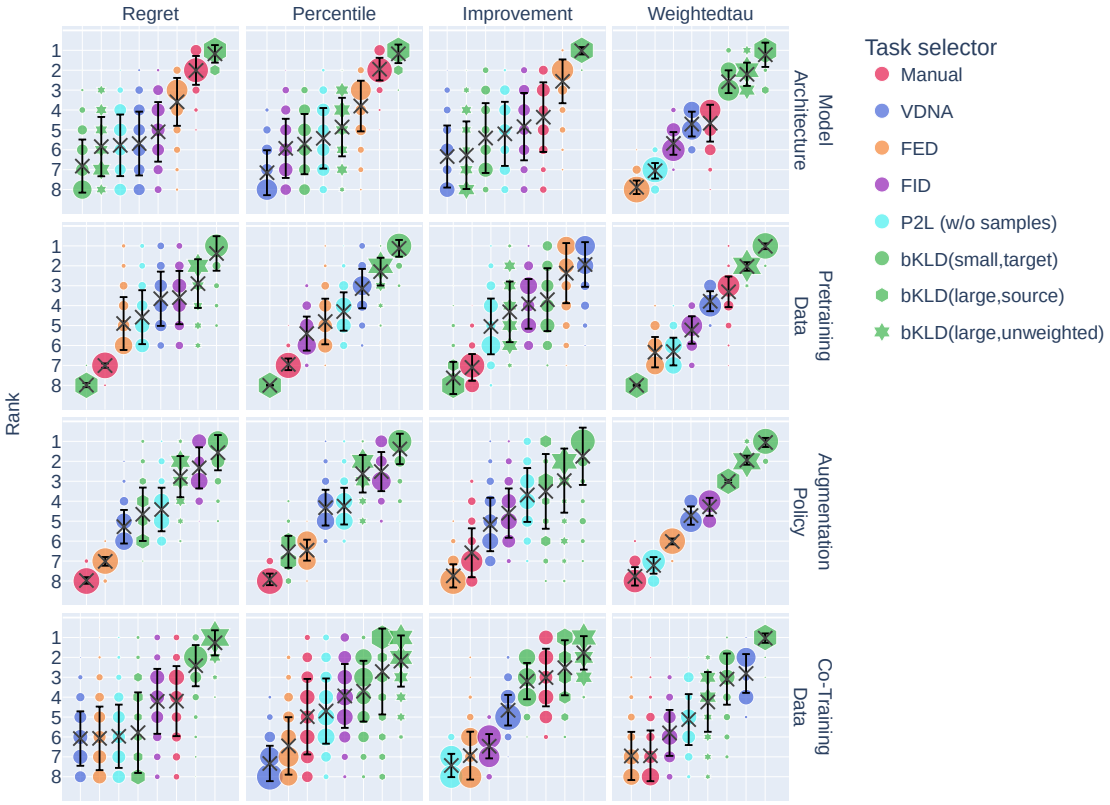


Figure 5.8: binned Kullback-Leibler Divergence (bKLD) outperforms previously proposed methods for knowledge transfer. Uncertainty-aware ranking of our proposed bKLD methods for task fingerprinting versus VisualDNA (VDNA) [313], Predict To Learn (P2L) [32], Fisher Embedding Distance (FED) [4, 136], Fréchet Inception Distance (FID) [97, 166] and manual task selection. Columns represent four meta metrics to compare task fingerprinting methods, whilst rows correspond to four knowledge transfer scenarios. We average across the top three suggestions by each method (*multi shot*), except for *weightedtau*. Blob size shows frequency of rank across 1000 bootstrap samples from 258 setups (two performance measures, 43 validation tasks, three repetitions). X marks the mean rank and whiskers the standard deviation. Plot is inspired by [420] and follows the *mean then rank* assessment method (see Def. 5.21). Published as a preprint in Godau et al. [137].

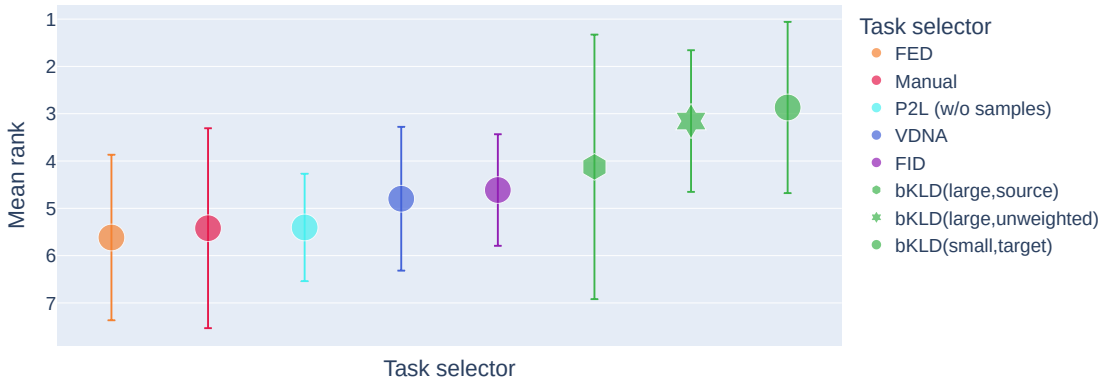


Figure 5.9: Aggregated summary of fingerprint method comparison. Uncertainty-aware ranking of our proposed bKLD methods for task fingerprinting versus VisualDNA (VDNA) [313], Predict To Learn (P2L) [32], Fisher Embedding Distance (FED) [4, 136], Fréchet Inception Distance (FID) [97, 166] and manual task selection. This figure is a summary of the 16 subplots in Fig. 5.8. The marker position refers to the mean rank over each of the 16 individual bootstrapped mean rankings (X marks) and whiskers indicate standard deviation. Published as a preprint in Godau et al. [137].

Comparative assessment

To compare our fingerprinting technique to previously proposed task similarity measures, we computed results for VisualDNA (VDNA) [313] (see Def. 5.7), Predict To Learn (P2L) [32] (see Def. 5.12), Fisher Embedding Distance (FED) [4, 136] (see Def. 5.11) and Fréchet Inception Distance (FID) [97, 166] (see Def. 5.9). Furthermore we add the manual baseline (see Def. 5.18) and compare all of them to the three configurations of our proposed bKLD fingerprinting. Fig. 5.8 shows a detailed analysis of the rankings for all those methods with respect to meta metris and knowledge transfer scenarios. Here we follow the bootstrapped *mean then rank* approach (see Def. 5.21). Within all four knowledge transfer scenarios, for at least three out of four meta metrics, a single variant of bKLD performed best. The 16 individual results are summarized in Fig. 5.9, where the three bKLD variants obtain the top three positions. The bootstrapped alternative aggregation scheme *rank then mean* can be found in Fig. 5.10. In addition Tab. 5.3 shows the *win rate* (see Def. 5.20) of each fingerprinting method. For our experiments, the *setup stability* scores (see Def. 5.19) were 0.61 for the meta metrics, 0.32 for the performance measures, 0.12 for the random seed, 0.02 for the target tasks, and only 0.01 for the knowledge transfer scenario, confirming the similarities given in Tab. 5.2.

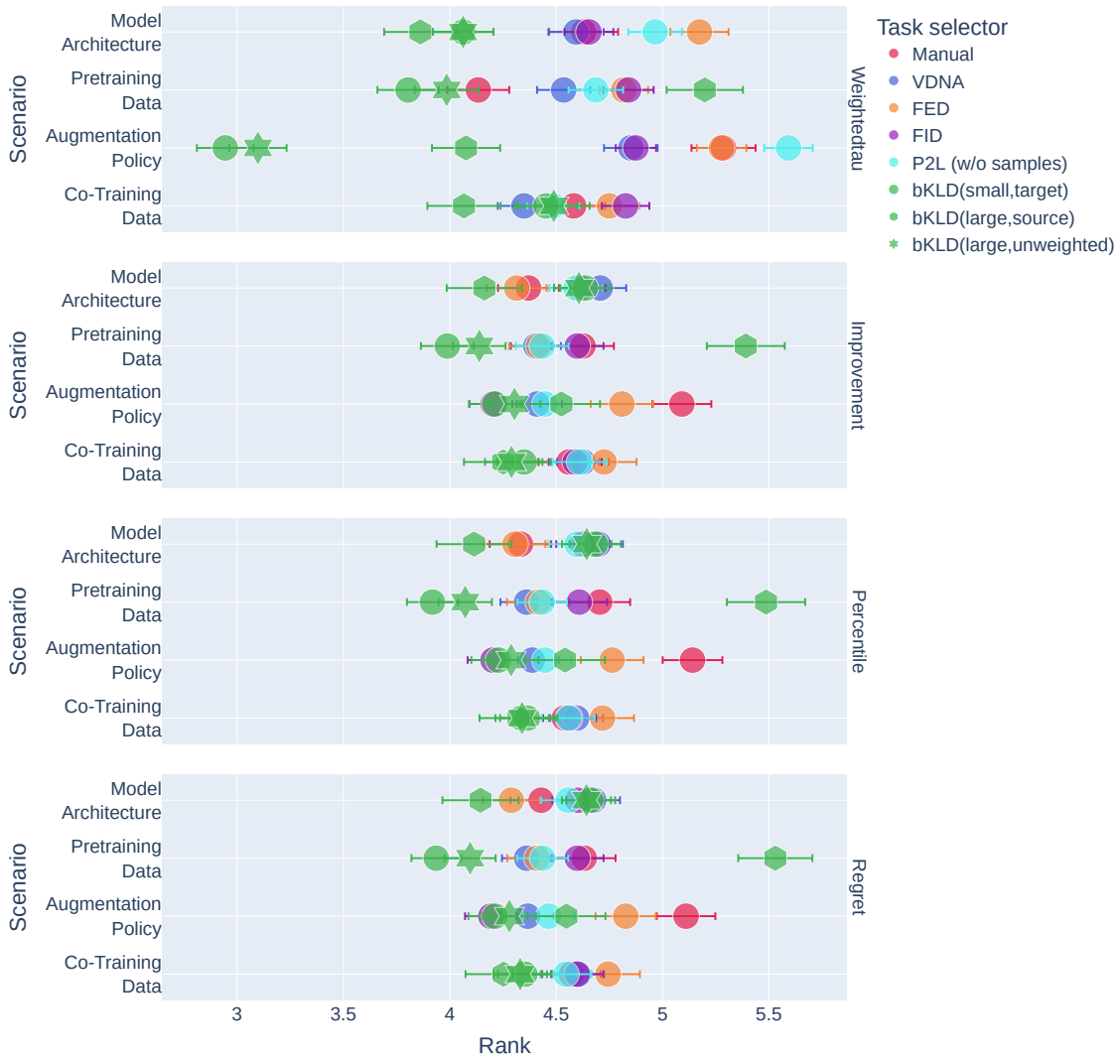


Figure 5.10: binned Kullback-Leibler Divergence (bKLD) outperforms previously proposed methods for knowledge transfer. In contrast to the evaluation shown in Fig. 5.8, this figure shows the evaluation according to *rank then mean* [420], using 258 setups (two performance measures, 43 validation tasks, three repetitions). The marker position refers to the mean over 1000 bootstraps, with the whiskers indicating standard deviation. For each setup, the *improvement*, *percentile*, and *regret* of the top three suggestions are averaged, while *weightedtau* is evaluated on the full ranking of suggested knowledge transfer sources. Published as a preprint in Godau et al. [137].

Table 5.3: Win rates for task fingerprinting methods. *Win rates* [9] for task fingerprinting methods in percent. Shows the fraction of 1032 individual setups (43 tasks, three repetitions, four meta metrics, two performance measures) that a specific fingerprint method performs best. Columns may sum above 100 because of ties, to reduce such occurrences we averaged the meta metrics of top 3 suggestions. Next to the win rates for four knwoledge transfer scenarios we also provide the average across them. Best value per column is marked boldface. Published as a preprint in Godau et al. [137].

	M. A.	P. D.	A. P.	C. D.	mean
FID	15.02	13.86	12.60	10.66	13.03
P2L	15.21	14.92	13.66	10.17	13.49
FED	15.89	15.21	11.14	12.79	13.76
VDNA	13.66	18.41	16.09	12.21	15.09
Manual	20.64	19.86	11.34	16.96	17.20
bKLD(small,target)	14.53	**25.58**	23.84	17.25	20.30
bKLD(large,unweighted)	14.53	25.39	22.19	20.64	20.69
bKLD(large,source)	**34.98**	19.86	**27.71**	**32.66**	**28.80**

Computational robustness

We examined the bKLD fingerprint generation process using varied sample sizes, given its non-deterministic nature and particular relevance for small datasets in situations of data scarcity. While our previous experiments consistently utilized 10 000 samples per task for fingerprint embedding computation – noteworthy for all fingerprinting methods (except manual fingerprinting) – we tested substantially reduced sample sizes down to just 10 samples. Fig. 5.11 demonstrates the robustness of bKLD computation in identifying beneficial source tasks in a best-of-three *multi shot* scenario, where random selection would yield an expected value of 0.75. Notably, both **bKLD(large,unweighted)** and **bKLD(small,target)** maintained reliable task matching performance even with fingerprints generated from only 10 samples. Interestingly, knowledge transfer in the pretraining scenario showed the best *relative* task matching among the four investigated scenarios, typically selecting source tasks around the 88th percentile. This contrasts with the *absolute* performance improvements shown in Fig. 5.7, where pretraining knowledge transfer provided the least benefit across all scenarios studied.

5.3 Discussion

Our intensive experiments addressing **(RQ2)** have yielded valuable insights on the utility of task fingerprinting for knowledge transfer for medical imaging, particularly through

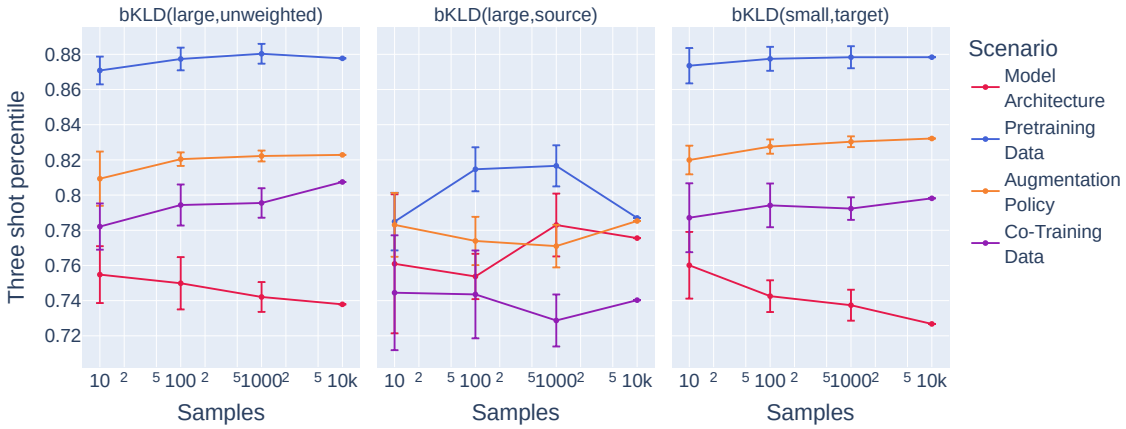


Figure 5.11: binned Kullback-Leibler Divergence (bKLD) is robust with respect to dataset size. Percentile of best 3-shot source task knowledge transfer according to Balanced Accuracy (BA) averaged over n=43 validation tasks in four scenarios applying three proposed bKLD variants. X-axis shows the number of samples used from a task to generate its fingerprint in log scale. Error bars indicate standard deviation over 10 resamplings. Published as a preprint in Godau et al. [137].

our newly introduced bKLD method. We demonstrated that bKLD:

(i) enables knowledge transfer without sharing sensitive patient data, outperforming simulated manual knowledge transfer (see Fig. 5.7),

(ii) adapts flexibly to various knowledge transfer strategies from the *Meta Learner* (see Fig. 5.8),

(iii) computes robustly with minimal data requirements, making it suitable for scenarios with limited task data (see Fig. 5.11).

Notably, bKLD outperformed previous approaches in what is, to our knowledge, the most comprehensive evaluation of task transferability estimation in biomedical imaging – combining the largest heterogeneous task set with the broadest assessment across knowledge transfer scenarios.

Interpretation

A fundamental goal of our Lifelong Learning system was to continuously build upon previous experiences to accelerate and enhance the learning of new tasks (see Def. 2.90). This characteristic directly counters the *isolated learning* paradigm dominating current healthcare AI algorithms. Our work demonstrates that task fingerprinting serves as an effective tool for AI democratization, addressing the concerning trend of AI expertise

concentration [10]. The shareable task embeddings we developed enable knowledge transfer estimation, allowing researchers to collaboratively accumulate and leverage insights during NN experiments.

Our extensive experiments confirm an emerging insight within Transfer Learning research: there is no 'one-size-fits-all' solution for transferability estimation [3, 9, 58, 367]. The low *setup stability* score (0.01) for knowledge transfer scenarios and the comparison of optimal source task rankings (Tab. 5.2) reveal the inconsistency across scenarios, suggesting that no single non-parametrized fingerprinting method can adequately address all transfer scenarios. The two hyperparameters B (for the number of bins) and w (for the weighting of feature dimensions) in bKLD resolve this challenge by controlling both granularity (B) and attention to scenario-relevant features (w).

This adaptability is clearly demonstrated in our comparison with the P2L [32] approach (see Def. 5.12), which computes KLD on averaged features. The binning plus weighting strategy in bKLD allows for more granular and controlled comparison of image feature distributions, resulting in the superior performance of bKLD variants over P2L across our experiments (see Figs. 5.8 and 5.10). Importantly, we derived hyperparameter variants for bKLD on a separate pool of target tasks to avoid overfitting – a methodological rigor often not explicitely stated in related literature.

Scenario-specific insights

Model Architecture Setting weights w according to dominant source task features – as in **bKLD(large,source)** – proved especially beneficial for model architecture transfer (see Fig. 5.8), though this scenario obtained the lowest *percentile* score, just slightly above random guessing (see Fig. 5.11). Manual task selection performed better in this scenario than others, suggesting that granular feature distributions play a subordinate role in model architecture fitting. This specific scenario, closely related to 'source-free model transferability estimation', remains an open challenge despite numerous proposed solutions [98].

Pretraining Data For pretraining, **bKLD(small,target)** proposed the highest-ranked source tasks across all scenarios (see Fig. 5.11). Interestingly, while bKLD performed well in *relative* task matching (see Fig. 5.11), the *absolute* performance improvement remained modest compared to other scenarios (see Fig. 5.7). We interpret this as reflecting a shortage of suitable pretraining tasks in our experimental pool, as ideal pretraining tasks require both large size and sufficient sample distribution variety [205, 248]. The consistently poor performance of **bKLD(large,source)** in this scenario (see Fig. 5.8) highlights the importance of feature distribution similarity, particularly regarding dominant features of the *target* task.

Augmentation Policy In transferring augmentation strategies, the semantic-based manual baseline performed worse than all data-driven fingerprinting methods (see

Fig. 5.8), indicating that granular comparison of visual features is essential for effective task matching in this context. Similar to pretraining, **bKLD(large,source)** performed worst among our proposed variants, while the target-weighted **bKLD(small,target)** consistently excelled across all meta metrics, reinforcing the necessity of focusing on dominant *target* task features.

Co-Training Data For Multitask Learning, the unweighted bKLD variant performed best, suggesting that focusing primarily on dominant features from one task does not improve transfer estimation in this scenario. However, the **bKLD(small,target)** and **bKLD(large,unweighted)** variants showed similar performance across most meta metrics (see Figs. 5.8, 5.10, and 5.11).

Research context

In general computer vision, task transferability estimation is an increasingly active field of research [98] with only few studies that focus on medical imaging in particular [58, 62, 310]. Consistent with our preliminary research [136] the FED fingerprinting method performs well for the model architecture and pretraining scenarios on certain meta metrics (see Fig. 5.8). However, it falls short on other meta metrics and knowledge transfer scenarios. The inconsistency of transferability estimations [9, 58] renders such studies of smaller scale and with less heterogeneity less meaningful.

Our aggregation across multiple performance measures and meta metrics was designed to counter the sensitivity of evaluations to specific metrics [317] and increase the assessment robustness [420]. We observed that the *weightedtau* meta metric provided the most stable separation of fingerprinting methods (see Figs. 5.8 and 5.10), supporting the current research trend toward focusing on this metric for evaluating transferability estimation [3, 9, 282, 433]. However, its interpretability remains challenging and its hyperbolic weighting scheme circumvents resource limitation-based cutoffs.

The emergence of FM [37] has amplified the importance of AI democratization as expertise requirements, computational demands, and data volume needs have grown substantially. While such models may bridge gaps between different medical imaging modalities through large-scale pretraining [21], they present two fundamental challenges: selecting the most suitable model and devising effective adaptation strategies for specific downstream tasks. Task fingerprinting offers promising solutions to both challenges.

Limitations

Despite our comprehensive assessment, several important parameters for knowledge transfer remain unexplored, for example the size of the source task[12], which plays an

[12] Bhattacharjee et al. [32] propose a mechanism to combine task similarity and source task features. Some experiments with such approaches have been conducted by us and can be found in our public repository [133].

important role specifically for pretraining [205]. Furthermore, we prioritized using consistent hyperparameters ω in our model training experiments over intense task-specific hyperparameter tuning, which would better represent real-world practice. While this decision does not undermine our broader conclusions about improving models for target tasks, future work should test additional problem categories (e. g., SemS), expand the number of target tasks, and explore knowledge transfer for other training pipeline components.

The practical implementation of a 'Knowledge Cloud' – a cross-institutional *Knowledge Base* – remains a future challenge. Existing infrastructure like the 'Joint Imaging Platform' [338], already supporting cross-institutional medical imaging research, offers an excellent starting point. Additionally, our experiments treated each transfer scenario in isolation; investigating the impact of transferring multiple pipeline components simultaneously ('entangled transfer scenarios') represents a crucial next step.

Conclusion

We consider the broader implications of our research to be significant: widespread adoption of *Knowledge Clouds* based on task fingerprinting could democratize AI research by facilitating collaboration and knowledge sharing. This shared knowledge can reduce model development times and decrease carbon emissions associated with extensive training processes. Long-term, the ability to quantify distances between tasks may enable the full potential of Lifelong Learning systems in biomedical imaging.

Our study provides compelling evidence that task fingerprinting effectively overcomes knowledge silos and enhances knowledge transfer in medical image analysis. The proposed bKLD method demonstrates clear advantages across various knowledge transfer scenarios with the flexibility to address a wide range of use cases. Future research will focus on expanding the framework's capabilities and exploring its broader impact on medical AI research.

6 Deployment of Classification Algorithms under Prevalence Shifts

Disclosure

Parts of the results of this chapter have been published at the *Medical Image Computing and Computer Assisted Interventions (MICCAI) conference* [138] and in *Medical Image Analysis* [139]. See App. A for full disclosure.

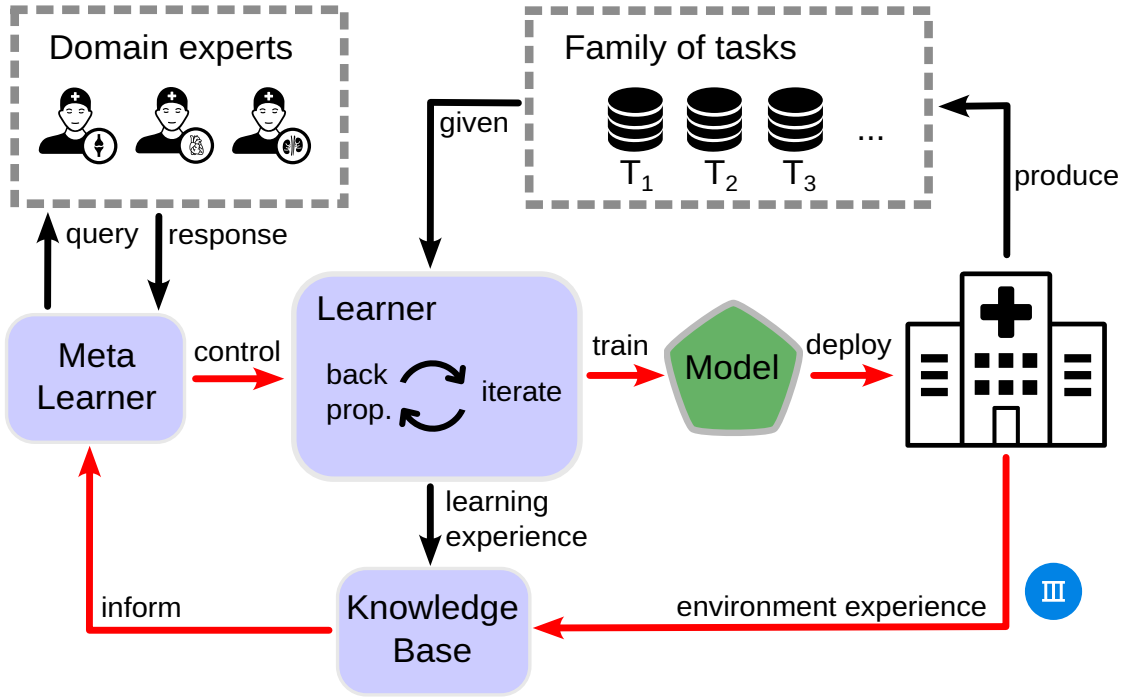


Figure 6.1: Environment-learning loop. Anchoring of this chapter in the overall Lifelong Learning system (see Fig. 1.1). Given a task, the *Meta Learner* leverages the specific experience from the environment to determine the optimal modifications for the *Learner* to update a model. Loop is highlighted in red.

This chapter addresses the third research question of designing an **environment-learning loop** as part of the **Deploy** phase in the AI lifecycle (see Fig. 6.1):

Research Question 3

What mechanisms enable biomedical imaging models to detect and compensate for prevalence shifts in deployment?

In Sec. 2.1 we elaborated on the issues with *distribution shifts* faced during model deployment and the underlying causal reasoning in medical imaging. Sec. 3.4 summarized the gap in literature with respect to comprehensive impact quantification and best practices for deploying models when *prevalence shifts* are present. In this chapter we want to present our analysis of prevalence shift implications and options to mitigate negative impacts. We aim to enable our Lifelong Learning system to regularly update models in deployment based on feedback from the environment. More precisely, for prevalence shifts, it should regularly quantify the prevalences faced in the environment, re-calibrate the model post-hoc (see Def. 2.67) and re-configure its decision rule (see Def. 2.9). In Sec. 6.1 we develop synthetic environments of prevalence shifts and present our proposed workflow. Subsequently, Sec. 6.2 presents our experimental observations for the concrete impact of prevalence shifts on models and test our workflow. Finally, Sec. 6.3 closes the chapter with a discussion of our results.

6.1 Methods

As depicted in Fig. 6.1, we want to gain 'experience' during the deployment of a model. However, unlike the **Design** and subsequent **Develop** phases, we will not assume that we will be given labels to accompany the images we see. This is because, although (i) annotating a dedicated deployment set may be feasible in some cases, and (ii) in other cases the diagnostic procedures will automatically generate labels in the patient's Electronic Health Record (EHR), we expect the former to be severely delayed and the latter to be particularly noisy. In both cases, additional dependencies of the Lifelong Learning system would be introduced. To compensate for this lack of information, the setup provides us with a trained model that can continuously compute predictions on the seen images - which are then fed back to the *Knowledge Base*. The first task of the *Meta Learner* is to use these predictions in order to detect and quantify potential distribution shifts.

6.1.1 Prevalence shifts

We continue to define distribution shifts more formally. Recall the associated joint distribution $p_{\mathcal{T}}(X, Y)$ for some task $\mathcal{T}$ from Def. 2.3.

Definition 6.1. Let $\mathcal{T}_1$, $\mathcal{T}_2$ be two tasks with the same number of classes and $p_1(X, Y)$, $p_2(X, Y)$ their respective joint distributions. If $p_1(X, Y) \neq p_2(X, Y)$ we call this scenario a **distribution shift** [382] (also **dataset shift** [261, 306]).

Obviously, dataset shifts may affect the predictive performance when a model φ is trained and validated on $\mathcal{T}_1$, while it is confronted with $\mathcal{T}_2$ data during deployment. To estimate this impact and determine appropriate mitigation strategies, it is necessary to identify the causal relationships in the shift. For this we start with a lightweight definition of causality and refer the interested reader to the literature for more comprehensive assessment [290, 296].

Definition 6.2. Let X, Y be random variables. We say X **causes** Y, written $X \to Y$, if an 'intervention' on X, i. e., forcing it to different values, changes the likelihoods on Y [54].

The perspective of causality exceeds mere probability, due to its reliance on interventions, structural assumptions, and counterfactual reasoning, which go beyond purely statistical associations. From probability theory the following two decompositions of a joint probability distribution $p(X, Y)$ hold equally true:

$$p(X, Y) = p(Y|X)p(X)$$
$$p(X, Y) = p(X|Y)p(Y)$$

From the causal perspective we are interested whether the mechanisms $p(Y|X)$ (respectively $p(X|Y)$) are 'invariant', i. e., if it is possible to change $p(X)$ (respectively $p(Y)$) by intervention without changing $p(Y|X)$ (respectively $p(X|Y)$) [296]. If $Y \to X$, then by the principle of 'independence of cause and mechanism' [296] this allows to examine $p(Y)$ (the cause) and $p(X|Y)$ (the mechanism) independently, i. e., after an intervention on $p(Y)$ we may still decompose $p(X, Y) = p(X|Y)\tilde{p}(Y)$ with the *original* $p(X|Y)$ and the modified $\tilde{p}(Y)$. We distinguish two cases of causal relationships for a task $\mathcal{T}$.

Definition 6.3. Let $\mathcal{T}$ be as task and $p_{\mathcal{T}}(X, Y)$ the associated joint distribution. We call $\mathcal{T}$ a **causal** (respectively **anticausal**) task, if $X \to Y$ (respectively $Y \to X$) [54].

We already hinted at the embedding of such relationships in the context of medical imaging in Fig. 2.1. Such causal diagrams are acyclic directed graphs, that connect nodes representing random variables with edges that represent a causal dependency. When simplifying the generic causal diagram given in Fig. 2.1, one may distinguish several common types of dataset shifts in medical imaging, as depicted in Fig. 6.2. Therein the

interventions are caused by the change in environment E.

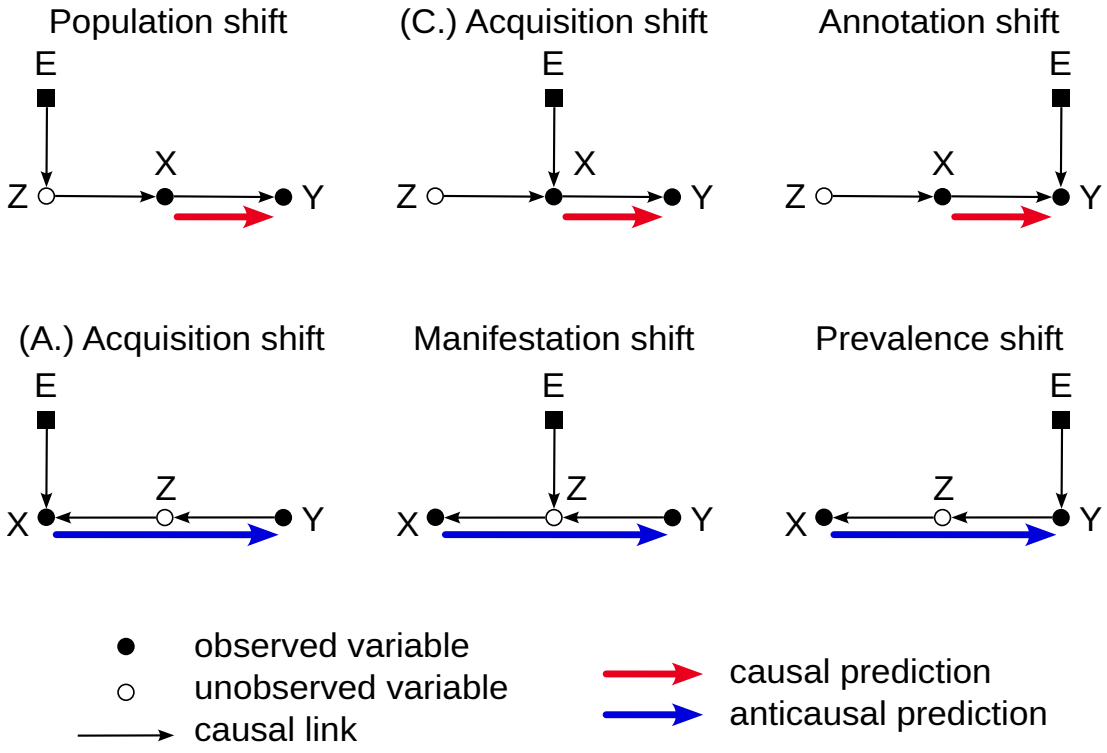


Figure 6.2: Variants of dataset shifts. Simplified variants of the full causal imaging workflow from Fig. 2.1. Depending on whether the environmental factor E causally affects the image X, the label Y, or the unseen anatomy Z, as well as the causal direction of the prediction, we distinguish the three causal (top row) and anticausal (bottom row) dataset shifts. Adapted from Castro et al. [54].

In most real-world deployment scenarios, multiple dataset shifts would occur in parallel. For the remainder of this chapter, we will focus on one specific dataset shift in isolation.

Definition 6.4. Let $\mathcal{T}_1$, $\mathcal{T}_2$ be two tasks with the same number of classes and $p_1(X, Y)$, $p_2(X, Y)$ their respective joint distributions. Assume further $\mathcal{T}_1$ and $\mathcal{T}_2$ are anticausal, i.e., we may decompose $p_i(X, Y) = p_i(X|Y)p_i(Y)$ for $i = 1, 2$ [54, 261]. We then call a distribution shift between $\mathcal{T}_1$ and $\mathcal{T}_2$ a **prevalence shift** [54] (also **prior probability shift** [99, 261], **label shift** [230] or **target shift** [443]) if $p_1(Y) \neq p_2(Y)$ but $p_1(X|Y) = p_2(X|Y)$.

Common causes of prevalence shifts are institutional differences [439], temporal changes [440], or sample selection bias [99]. The latter may also be subtly hidden by the

data sampling strategy (see Def. 2.77), especially since a balanced sampling strategy is common for highly imbalanced datasets [187].

6.1.2 Common misconceptions in medical image classification

Our study was partly inspired by the anecdotal observation that the impact of prevalence shifts on deployed models is largely unknown. More specifically, while reviewing papers, inspecting code repositories, or in professional discussions, we encountered the following misconceptions along the model output flow (see Fig. 2.11):

Logit transformation '*Temperature scaling typically corrects for model miscalibration*'
Temperature scaling [151] (see Def. 2.67) is amongst the most common re-calibration methods. While it is well suited for compensating for over- and underconfidence, it does not solve for miscalibration due to prevalence shift. By design, its single parameter $t \in \mathbb{R}$ lacks the power to compensate for a shifted class prevalence $\mathcal{P} \in \sigma_{C-1}$. Since all logits are uniformly multiplied by t^{-1}, temperature scaling preserves the ranking of class probabilities compared to the softmax transformation σ.

Decision rule '*The* argmax *operator is the optimal decision rule*'
While it may seem intuitive to choose the class with the highest score, the argmax operator does not necessarily yield the desired results. While for a calibrated model (see Def. 2.59) the argmax operator is the optimal decision rule for AC [33, 113, 157], a prevalence shift can lead to severe miscalibration of the predicted class scores. In such a case, or when using a different metric, using argmax may lead to suboptimal decisions.

Performance measure '*Test set results mirror real-world application performance*'
Since data collection is often prone to sample selection bias [54], the prevalences of the test set may not match the prevalences observed in deployment. Prevalence-dependent metrics (see Tab. 4.1) are inherently susceptible to prevalence shifts [317] and thus not well suited for comparative performance analysis across datasets.

The subliminal presence of these misconceptions will be supported by our literature analysis in Sec. 6.2.1. Thus, in addition to our core research question **(RQ3)**, we are interested in the consequences of failing to detect or mishandling prevalence shifts.

6.1.3 Prevalence-aware deployment workflow

We propose the following five-step workflow to detect and properly handle prevalence shifts. First, (I) the deployment prevalences $\mathcal{P}_{\text{dep}}$ are estimated, in a process called *quantification*. Next, the estimated prevalences $\hat{\mathcal{P}}_{\text{dep}}$ are used to (II) re-calibrate the model and (III) configure the performance measure (in our case, EC). Then, (IV) the decision

rule for making categorical decisions is adjusted accordingly. Finally, (V) an (external) validation can be performed on the deployment data.

Crucially, the whole process requires only three ingredients for the Lifelong Learning system: (i) the trained model φ, (ii) the small labeled calibration task $\mathcal{T}_{\text{cal}}$, and (ii) the *images of* the deployment task $\mathcal{T}_{\text{dep}}$. No access to the training task $\mathcal{T}_{\text{train}}$ or the deployment task labels $Y_{\mathcal{T}_{\text{dep}}}$ is required. Note also that in order to ensure that our workflow is applicable, the preconditions of a prevalence shift must be met (see Def. 6.4), i. e., we rely on an anticausal task with the assumption that $p_{\text{dev}}(X|Y) = p_{\text{dep}}(X|Y)$. There are two more mild assumptions we have to make [230, 443]:

(i) The support of $p_{\text{dep}}(Y)$ is a subset of the support of $p_{\text{dev}}(Y)$, i. e., $\forall k \leq C : \mathcal{P}_{\text{dep}}(k) > 0 \Rightarrow \mathcal{P}_{\text{dev}}(k) > 0$.[1]

(ii) The $\mathcal{T}_{\text{dep}}$ is not 'degenerated', i. e., there is a unique $p(Y)$ that explains $p(X)$.[2]

The first assumption guarantees that no new classes will appear during deployment, which would cause problems because we would have no information from the development phase to transfer to them. The second assumption is more interesting and may seem confusing at first, but consider the following toy example.

Example 6.5. Let $\mathcal{T}$ be a task with $C = 3$ classes that only contains two types of images ($x = 0$ and $x = 1$ for simplicity) and the underlying anticausal data generation mechanism $Y \to X$ given by

- $x = 0$ for $y = 1$,
- $x = 1$ for $y = 3$,
- and for $y = 2$, x is uniformly chosen from $\{0, 1\}$.

Now suppose we observe $p(X) = (0.5, 0.5)$. There is no way to deduce whether $p(Y) = (0, 1, 0)$, or $p(Y) = (0.5, 0, 0.5)$, or $p(Y) = (0.25, 0.5, 0.25)$, etc. It is therefore impossible to deduce $\mathcal{P}_{\mathcal{T}}$ from a series of observed images $X_{\mathcal{T}}$.

Assuming that the original task is well formulated by domain experts, i. e., all relevant classes of a problem are described (first assumption) and images are indeed indicative for the corresponding label (second assumption) we describe the individual steps of our workflow in more detail.

[1] This is assumption A.2 in the work from Lipton et al. [230] and assumption A_2^{TarS} in the work from Zhang et al. [443].

[2] This is indirectly formulated as assumption A.3 in the work from Lipton et al. [230] with the help of an 'informed classifier' and more formally formulated as assumptions A_3^{TarS} and A_4^{TarS} in the work from Zhang et al. [443].

Step 1: Estimate the deployment prevalences

To detect prevalence shifts, a Lifelong Learning system must monitor the environment and quantify the class prevalences it encounters. While this estimation can theoretically be supported by medical records or epidemiological research, in this work we focus on fully data-driven approaches.

Definition 6.6. Let $\mathcal{T}$ be a task. The problem of **quantification** [141] refers to the estimation of the prevalence $\mathcal{P}_\mathcal{T}$ using only the images $X_\mathcal{T}$ of $\mathcal{T}$ but not the actual labels $Y_\mathcal{T}$.

Quantification methods can estimate the deployment prevalences based on categorical model outputs on the training data and unlabeled deployment data [118, 230, 263, 333], or from the data distributions themselves [142]. Based on our experiments (see Sec. 6.2.2), we recommend KDEyHD [263], but there are several alternatives (e. g., ACC [118, 171], HDy [142]).

Step 2: Perform prevalence-aware re-calibration

Suppose the first step detected a change in prevalences $\mathcal{P}$ between the development $\mathcal{T}_{\text{dev}}$ and deployment $\mathcal{T}_{\text{dep}}$. Ignoring potential manifestation and acquisition shifts during deployment (see Fig. 6.2), and under the mild assumptions discussed previously, there is a theoretical *optimal* solution to minimize the expected loss on $\mathcal{T}_{\text{dep}}$ [352, 443].

Definition 6.7. Let the supervised classification problem of $\mathcal{T}$ be given with loss function $\mathcal{L} : \Delta_{C-1} \times Y_\mathcal{T} \to \mathbb{R}_{\geq 0}$ (see Def. 2.75). Let $\beta : Y_\mathcal{T} \to \mathbb{R}$ be a **reweighting function**. The process of **sample reweighting** (also **importance reweighting** [352, 443]) refers to solving the same supervised classification problem with the slightly modified loss function $\mathcal{L}^\star(p, y) := \beta(y) \cdot \mathcal{L}(p, y)$.

Recall the weighted cross entropy loss $\mathcal{L}_{CE}(p, y) = -w_y \ln(p_y)$, from Def. 2.76. Following Shimodaira [352] we let $w_k = \hat{\mathcal{P}}_{\text{dep}}(k)/\mathcal{P}_{\text{dev}}(k)$, where $\hat{\mathcal{P}}$ are the estimated prevalences from the previous step. Instead of fully retraining the model φ we only learn the $(C + 1)$ parameters of the following post-hoc transformation on the small subset $\mathcal{T}_{\text{cal}}$ in order to re-calibrate φ.

Definition 6.8. Let $t \in \mathbb{R}_+$ be a positive real, and $b \in \mathbb{R}^C$ for some integer $C > 1$ then the function $f_{aff} : \mathbb{R}^C \to \Delta_{C-1}$, with

$$f_{aff}(p) := \sigma(t^{-1}p + b)$$

is called **affine scaling** [113] (also **Platt scaling** [298] or **bias-corrected temper-**

ature scaling [11]).

Step 3: Configure validation metric with deployment prevalences

Prevalence-dependent metrics, such as AC, MCC, or F1, are widely used in image analysis [240]. However, they reflect the performance of a model only with respect to the specific, currently given prevalences. For AC we showed the decomposition into prevalences and TPR in Prop. 2.20. EC as a generalization of AC had a similar decomposition as given in Prop. 2.25:

$$\mathbf{EC} = \sum_{i \leq C} \mathcal{P}(i) \sum_{j \leq C} c_{ij} R_{ij}.$$

To use EC as a robust estimator of deployment performance, we propose replacing the prevalences $\mathcal{P}_{\text{dev}}$ with those previously estimated: $\hat{\mathcal{P}}_{\text{dep}}$ [113].

Step 4: Set prevalence-aware decision rule

Different metrics require different optimal decision rules, and the use of the argmax operator is not recommended in general. When using a counting metric it is often necessary to tune the decision rule during model development (see Sec. 4.1.3), which is likely to introduce a dependence of the resulting decision rule on the prevalences $\mathcal{P}_{\text{dev}}$. Such a rule is therefore unlikely to generalize to the deployment setting. Using EC as the primary performance measure provides an elegant solution to this problem. Provided that the class scores are calibrated, one can derive the theoretically optimal decision rule for EC [33, 157], which can be applied without any tuning (see Def. 2.26):

Definition 6.9. Let $C > 2$ and $\{c_{ij}\}_{i,j \leq C}$ be some confusions costs. The **optimal decision rule for EC** [113] is defined for any $p \in \Delta_{C-1}$ by

$$\rho_{\mathbf{EC}}(p) = \operatorname{argmin}_k \sum_j c_{jk} p_j.$$

Proposition 6.10. The optimal decision rule for EC actually coincides with the argmax operator if 0-1-costs are used.

Proof.

$$\begin{aligned}\rho_{\mathbf{EC}}(p) &= \operatorname{argmin}_k \sum_j c_{jk} p_j \\ &\overset{2.23}{=} \operatorname{argmin}_k \sum_{j \neq k} p_j \\ &= \operatorname{argmin}_k (1 - p_k) \\ &= \operatorname{argmax}_k p_k\end{aligned}$$

□

Step 5: External validation

Despite the strong theoretical guarantees provided by the previous steps, additional validation on the data from the deployment environment is critical for monitoring [335]. We recommend that the predictions be validated by clinicians as part of an integrated feedback loop [116, 197]. In addition, prevalences need to be re-estimated periodically to compensate for further shifts.

6.1.4 Experimental design

The purpose of our experiments was to

(i) show the prevalence of the misconceptions we observed (see Sec. 6.1.2),

(ii) demonstrate the negative consequences of ignoring prevalence shifts, and

(iii) exhibit the capabilities of the proposed workflow to circumvent these implications (see Sec. 6.1.3).

The following paragraphs describe the data used and the experiments performed. The code for our experiments can be found online [134].

Simulating prevalence shifts

For our experiments, we will synthetically generate prevalence shifts. For this we use the mechanisms of (i) sample selection to compile tasks and (ii) sampling strategy as part of the training process. For a given task $\mathcal{T}$, we derive subsets $\mathcal{T}_{\text{train}}$, $\mathcal{T}_{\text{cal}}$, $\mathcal{T}_{\text{dev}}$, and $\mathcal{T}_{\text{dep}}$, to simulate training, validation (also to be used for re-calibration), and test data for model development respectively deployment. These subsets serve the following purposes:

$\mathcal{T}_{\text{train}}$ The *train split* is used as training data during the model development, i. e., the model updates its weights by processing these samples as part of SGD (see Def. 2.77). We also chose balanced sampling [187] to compensate for the class imbalance present in some of our $\mathcal{T}_{\text{train}}$ subsets. In this way we (i) reflect its widespread use as an

approach to boost performance of typically very important but underrepresented classes, (ii) 'align' all tasks so that we can aggregate across them for different prevalence shifts, and (iii) capture the common biased data collection scheme that has balanced classes for design simplicity.

$\mathcal{T}_{\textbf{cal}}$ The *validation and re-calibration split* is also part of the training process. Hold-out validation splits are commonly used to inform the training process and prevent overfitting. We used three techniques to do this: (i) a "ReduceLROnPlateau" learning rate scheduler [15], where the learning rates $\{\gamma_i\}_{i\in\mathbb{N}}$ are reduced in case the validation loss stagnates, (ii) an "EarlyStopping" mechanism [109] to stop the entire training process if even learning rate reduction does not help to prevent overfitting, and (iii) to keep only the model with the best validation performance from regular checkpointing. This split is also used for post-hoc re-calibration of models (see Def. 2.67).

$\mathcal{T}_{\textbf{dev}}$ The *development test set* is used to estimate the model performance on the task at the end of the development phase. To be consistent with the assumed sample selection bias during development, we design this split with balanced classes.

$\mathcal{T}_{\textbf{dep}}$ The *deployment test set* serves as a basis to simulate the samples as they will be observed during deployment. Good performance on this split is *the ultimate goal* of the Lifelong Learning system. In order to densely assess the impact of prevalence shifts, $\mathcal{T}_{\text{dep}}$ is further subject to a sub-sampling strategy to generate subtasks $\mathcal{T}_{\text{dep}}(r)$ with IR r for some $1 \leq r \leq 10$.

For each original task $\mathcal{T}$, this design allows examining prevalence shifts from the balanced $\mathcal{T}_{\text{dev}}$ to different $\mathcal{T}_{\text{dep}}(r)$. The exact procedure for generating subsets is as follows:

Assume $\mathcal{T}$ has C classes, then the first $\lfloor 0.1 \cdot |\mathcal{T}|/C \rfloor$ samples from each class are randomly drawn from $\mathcal{T}$ to make up $\mathcal{T}_{\text{dev}}$. Note that $\mathcal{T}_{\text{dev}}$ has balanced classes and the IR r' of the remainder $\mathcal{T} \setminus \mathcal{T}_{\text{dev}}$ does not necessarily match the IR of $\mathcal{T}$. Next, for each class, one third of the corresponding samples within $\mathcal{T} \setminus \mathcal{T}_{\text{dev}}$ are randomly selected and assigned to $\mathcal{T}_{\text{dep}}$. This ensures that the IR of $\mathcal{T}_{\text{dep}}$ is again equal to r'. The procedure is repeated with one-sixth of the remaining samples, resulting in $\mathcal{T}_{\text{cal}}$, while the rest of samples are declared as $\mathcal{T}_{\text{train}}$. By design, the IRs of $\mathcal{T}_{\text{train}}$, $\mathcal{T}_{\text{cal}}$, and $\mathcal{T}_{\text{dep}}$ are all equal, since each class was reduced by the same factor during the process. When subsampling $\mathcal{T}_{\text{dep}}(r)$ from $\mathcal{T}_{\text{dep}}$, we distinguish two cases: If $r \geq r'$, we subsample all but the majority class, that is, we randomly select a fraction of r'/r cases from them. Otherwise, if $r < r'$, we subsample all but the minority class. More precisely, we linearly interpolate between the number of cases in the minority class m and the n total available samples of that class: $m + \frac{r-1}{r'-1} \cdot (n - m)$. The linear interpolation of all classes ensures a continuous mapping from a given IR r to the prevalences of $\mathcal{T}_{\text{dep}}(r)$ and prevents the majority class

from changing as a consequence of subsampling. These features allow us to align the tasks in our results.

This procedure makes some demands on the availability of samples in the tasks. Therefore, we filtered our full task pool (see Tab. 2.1) according to the following criteria: (i) a total number of at least 1000 samples, and (ii) the smallest class having at least 30% of the average class size (to avoid introducing a very strong artificial class imbalance in the training based on our data splitting procedure). Finally, we excluded one task that had only a few samples (nine) in the minority class of the deployment set for our maximum considered IR. This resulted in 30 tasks covering the modalities:

- colonoscopy (T01, T05)
- laparoscopy (T07-T10, T12)
- X-ray (T29, T30, T32, T33, T35, T37-T39, T42, T44-T51)
- ophthalmic microscopy (T53)
- MRI (T57, T58)
- fundus photography (T63, T64)
- capsule endoscopy (T69)

Prevalence of observed misconceptions

For our literature analysis, we used the *Google Scholar* feature to search for terms in all publications that cite a given paper. Our search was conducted on 07/27/2024. Some of our tasks share underlying dataset references, and we did not distinguish explicit use of a given task, only citation of the dataset source. Note that T44 is a derivative of T42 and thus not listed separately. We used the primary reference given in Tab. 2.1. By design, a given publication may cite more than one of the datasets – and so do we – so we report the average of term frequency per dataset.

Model training details

We trained neural networks for all 30 classification tasks. In the interest of better reproducibility and interpretability, we focused on a homogeneous workflow (e. g., by fixing hyperparameters across tasks) rather than aiming for the best possible AC for each individual task. All models used the ResNet34 [161] architecture and were implemented in *pytorch* [15]. For faster convergence, we used pretrained weights based on the ImageNet [93] provided by the *timm* library [421]. We performed an automated search for the initial learning rate [359] provided by the *lightning* framework [109] and reduced the learning rate by a factor of 0.1 when the validation loss plateaued for 5 epochs ('ReduceLROnPlateau'). As discussed earlier, we chose balanced sampling [187] for training. We add lightweight augmentations provided by the *albumentations* library [50] as follows: The deterministic series of *SmallestMaxSize (256)*, *PadIfNeeded (288x288)* and *Resize (256x256)* unified all of our image samples (see Def. 2.83). During training, we used *RandomCrop (224x224)* and *HorizontalFlip (p=0.5)* (see Def. 2.82). The batch size was 300.

Models were updated using the Adam optimizer [200] and (unweighted) Cross Entropy Loss (see Def. 2.73). Training was stopped either after a maximum of 40 epochs, or an early termination was signaled by no improvement in validation loss for 7 epochs. While monitoring validation loss at the end of each epoch, we kept and eventually used the model weights with the lowest value.

For all experiments, the model was trained on $\mathcal{T}_{\text{train}}$ while monitoring performance on $\mathcal{T}_{\text{cal}}$. Logits were then generated for $\mathcal{T}_{\text{cal}}$ (for re-calibration) as well as for $\mathcal{T}_{\text{dev}}$ and $\mathcal{T}_{\text{dep}}$ (for evaluation). To mimic a prevalence shift, we subsampled tasks $\mathcal{T}_{\text{dep}}(r)$ from the deployment test sets $\mathcal{T}_{\text{dep}}$ according to IRs $r \in [1, 10]$ with a step size of 0.5.

Workflow evaluation and implications of ignoring prevalence shifts

To avoid overoptimistic results in the later steps of our workflow, we did not use the best performing quantification method from Step 1 as the basis for our experiments in the following steps. Instead, we consistently chose the popular "Adjusted Classify and Count (ACC)" [118, 171], which estimates the deployment prevalences by performing a simple adjustment to the prevalences trivially estimated using model predictions on the deployment set (so called "Classify and Count" [118]). The following experiments were conducted:

Step 1 (quantification) To assess the ability of different quantification methods to estimate the deployment prevalences $\mathcal{P}_{dep(r)}$, we computed the L1 distance between estimated $\hat{\mathcal{P}}_{dep(r)}$ and exact prevalences $\mathcal{P}_{dep(r)}$ for varying IR r. In our experiment, we tested a broad variety of quantification methods. First, we considered methods based on aggregation of model outputs. These included "Classify and Count" (CC) [118], its simple adaptation "Adjusted Classify and Count" (ACC) [118, 171], the similar "Black Box Shift Estimation" (BBSE) [230], and "Probabilistic Adjusted Classify and Count" (PACC), a variant based on probabilistic model outputs [26]. Second, we analyzed approaches based on distribution matching techniques. Namely, the seminal "Expectation Minimanization Quantification" (EMQ) [333], a method based on minimizing the Hellinger Distance (HDy) [142], and its recent extensions using kernel density estimation with a Monte Carlo estimate (KDEyHD), the maximum likelihood framework (KDEyML), and computing a closed-form solution (KDEyCS) [263]. Implementations for most quantification methods are provided by the *QuaPy* library [262].

Step 2 (re-calibration) To assess the effect of prevalence shifts on model calibration, we measured miscalibration on the deployment test set $\mathcal{T}_{\text{dep}}(r)$ as a function of the increasing IR r. We compared our proposed post-hoc importance reweighted affine scaling transformation to:

(i) no re-calibration (i. e., applying softmax σ),

(ii) temperature scaling (see Def. 2.67) with and without importance reweighting (see Def. 6.7),

(iii) affine scaling without importance reweighting (see Def. 6.8),

(iv) and the full retraining of φ with importance reweighting.

For any applied importance reweighting we provide results for both the true deployment prevalences $\mathcal{P}_{dep(r)}$ and their estimates thereof $\hat{\mathcal{P}}_{dep(r)}$ from Step 1 (based on ACC).

Step 3 (performance measure) In addition, to assess the impact of prevalence shifts on the generalizability of validation results, we measured the absolute difference between the metric scores obtained on the development test data $\mathcal{T}_{\text{dev}}$ and those obtained on the deployment test data $\mathcal{T}_{\text{dep}}(r)$ with varying IR r. We chose popular prevalence-dependent metrics AC, MCC, and F1 for performance assessment and compared them to our proposed prevalence-adjusted EC with standard 0-1-costs (see Def. 2.23). The scores were calculated for the argmax decision rule for both non-re-calibrated and re-calibrated predicted class scores according to Step 2.

Step 4 (decision rule) We also evaluated the effect of prevalence shifts on the decision rule for all 24 binary tasks – with and without re-calibration according to Step 2. To do this, we computed the differences between the metric values on $\mathcal{T}_{\text{dep}}(r)$ corresponding to an optimal decision rule (found by sweeping) and two other decision rules: argmax and the threshold operator ρ_τ (see Def. 2.9) that was tuned on $\mathcal{T}_{\text{cal}}$ (also found by sweeping). This difference was computed as a function of the IR. We used the same performance measures as in the previous step. Note that for the standard 0-1-costs our proposed adjusted decision rule for EC coincides with the argmax operator.

6.2 Results

This section presents the results of the experiments explained in Sec. 6.1.4. We begin with our literature analysis to support the misconceptions from Sec. 6.1.2 in Sec. 6.2.1. Next, we assess the quality of quantification methods to detect and quantify prevalence shifts in Sec. 6.2.2. We will then first show the impact of prevalence shifts on the optimal decision rule in Sec. 6.2.4, and conclude with an assessment of performance measures to adequately reflect deployment performance in Sec. 6.2.5.

6.2.1 Literature analysis on misconceptions

To estimate the susceptibility of research models to the misconceptions of Sec. 6.1.2, we scanned all publications citing any of the datasets that meet the requirements for this chapter. Our analysis is based on the average frequency of search terms corresponding to the topics of calibration, metrics, and decision rules. The results are given in Tab. 6.1.

Table 6.1: Search term frequency within the literature citing any of the datasets we used for the experiments of this chapter. Originally published in Godau et al. [139].

	Search term							
Dataset (citation count)	re-calibration OR recalibration OR calibration OR calibrated OR calibrate	thresh-old	"decision threshold" OR "decision rule" OR cutoff OR "classifica-tion threshold"	Accu-racy	F1-Score OR "F1 Score"	MCC OR "Matthews Correlation Coefficient"	"Area Under Receiver Operator Characteris-tic" OR AUROC OR AUC	"Balanced Accuracy"
Kaggle Brain Tumor Cls. (2)	0.0% (0)	0.0% (0)	0.0% (0)	100.0% (2)	100.0% (2)	0.0% (0)	0.0% (0)	0.0% (0)
Hyperkvasir (329)	15.8% (52)	35.0% (115)	3.6% (12)	85.7% (282)	45.9% (151)	16.1% (53)	31.6% (104)	4.0% (13)
Brain Tumor Type Cls. (700)	5.9% (41)	34.3% (240)	2.7% (19)	95.1% (666)	40.6% (284)	6.3% (44)	22.4% (157)	1.9% (13)
CatRelComp (17)	41.2% (7)	5.9% (1)	82.4% (14)	52.9% (9)	0.0% (0)	0.0% (0)	5.9% (1)	0.0% (0)
CheXpert (2481)	15.2% (378)	42.3% (1050)	5.6% (140)	83.4% (2070)	31.6% (783)	2.9% (71)	60.1% (1490)	2.2% (55)
Zhang Chest X-Ray Images (4032)	9.6% (388)	35.2% (1420)	4.9% (196)	91.0% (3670)	33.0% (1330)	3.3% (132)	45.6% (1840)	1.5% (59)
LapGyn4 (45)	11.1% (5)	31.1% (14)	2.2% (1)	71.1% (32)	31.1% (14)	6.7% (3)	15.6% (7)	6.7% (3)
DeepDRiD (92)	18.5% (17)	30.4% (28)	4.3% (4)	82.6% (76)	37.0% (34)	1.1% (1)	63.0% (58)	4.3% (4)
Nerthus (84)	13.1% (11)	39.3% (33)	2.4% (2)	86.9% (73)	51.2% (43)	38.1% (32)	28.6% (24)	1.2% (1)
MURA (67)	13.4% (9)	49.3% (33)	6.0% (4)	94.0% (63)	26.9% (18)	1.5% (1)	43.3% (29)	32.8% (22)
Kvasir-Cap. (144)	10.4% (15)	37.5% (54)	4.2% (6)	86.8% (125)	43.8% (63)	16.0% (23)	30.6% (44)	2.1% (3)
Cholec80 (879)	12.4% (109)	36.3% (319)	2.7% (24)	81.6% (717)	24.1% (212)	1.3% (11)	7.4% (65)	2.2% (19)
Mean	13.9%	36.3%	3.7%	86.7%	43.2%	7.8%	29.5%	4.9%
Median	12.7%	35.8%	3.9%	86.3%	38.8%	3.1%	29.6%	2.1%

Logit transformation '*Temperature scaling typically corrects for model miscalibration*'

Only 13.9% of the publications explicitly mentioned the term "calibration" or a synonym thereof, demonstrating reduced awareness of model miscalibration. While a deeper analysis of the applied re-calibration techniques would be necessary to fully support our misconception, temperature scaling and its variants remain among the most cited re-calibration methods we are aware of.

Decision rule '*The* argmax *operator is the optimal decision rule*'

Common metric libraries for ML use argmax as the default decision rule (e. g., *torchmetrics* [94], *scikit-learn* [292]), while reporting on specific decision rules only happens between 3.7% (terms "decision rule", "classification threshold", "cutoff" or "decision threshold") and 36.3% (generic "threshold" search) of cases.

Performance measure '*Test set results mirror real-world application performance*'

Consistent with previous work [240], prevalence-dependent metrics remain the primary performance estimates, with AC (86.7%) being by far the most commonly used. F1 (43.2%) and MCC (7.8%) are also reported frequently. On the contrary, metrics that are prevalence-independent, such as AUROC (29.5%) or BA (4.9%), are less prevalent.

6.2.2 Quality of data-driven prevalence estimation

Our experiments show that deployment prevalences can be estimated well with some quantification methods. Following the recommendations of Sebastiani [347] we rely on the 'Absolute Error' as the primary evaluation metric (see Fig. 6.3). One of the main shortcomings of the Absolute Error is the lack of an upper bound, and in turn some samples will exert a higher influence on the results obtained [347]. Therefore, we also provide a secondary, somewhat complementary assessment measure to ensure that our conclusions are also valid from this second perspective.

Definition 6.11. Let $p, \hat{p} \in \Delta_{C-1}$, then the **Normalized Kullback-Leibler Divergence (NKLD)** [347] is given by

$$\mathbf{NKLD}(\hat{p}, p) = 2\frac{e^{\mathbf{KLD}(\hat{p},p)}}{e^{\mathbf{KLD}(\hat{p},p)} + 1} - 1.$$

The normalization ensures a value range of $[0, 1]$, with lower values being preferred. The results are shown in Fig. 6.4. In our experiments the recent Kernel Density Estimation method KDEyHD [263] gave the closest estimates for both assessment approaches. The 'outlier' in the box plot (Fig. 6.3 right) is the task T63 (DeepDRiD dr level), being the smallest of our selected tasks with 5 classes and a high intrinsic IR, it naturally presents

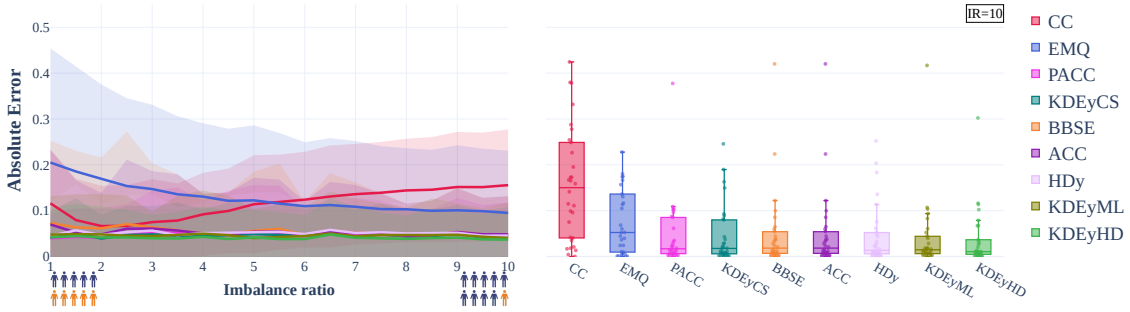


Figure 6.3: Prevalences can be accurately and robustly estimated in order to detect prevalence shifts. The 'Absolute Error' (=L1 Distance) of estimated prevalences can be held constant with an increasing prevalence shift from the development (balanced) to the deployment test set for a variety of quantification methods. Left: Mean (line) and standard deviation (shaded area) obtained from n = 30 medical classification tasks. Right: Absolute Error values for all tasks at imbalance ratio 10 (rightmost point from left figure). Each box ranges from the first quartile (Q1) to the third quartile (Q3). The second quartile (Q2) is marked by a line inside the box. The whiskers correspond to the edges of the boxes +/- 1.5 times the interquartile range (IQR: Q3-Q1). The nine different quantification methods are characterized in Sec. 6.1.4. Originally published in Godau et al. [139].

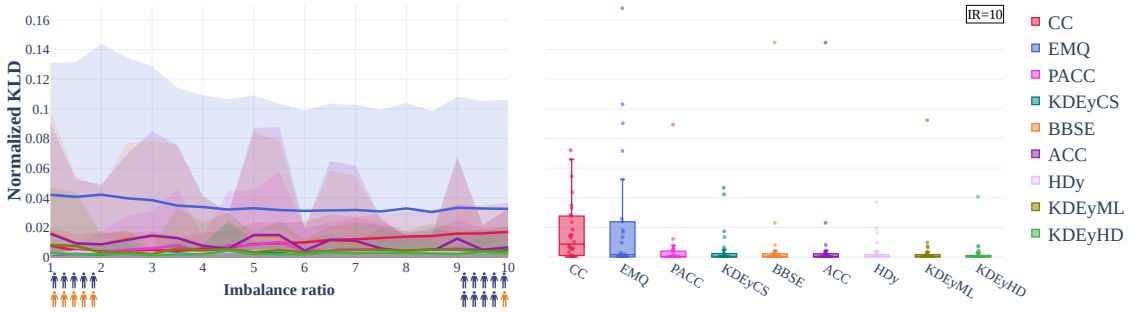


Figure 6.4: Complementary quantification assessment with Normalized Kullback-Leibler Divergence (NKLD). The NKLD of estimated prevalences can be held constant as the prevalence shifts from development (balanced) to deployment test set for a variety of quantification methods. The nine different quantification methods are characterized in Sec. 6.1.4. Originally published in Godau et al. [139].

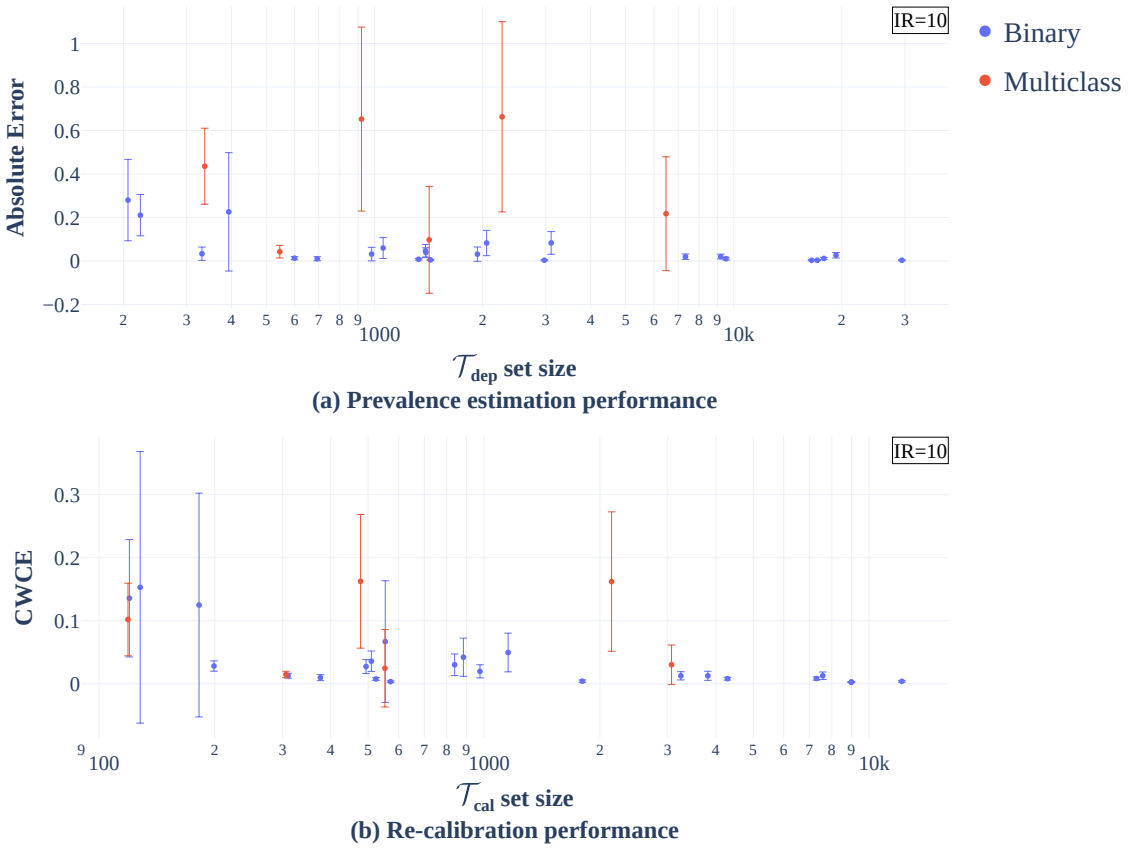


Figure 6.5: Quantification and re-calibration performance improves for larger dataset sizes. Effect of dataset size on prevalence estimation and re-calibration quality for all 30 datasets at IR=10. For each dataset, we used 15 repetitions with different random seeds for computations (10) and data partitioning (5). Each bar represents one dataset with the mean as the center and the standard deviation as range. Multiclass datasets are colored red in contrast to binary tasks (blue). (a) The Absolute Error of prevalences estimated with KDEyHD as a function of the size of the $\mathcal{T}_{dep}$ split used for prevalence estimation. No normalization was applied to correct for the different dimensionality of multiclass tasks. (b) The Class-wise Calibration Error (CWCE) for the re-calibration with importance reweighted affine scaling using the estimated prevalences relative to the size of the $\mathcal{T}_{cal}$ split used for re-calibration. Originally published in Godau et al. [139].

a challenging case for quantification. Fig. 6.5 (a) further illustrates the Absolute Error for prevalence estimation using KDEyHD at IR=10 as a function of the size of $\mathcal{T}_{dep}$.

6.2.3 Effects of prevalence shifts on model calibration

The CWCE of the raw scores produced by the systems increases as the shift between the prevalences of the development and the deployment settings increases (Fig. 6.6 top, no

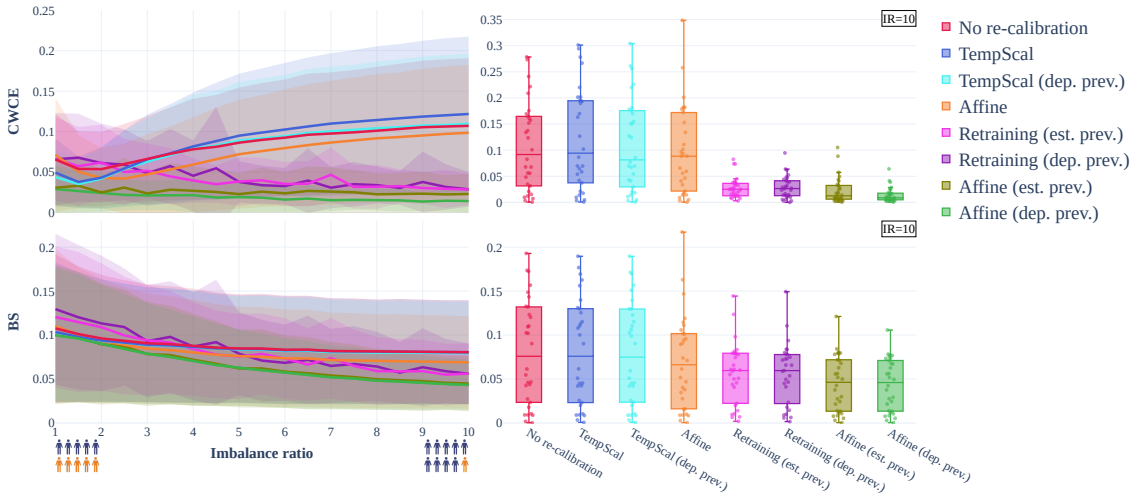


Figure 6.6: Prevalence shifts have a strong impact on calibration quality. Class-wise Calibration Error (CWCE) at the top to measure model calibration and BS at the bottom to measure the overall quality of the predicted class probabilities (lower is better for both) when shifting from a balanced deployment scenario to an Imbalance Ratio (IR) of 10. Left: Mean (line) and standard deviation (shaded area) obtained from n = 30 medical classification tasks. Right: Values for all tasks at IR=10 as box plot. Temperature scaling (blue, see Def. 2.67) as a commonly used re-calibration method does not address the miscalibration, nor does affine scaling (see Def. 6.8) without proper reweighting (orange). Retraining with importance reweighting (see Def. 6.7) largely compensates for the effect (pink & purple). The best results are achieved with the proposed approach of importance reweighted affine scaling (green). Originally published in Godau et al. [139].

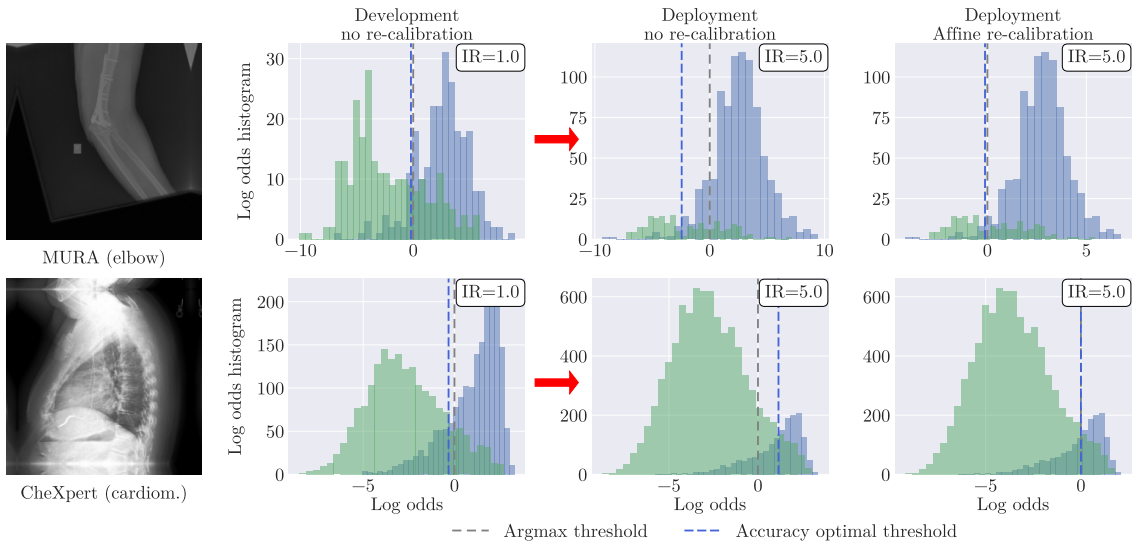


Figure 6.7: In the presence of prevalence shifts, the argmax **operator may lose its property as an optimal decision rule for AC.** Histogram of log odds for two binary sample tasks on the (balanced) development test set (second column) and on the imbalanced deployment data at Imbalance Ratio (IR) 5, without (third column) and with re-calibration (fourth column). The optimal decision threshold for AC (dashed blue line) is strongly affected by the prevalence shift and the miscalibration. Originally published in Godau et al. [139].

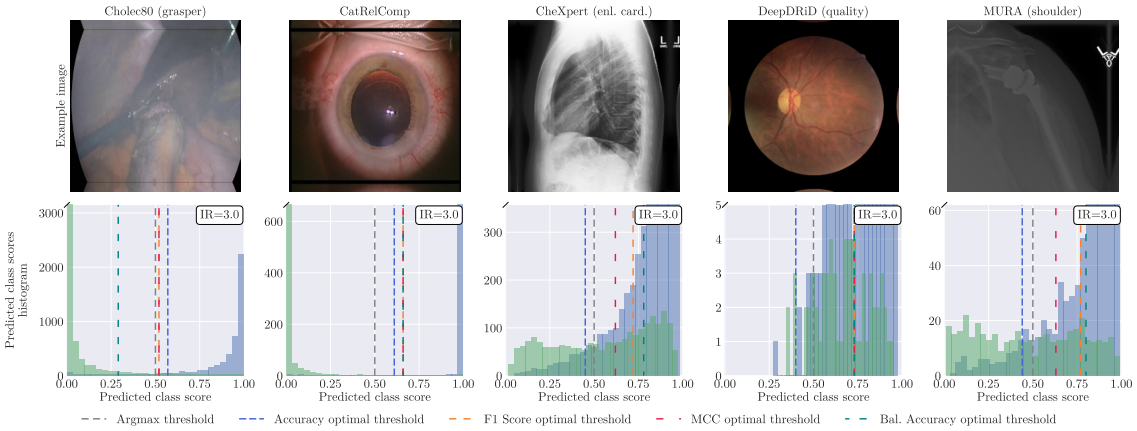


Figure 6.8: Optimal decision thresholds depend on the evaluation metrics. Histogram of predicted class scores for five binary example tasks (from left to right: T10 from Cholec80, T53 from CatRelComp, T38 from CheXpert, T64 from DeepDRiD, and T46 from MURA) on the deployment data at Imbalance Ratio (IR) 3 after re-calibration. Dotted lines mark optimal decision thresholds for several validation metrics: Accuracy (AC), F1-Score (F1), Matthews Correlation Coefficient (MCC) plus the common argmax threshold (at 0.5 for these binary tasks). Originally published in Godau et al. [139].

re-calibration curve). This miscalibration can be corrected by re-calibration. However, it is important to note that re-calibration is only successful if the transform includes a bias term (affine) and is trained with loss weights adapted to the deployment prevalences (est. prev., the prevalences estimated with the ACC method, or dep. prev., the actual deployment prevalences). Crucially, a simple temperature scaling based method is not sufficient under prevalence shifts, even if the loss weights are adjusted accordingly. Completely retraining a classifier with adjusted weights in the loss function mitigates miscalibration due to prevalence shift. Yet, miscalibration may still be present in these retrained models due to overfitting. This is the reason why the affine dep. prev. results are consistently better than the retraining results: Re-calibration with affine scaling and importance reweighting fixes the miscalibration caused by both prevalence shifts and overfitting. As an additional advantage, affine scaling requires fitting only a few parameters, which is much more computationally efficient than retraining.

The CWCE measures only the (marginal) miscalibration of the systems. However, the overall quality of the predicted class scores includes both calibration and discrimination components (see Sec. 2.6). Since for most methods not only the calibration but also the discrimination of the systems change, a better way to compare approaches is through a metric that reflects the overall performance of the scores, i. e., a proper scoring rule such as BS (see Def. 2.74). Our results show (Fig. 6.6 bottom) that the BS is minimized, for each IR, for our proposed re-calibration approaches with both methods for adjusting the prevalences (dep. or est.). Fig. 6.5 (b) additionally shows the CWCE for the (estimated) importance reweighted affine re-scaling method at IR=10 as a function of the size of $\mathcal{T}_{\text{cal}}$.

6.2.4 Effects of prevalence shifts on the decision rule

To illustrate how a prevalence shift can affect the performance of the argmax decision rule, we plot the log odds of model predictions along with the argmax decision threshold and the optimal threshold for AC for two binary tasks in Fig. 6.7. Under a prevalence shift, argmax is no longer the optimal decision rule, which can be mitigated by re-calibration. Additionally, it is important to note that the optimal decision threshold depends on the metric of interest, as illustrated in Fig. 6.8. Fig. 6.9 supports our proposal: An argmax-based decision made with re-calibrated class scores (top right) and assessed with EC (identical to the blue AC line in this case, since we use standard 0-1-costs) yields optimal results regardless of prevalence shifts. In fact, this approach substantially improves the quality of the decisions when compared to a baseline without re-calibration, as indicated by an average relative decrease in EC of 1% (IR=1), 18% (IR=4), 32% (IR=7) and 42% (IR=10), respectively, depending on IR. This is similar for EC using prevalences estimated with ACC. The results further demonstrate quantitatively what the examples in Fig. 6.8 suggested: argmax is not the best decision rule for F1 and MCC (Fig. 6.9 top). Importantly, decision rules optimized for AC, F1, or MCC on a developmental dataset do not generalize to unseen data under prevalence shifts (Fig. 6.9 bottom).

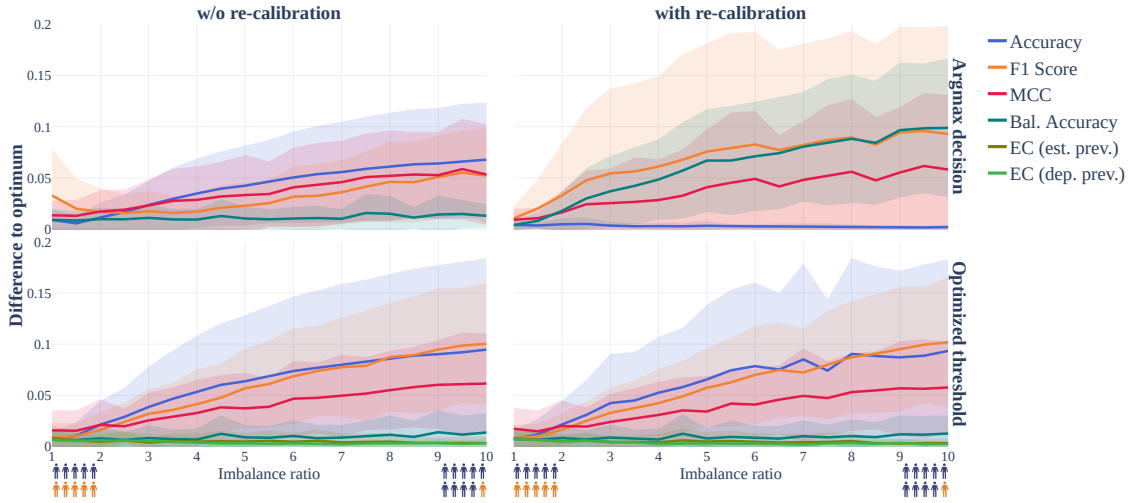


Figure 6.9: Optimal decision rules do not generalize to datasets with different prevalences. The difference between the actual and the optimal metric value on the deployment data is shown as a function of the Imbalance Ratio (IR) for non-re-calibrated (left) and re-calibrated (right) models for two decision rule strategies: argmax (top) and threshold operator ρ_τ (see Def. 2.9) optimized on the development test set (bottom). Affine scaling trained with (optimal) importance reweighting is used for all metrics except EC (est. prev.) for which the weights are adapted using the estimated prevalences $\hat{\mathcal{P}}_{dep(r)}$ for consistency (right). The optimal metric value is obtained using the optimal decision rule for the deployment data. Mean (lines) and standard deviation (transparent area) obtained from n=24 binary tasks. Originally published in Godau et al. [139].

6.2.5 Effects of prevalence shifts on the generalizability of validation results

The evaluation of our last experiment yields large deviations of metric values observed in deployment settings from those obtained on the development test data (Fig. 6.10). Prevalence-dependent metrics can vary widely, such as AC up to 0.18/0.41, F1 up to 0.46/0.18, and MCC up to 0.32/0.13 in the non-re-calibrated/re-calibrated case. BA as a prevalence-independent metric deviates only up to 0.08/0.05, as does our proposed variation of EC, which yields maximum discrepancies of 0.05/0.07 in the case of estimated prevalences and 0.05/0.02 for exact prevalences. Thus, EC allows for reliable estimation of performance irrespective of prevalence shifts, even when only estimated prevalences are available.

6.3 Discussion

Our investigation into **(RQ3)** has yielded several significant findings, particularly regarding prevalence shifts in biomedical image classification. Through extensive experimenta-

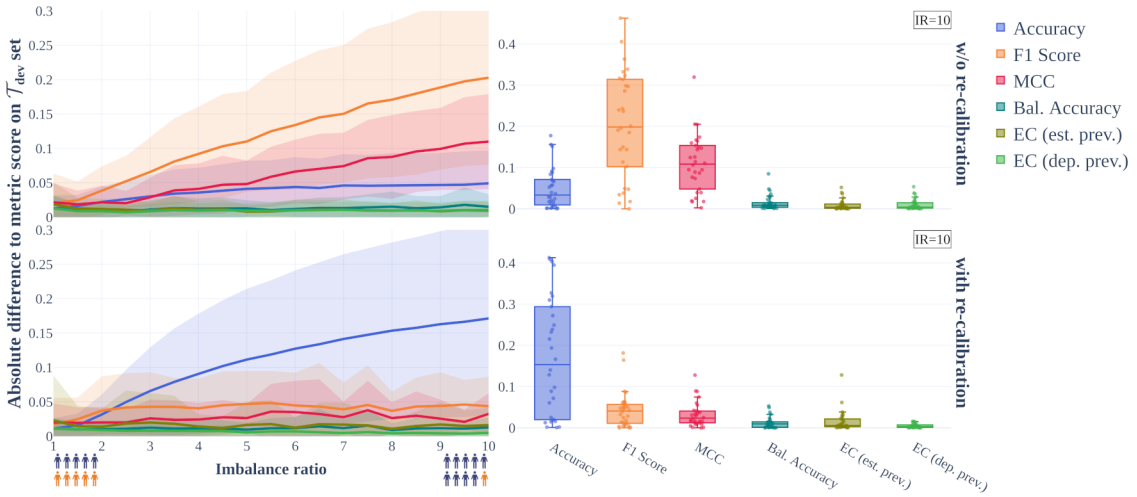


Figure 6.10: Performance estimates based on development data may become invalid under prevalence shifts. The absolute difference of the metric score computed on the deployment data to that computed on the development test set is shown as a function of the IR for non-re-calibrated (top) and re-calibrated (bottom) models. The scatter and box plots show the results for all n=30 tasks at a fixed IR of 10. Only metrics that are agnostic to prevalence – here Expected Cost (EC) configured with target prevalence and Balanced Accuracy (BA) – allow comparison of conclusions. Originally published in Godau et al. [139].

tion on 30 different image classification tasks, we demonstrated that[3]:

(i) Class prevalences can be accurately estimated from unlabeled medical imaging data using previously proposed quantification methods, with KDEyHD [263] performing the best.

(ii) Prevalence shifts can cause severe model miscalibration, rendering the standard argmax decision rule suboptimal even for straightforward metrics such as AC.

(iii) Re-calibration via importance reweighted affine scaling based on estimated deployment prevalences effectively compensates for miscalibration caused by prevalence shifts, surprisingly outperforming retraining with known prevalences.

(iv) EC, when adjusted for prevalences, provides a robust framework for comparative validation across tasks with different prevalence distributions.

[3]It should be noted in this context that our findings were confirmed by repeated experiments using multiple random seeds for dataset splitting and model training. Full results are available in the code repository [134].

Interpretation

We have shown that recent advances in quantification have indeed improved prevalence shift estimation (see Fig. 6.3) and demonstrated that such estimates are sufficient to adapt models (see Fig. 6.6, 6.9 and 6.10). KDEyHD [263] combines Kernel Density Estimation (KDE) on the probability simplex Δ_{C-1} with a Monte Carlo approach to minimize the Hellinger Distance (HD), which is computationally more demanding than previous methods[4], but still acceptable in a periodic prevalence shift detection routine of the Lifelong Learning system. Notably, we used the default values of the few hyperparameters of this approach (kernel bandwidth, number of Monte Carlo trials, and a numerical stabilizing epsilon).

In addition to quantification, also the proposed post-hoc re-calibration is also computationally extremely efficient, outperforming even the much more expensive full retraining procedure. It should also be noted that other common post-hoc adaptation techniques in model deployment, in particular ensemble methods [449], e. g., test-time augmentation [20], are not suitable for resolving prevalence shifts, as all ensemble members are likely to suffer from the same systematic prior probability bias. The task size stratification shown in Fig. 6.5 allows cautious conclusions about the number of samples required for quantification and re-calibration. For binary tasks, in both cases 1000 samples seems to be a rough order of magnitude to ensure good performance of our workflow. However, we also observe that multiclass tasks are obviously more challenging.

EC as a generalizing performance measure (see Tab. 4.2) turns out to be well suited and easily configurable along our workflow. The flexibility in assigning individual class confusion costs, the theoretically given optimal decision rule (see Def. 6.9), and the prevalence replacement strategy make it stand out not only against prevalence-dependent metrics, but also against common prevalence-independent alternatives such as BA. The magnitude of the performance degradation we observed due to prevalence shifts has been largely underexplored in the medical image analysis literature. A 0.1 performance drop in F1 resulting from an inappropriate decision rule (see Fig. 6.9) can significantly alter the comparison of algorithm results – essentially becoming a 'game changer' in competitive evaluations.

The lack of attention to calibration assessment and decision rule strategies in the literature is particularly concerning in light of our findings. Our systematic review in Sec. 6.2.1 suggests that many comparative studies in medical image analysis may be vulnerable to severe performance deviations under potential prevalence shifts that naturally occur in clinical environments.

Our results confirm theoretical predictions, but quantify their practical impact in real-world biomedical imaging tasks. The five-step workflow for detecting and adapting to prevalence shifts that we propose in Sec. 6.1.3 represents a significant advance over current practices. While domain adaptation methods have been explored previously,

[4]Anecdotally speaking, KDEyHD is about an order of magnitude slower than KDEyML and KDEyCS, which are themselves an order of magnitude slower than most previous approaches.

many may be ineffective in the presence of prevalence shifts, as noted by Arjovsky et al. [18]. Our approach differs by specifically targeting prevalence shifts and providing a mechanism for automatic adaptation without requiring new annotations.

Research context

To our knowledge, the study in this chapter represents the first comprehensive investigation of prevalence shift effects on biomedical image classification algorithms. Although EC is promoted by the *Metrics Reloaded initiative*, our literature analysis revealed that it appears in less than a dozen papers in our corpus (not yet accounting for synonyms). Our work provides compelling evidence of its advantages over established metrics in changing environments. Current practice often relies on prevalence-dependent metrics such as F1 and MCC, which we show do not provide robust performance estimates under prevalence shifts. Furthermore, the common practice of using argmax indiscriminately as a decision rule fails to account for the theoretical underpinnings of optimal decision making. Our results are consistent with the broader movement toward robust AI in healthcare, and extend previous work on acquisition shifts by Roschewitz et al. [323], which can also be used along a simultaneous prevalence shift (given a known deployment prevalence). However, the authors advise against using it for automatic re-calibration when the cause of the shift is unknown. Moreover, solving manifestation shifts requires additional assumptions about the nature of the shift [54] and cannot be solved as generally as prevalence shifts.

Limitations

Several limitations are worth discussing. First, due to computational complexity, our decision rule experiments were conducted only for binary classification tasks, although multiclass problems are common in clinical practice. Second, our experiments validated the workflow only for tasks with IRs up to 10, whereas larger imbalances may occur in practice. While our theoretical foundation should generalize to these cases, experimental validation would strengthen our claims. In addition, our analysis focused on prevalence shifts occurring in isolation. Real-world deployment often involves multiple simultaneous shifts (see Fig. 6.2), including acquisition shifts (e. g., changes in technology or imaging protocols) and manifestation shifts (e. g., changes in population demographics or clinical settings). The interplay between these different shift types remains an open research question. Future work should evaluate the performance of our proposed methods when more complex shifts are present, in comparison to methods specifically designed to combat such shifts.

Conclusion

This chapter demonstrates that models can indeed detect changing environmental conditions and adapt accordingly during deployment, specifically for prevalence shifts in

biomedical image classification tasks. Our five-step workflow enables automatic detection and adaptation to changing class distributions without requiring new annotations, a critical capability for a Lifelong Learning systems in healthcare. The proposed approach of quantifying prevalence shifts, adjusting decision rules, and performing re-calibration provides a practical framework for maintaining model performance in dynamic clinical environments. EC, with its strong theoretical foundation and flexibility [113], emerges as a recommended default metric for image classification tasks where deployment conditions may differ from development settings (keeping in mind that prevalence-dependent metrics may still be necessary to reflect the clinical interest as argued in see Sec. 4.2). These findings contribute significantly to the third metacognitive loop of our Lifelong Learning system: adaptation during the deployment phase. By autonomously addressing the challenge of shifting distributions across clinical environments, we enable AI systems to continuously evolve in changing healthcare contexts without constant human intervention.

Part IV

Perspective

7 Discussion

The three core chapters of this thesis contain individual discussions: One in Sec. 4.3 for the proposed **reward-learning loop** on the determination of appropriate performance measures based on a systematic interview, one in Sec. 5.3 for the presented **pipeline-learning loop** on the identification and reuse of relevant prior knowledge, and one in Sec. 6.3 for the suggested **environment-learning loop** on the self-adaptation of models according to detected prevalence shifts during deployment. This chapter serves as an overarching discussion of the work of this thesis in relation to our vision of a Lifelong Learning system, and aims to place the achieved results in a broader research context. Note that a conclusion and outlook are provided in Chap. 8.

Synthesis

The evolution of the research questions presented in this dissertation followed a trajectory that merits examination. The final research direction diverged severely from initial hypotheses. The preliminary conceptualization of **(RQ2)** initially centered on a system, that learns across the boundaries of datasets, incorporating some concept of *Meta Learning* (see Def. 2.89) and primarily solving the data sparsity problems in Surgical Data Science (SDS). We realized early on that the problem formulation of *Continual Learning* (see Def. 2.88) with its strong focus on the development of a *single* model did not fit this vision. It also became clear that the dominant approaches to *Meta Learning*, e. g., MAML [117], required a very homogeneous set of tasks and might not be well suited to overcome data scarcity. We were strongly inspired by the work of Achille et al. [4] (referred to as FED in Sec. 5.1, see Def. 5.11), which was accompanied by interesting theoretical considerations [5]. This led to a shift in focus to a better understanding of the relationships between tasks, how to measure them and how they relate to knowledge transferability. The term *task fingerprinting* (see Def. 5.1) for this was coined along our first conference submission [136]. Along the way, there was a recurring pattern that repeatedly caught our attention, triggered the other research questions, and contributed to the final experimental design for **(RQ2)**: The theme of 'pitfalls in validation'.

Anecdotal observations revealed systematic inconsistencies in performance measure selection for international competitions, manifesting as an implicit adherence to conventional metrics regardless of their appropriateness for specific evaluation contexts. These

observations led to the *Metrics Reloaded* initiative described in Sec. 4.1 and, in the context of this thesis, to our **(RQ1)**. Similarly, for **(RQ2)** there is a crucial statistical shift in the evaluation: the sample size 'n' moves from the number of samples in the test set of a given task $\mathcal{T}$ during the evaluation of *isolated learning* to the **number of tasks** evaluated on. This also implies a shift in the practice of splitting data: while normally each task is split into a development set, for training models and tuning hyperparameters, and a test set, for accurately estimating the empirical risk, for scenarios where 'tasks become the instances of interest' it is necessary to perform this split along the *set of tasks* (as done in Sec. 4.1). This necessary separation has not been reported for much of the literature in knowledge transfer estimation. Meanwhile, the inherently challenging research question of robustly assessing transferability estimates has also been described by Agostinelli et al. [9] and confirmed for medical imaging by Chaves et al. [58]. These methodological challenges necessitated a multidimensional expansion of the experimental design across several parameters (number of tasks, number of knowledge transfer scenarios, number of validation metrics) and the uncertainty-aware evaluation presented in Sec. 5.2. Lastly, triggered by the insights generated during *Metrics Reloaded*, especially the prevalence dependency of many metrics, we naturally came to the question of understanding 'further implications' of prevalence dependency. A significant methodological inconsistency exists wherein numerous studies employ *biased* sample selection protocols while nonetheless deriving application-specific conclusions. The observed disparity between established theoretical foundations and community implementation practices represents an unexpected finding (see Sec. 6.2.1). The quantification step (see Def. 6.6) in our proposed methodology, was actually only added later, but integrates nicely with the goal of an environmental feedback loop and adds the 'detection' of changing environments to **(RQ3)**.

Remarkably, all three of our learning loops can be used independently. They individually address complementary issues related to application alignment, data sparsity, and distribution shifts. Moreover, they are not limited to the Lifelong Learning scenario alone (see Def. 2.90). The *Metrics Reloaded* recommendations are universally designed for (multi-level) image classification tasks and have already been applied in a variety of scenarios. The *task fingerprinting* [133] and *prevalence-shifts* [134] repositories are independent plugins for the common *Medical Meta Learner* [132] framework developed and published in the context of this thesis.

Limitations

We must not ignore the limitations of our methodology. The focus on *Image-level Classification (ImLC)* was motivated in Sec. 1.2 for reasons of a well-researched problem type and the experimental efficiency. However, it is neither the only nor the most prevalent problem type in medical imaging [240]. Also, our experiments only included tasks that are based on two-dimensional RGB images – leaving out modalities such as hyperspectral imaging and three-dimensional MRI volumes. *Metrics Reloaded* covers the most common problem types that are both *discriminative* and *image-based*. The need to extend

to *generative* problem types as well as *video-based* models has already been recognized and triggered a subsequent iteration of the recommendation generation process. Our proposed task similarity measure bKLD does not rely on a specific problem type, although it has only been evaluated on one. It will be particularly challenging to find a similarity measure that performs well across problem types, as it has been shown that requirements of source task differ across problem types [224]. The hierarchical structure from pixels, to objects, to full images also complicates the handling of prevalence shifts for SemS and ObD. There, prevalence shifts can occur at multiple levels: e. g., there could be *more objects* from all classes per image *and* a shift in the *class distribution*.

We mentioned in the synthesis section, that the sample size for Lifelong Learning systems is measured by the number of tasks. The sample sizes used in this work are not yet sufficient to draw rock solid conclusions about the applicability of our approaches. The 73 international experts within *Metrics Reloaded*, including a variety of clinicians, ensure the coverage of a multitude on use cases. But with the sometimes very specific requirements for predictions on medical images, there are certainly white spots and open questions. The 43 heterogeneous validation tasks we used to evaluate bKLD are unprecedented in combination with the four transfer scenarios. Furthermore, some meta metrics used consider all possible *pairs* of tasks, which increases the respective sample size to about 43×70 or slightly below 3000 because of the partially overlapping tasks that we excluded as transfer candidates. Still, these numbers are below the expectations for truly robust scientific evidence. Fortunately, we have strong theoretical guarantees for the prevalence shift experiments, which partially compensates for our reduced number of 30 tasks. This reduction was necessary to ensure that enough samples were available to meaningfully measure performance after the splitting and subsampling strategy. However, we saw that our workflow struggled with some smaller multiclass tasks. This limitation requires further investigation.

There are also practical considerations of a Lifelong Learning system that we have not yet addressed. Repeatedly retraining models is a resource-intense paradigm that requires the necessary hardware, has ongoing costs, and increased maintenance requirements compared to a 'deploy and forget' solution. Regulatory demands for reliability and accountability have also been left untouched, such as how the general lifelong learning system would be approved by regulators as opposed to a specific medical device. The integration of such a system into the disparate infrastructures of medical centers will also require considerable effort. We have not mentioned the dilemma of 'exploration versus exploitation', which in a way deals with the willingness of the system to take risks. The parameters for such behavior, and many others necessary to configure a large-scale system to individual needs, have been wiped away in our analysis. Nor have we touched on the specifics of human-machine interaction, especially issues such as explainability, which could be crucial for the acceptance of AI in healthcare. It is also worth mentioning that a continuously evolving system offers different vectors for potential adversarial attacks compared to standalone solutions. Again, the security aspects for such a centralized and networked system would place a heavy burden on development and maintenance.

Current research directions

DL has become an incredibly fast-paced research environment in recent years, and filtering out the most important trends has therefore become an increasing challenge in itself[1]. Nevertheless, we try to provide some high-level context of advances in computer vision and potential implications for our work.

Recently, Foundation Models (FMs) – predominantly large-scale transformers – have dominated research [37]. Key strategies include using self-supervised learning on massive datasets to develop general visual understanding, integrate speech and other modalities, and reducing the fine-tuning efforts through Parameter-Efficient Fine-Tuning (PEFT) [155]. FM are also well suited to the concept of '*knowledge distillation*', where larger models act as teachers and their predictions are used to train smaller student models [168]. In particular, generative teacher models can be used to produce specific synthetic data as the underlying task for students. Knowledge distillation students tend to predict faster, are cheaper to operate, and often even more accurate than their teachers [427]. Overall, these methods emphasize the construction of strong *generalist* models and offer new approaches to knowledge transfer. In the context of our presented contributions, the systematic recommendation of performance measures is challenged by the evaluation of generative FM. Measuring their fitness for broad applicability, quantifying the effort required to optimize prompts, and the problem of hallucination are some of the issues that would need to be addressed. Our second contribution, transferability estimation via *task fingerprinting*, is challenged by the additional knowledge transfer scenarios such as PEFT and knowledge distillation. Remarkably, task fingerprinting does not have to directly match a FM with a target task, but can rely on similar tasks that have been solved via FM. The corpus of investigated models of our approach (see Sec. 5.1.3) needs to be extended to include FMs, although the computational complexity of this investigation would increase significantly due to their size. Our proposed procedure for overcoming prevalence shifts is probably the least affected. Although the degree of miscalibration in FMs seems less severe than in classical CNNs [251], the general applicability of our workflow is not compromised.

Video understanding has also emerged as a major focus of computer vision research, going beyond traditional frame-by-frame analysis to comprehend temporal dynamics and contextual relationships. Videos are particularly relevant for minimally invasive surgery, but could also prove useful for other medical imaging techniques, such as ultrasound. In addition to 'pure' video models such as three-dimensional CNNs, there are also hybrid approaches that combine, e. g., CNN backbones with temporal attention mechanisms. Introducing an additional dimension to the data naturally increases the complexity of the problem and thus the resource requirements. All three of our contributions need to

[1] According to my `scopus.com` search about 17 500 articles matching the "deep learning" query have been published in 2018, the year prior to the start of my Ph.D.. In 2024, the year prior to the completion of this dissertation, almost 122 000 articles have been published. On average, this is an increase of about 38% per year.

be carefully extended in order to capture all facets of this domain.

Conclusion 8

In this thesis, a vision for a Lifelong Learning system was outlined in Chap. 1. Subsequently, several advances for such a Lifelong Learning system were proposed and evaluated for biomedical image classification. A systematic process to align application needs with performance measures was introduced in Chap. 4. A framework for cross-institutional knowledge transfer between tasks was presented in Chap. 5. Lastly, a workflow for automatically mitigating prevalence shifts that occur during model deployment was proposed in Chap. 6. In total, tens of thousands of models have been trained and evaluated in biomedical applications.

In Sec. 8.1 of this chapter, we draw conclusions from our results with respect to our research questions from Sec. 1.2. Next, in Sec. 8.2, we summarize our scientific contributions and present the new knowledge we have gained. Finally, in Sec. 8.3, we provide an outlook for future research, touching on open challenges and opportunities.

8.1 Conclusions

Research Question 1

How can clinical objectives be systematically translated into appropriate Artificial Intelligence (AI) model validation metrics?

Our work has shown that appropriate metric selection for biomedical image classification requires knowledge from three categories: the abstract domain interest, the properties of the dataset, and details of the algorithm output. Systematically formalizing this knowledge as a *problem fingerprint*, enables automatic recommendation of appropriate performance measures, and avoids selection based on subjective preferences. An international consensus-building process involving 73 experts with 93% final agreement validated the feasibility of our recommendation workflow. The *problem fingerprint* allows domain experts to efficiently communicate relevant context to a Lifelong Learning system to derive the necessary performance measures for model development. For such systems, changes in two of the three categories (dataset properties and algorithm output) can be

updated internally, allowing adaptability to evolving clinical contexts without requiring continuous human oversight. The combination of complementary metrics ensures a holistic model evaluation and overcomes the weaknesses of individual metrics. While comprehensive, standard metric sets cannot address all specialized biomedical imaging applications, requiring optional integration of application-specific measures. For individual researchers – outside the context of a Lifelong Learning system – the structured selection process increases reproducibility and helps to avoid pitfalls without requiring in-depth metrics expertise. When multiple metrics seem appropriate, our detailed decision guides systematically resolve these tensions and provide pathways through edge cases.

Research Question 2

How to enable effective knowledge transfer across biomedical image analysis tasks?

Our proposed concept of *task fingerprinting* decouples task identifiers from task-specific experience, which enables collaborative and cross-institutional aggregation of knowledge. By enabling collaborative accumulation of insights, *task fingerprinting* fundamentally democratizes AI development, potentially reducing both development time and the environmental impact of redundant model training. Our novel proposed *task fingerprinting* method binned Kullback-Leibler Divergence (bKLD) meets the requirement of preventing the disclosure of sensitive patient data, while outperforming both manual selection and previous automated approaches. The parameterized nature of bKLD, with adjustable bin count and feature weighting, allows adaptation to specific transfer contexts, addressing the fundamental challenge that no single non-parameterized fingerprinting method adequately serves all transfer scenarios. bKLD works robustly with minimal sample sizes, making it practical for real-world medical imaging scenarios where data is often scarce. Our methodology bridges the gap between isolated learning paradigms and equips Lifelong Learning systems with a task-matching capability thatallows for targeted knowledge transfer between tasks. While the absolute performance improvement varies by transfer scenario, even modest gains represent meaningful progress toward AI systems that evolve efficiently in changing healthcare contexts without the need for human guidance. Our findings support the development of cross-institutional *Knowledge Clouds* where institutions can share task fingerprints and transfer experience without compromising patient privacy.

Research Question 3

What mechanisms enable biomedical imaging models to detect and compensate for prevalence shifts in deployment?

Our research demonstrates that biomedical imaging models can effectively detect and compensate for *prevalence shifts* in deployment environments through our novel five-step workflow. Accurate prevalence estimation during deployment is achievable with quantification methods such as KDEyHD that require only unlabeled samples. Affine re-calibration based on estimated prevalences outperforms even full model retraining with known prevalences, providing an efficient adaptation strategy for deployed models. Ignoring *prevalence shifts* can lead to severe model miscalibration that renders standard decision rules suboptimal, with performance degradations having significant clinical implications. Expected Cost (EC) emerges as a metric with a robust framework for compensating fot *prevalence shifts*, allowing for straightforward decision rule adjustments and reliable performance prediction. For binary classification tasks, approximately 1000 samples are sufficient to ensure both effective quantification and re-calibration, although multiclass problems require more samples due to their increased complexity. While *prevalence shifts* have been studied in the literature, our work directly addresses a significant gap where their profound impact is empirically underexplored and largely ignored in medical image analysis practice. Our workflow enables Lifelong Learning systems to autonomously adapt to changing deployment contexts without additional annotation effort.

8.2 Summary of contributions

This section list the contributions of new knowledge that this thesis makes.

Model validation

Our research makes several significant contributions to the field of biomedical image analysis validation: We formulated a comprehensive theoretical overview of performance measures, analyzing the relationships between common metrics, their mathematical properties, and corresponding pitfalls, thereby establishing a foundation for informed metric selection. Building on this theoretical foundation, we formalized the concept of the *problem fingerprint* – a systematic description of domain interest, dataset characteristics, and algorithm output properties that comprehensively determines appropriate validation. We introduced a novel, systematic workflow for recommending performance metrics that align with clinical goals, transforming subjective metric selection into a structured process guided by use case properties through the *problem fingerprint*. We developed detailed decision guides that address nuanced metric selection challenges, resolving tensions between competing metrics for edge cases and complex scenarios. We demonstrated the broad applicability of our framework across seven diverse biomedical imaging tasks, validating its effectiveness in translating clinical requirements into appropriate validation strategies. The results of *Metrics Reloaded* were published as two separate parts in *Nature Methods*, one describing metric pitfalls [317] and one presenting problem fingerprints

and the recommendation workflow [238].

Training in sparse data settings

Our research makes several important contributions to knowledge transfer in biomedical image analysis: We formalized the concept of *task fingerprinting* as a methodology for decoupling task identifiers from task-specific experience, providing a theoretical foundation for privacy-preserving knowledge transfer across institutions. We introduced a novel measure for quantifying task similarity, that provides parameterizable control over both feature granularity and dimensional weighting to enable scenario-specific configuration. We created a comprehensive evaluation framework for assessing task transferability estimation methods, defining a set of meta metrics that capture different aspects of transfer quality and enable robust method comparison. We conducted the largest known heterogeneous evaluation of task transferability estimation in biomedical imaging, spanning 71 tasks across 12 medical imaging modalities and four types of knowledge transfer, setting a new standard for validation scale in task transferability research. We empirically demonstrated that different knowledge transfer scenarios cannot be resolved by a single similarity measure, but require transfer scenario specificity. We have shown that our bKLD approach can work effectively with as few as 100 samples, making it viable for real-world medical imaging scenarios with limited data availability. The presented results are currently under review at *Nature Communications Medicine*, while preliminary results have been published at the *Medical Imaging Meets NeurIPS workshop* [342] and the *MICCAI conference* [136].

Prevalence shifts in algorithm deployment

Our research makes several significant contributions to understanding and addressing *prevalence shift* in biomedical image classification: We conducted the first comprehensive empirical analysis of *prevalence shift* effects on biomedical image classification performance, quantifying performance degradation across 30 different tasks and revealing substantial impacts that have been largely overlooked in the existing literature. We developed a systematic five-step workflow for detecting and compensating for *prevalence shifts* using only unlabeled deployment data, providing a practical framework for maintaining model performance in dynamic clinical environments. We provided empirical evidence that post-hoc re-calibration via importance reweighted affine scaling not only compensates for *prevalence shifts*, but surprisingly outperforms full model retraining with known prevalences, challenging common assumptions about adaptation strategies. We demonstrated that standard decision rules become suboptimal under *prevalence shifts* and quantified the substantial performance degradation this causes, providing evidence for the critical importance of decision rule adjustment in deployment contexts. We validated EC as an adjustable performance measure for managing *prevalence shifts* in biomedical image classification, demonstrated its advantages over traditional metrics, and provided

practical guidelines for its implementation. We highlighted the limitations of current practices in medical image analysis regarding *prevalence shifts*, demonstrating through a systematic literature analysis that this critical factor is routinely overlooked despite its significant impact on model performance. Our results were first published at the *MICCAI conference* [138] and as an extended version in *Medical Image Analysis* [139], winning the *Medical Image Analysis MICCAI 2023 Best Paper Award.*

8.3 Outlook

This thesis has introduced three metacognitive loops to address fundamental challenges in biomedical image classification through a Lifelong Learning framework. Each contribution advances the state of the art while opening new research directions at the intersection of AI and healthcare.

AI validation

The *Metrics Reloaded* framework provides a critical foundation for aligning AI validation with clinical goals, but significant challenges remain as the field evolves. The emerging capabilities of Foundation Models (FMs) require novel validation paradigms that go beyond traditional performance measures to capture transfer efficiency, determine generalization limits, and evaluate open-ended questions. Future research must address temporal validation challenges as AI systems continuously adapt in deployment, requiring dynamic evaluation that can distinguish beneficial adaptation from performance drift. As regulatory frameworks for AI medical devices mature, our *problem fingerprinting* methodology offers a standardized approach that could inform compliance requirements. Perhaps most importantly, the field must evolve from optimizing technical metrics to validating tangible clinical impact, requiring evaluation schemes that measure success through real-world outcomes. The true promise of our validation framework lies in its potential to ensure that biomedical imaging AI demonstrably improves healthcare rather than merely advancing technical benchmarks.

Knowledge transfer

The emergence of FMs has demonstrated that building large databases creates powerful advantages, but *task fingerprinting* offers a democratizing alternative by enabling collaborative knowledge accumulation and overcoming isolated data silos. Three promising research directions emerge from our work: First, combining *task fingerprinting* with Parameter-Efficient Fine-Tuning (PEFT) methods for FM could dramatically reduce computational requirements while leveraging the generalizability of FM. Second, assessing the configurability of bKLD for problem types other than Image-level Classification (ImLC). Third, extending *task fingerprints* to encompass multimodal data and exploring cross-domain knowledge transfer capabilities. Future research must also address the tension

between knowledge sharing and competitive advantage by developing incentive mechanisms that encourage contributions to shared *Knowledge Clouds*. Our fingerprinting methodology has the potential to transform thousands of isolated medical AI experiments into a coherent, accessible *Knowledge Cloud* that accelerates progress across healthcare domains worldwide.

Distribution shifts

As soon as healthcare AI deployment scales globally, the challenge of *distribution shifts* will intensify across diverse clinical environments and populations. While our workflow effectively addresses *prevalence shifts*, future research must tackle the complex interplay of multiple simultaneous shifts that characterize real-world deployment scenarios. Our approach could be extended to a continuous monitoring system for Lifelong Learning systems, that detects and distinguishes variants of *distribution shifts*. The integration of causal reasoning frameworks represents a promising direction for disentangling complex *distribution shifts* beyond prevalence alone. In addition, quantifying uncertainty in both prevalence estimation and adaptation decisions will be critical for maintaining clinical trust in autonomously adapting systems. Finally, our methodology for adapting EC provides a foundation for robust model validation under presumed *prevalence shifts*, enabling researchers to accurately report model performance in biased research settings, making potential deployment expectations more realistic.

Lifelong Learning systems

The true potential of the three metacognitive loops described in this thesis lies not only in their individual contributions, but in their integration within comprehensive Lifelong Learning systems. Future research should explore how these processes might interact with each other, with insights from one loop informing adjustments in others. For example, detected *distribution shifts* could trigger targeted knowledge transfer from similar historical contexts, while validation metrics could dynamically adjust to reflect the uncertainty introduced by environmental changes. As AI continues its evolutionary trajectory from isolated algorithms to continuous learning systems, the self-referential loops that have driven previous breakthroughs – from universal computation to FMs – are likely to become increasingly sophisticated. The methods developed in this thesis represent meaningful steps toward truly autonomous learning systems that can grow with the healthcare environments they serve: Ensuring alignment with evolving clinical goals, leveraging accumulated experience, and adapting to changing environments. By addressing open issues in medical imaging throughout the entire AI lifecycle, this work contributes to a future where AI systems no longer represent static capabilities, but have the ability to continuously interact, learn, and adapt in response to the dynamic challenges of real-world clinical practice.

Closing

This brings us to the end of my thesis. To wrap it up, I would like to end it as we began it, following the *self-referencing loops* of Douglas Hofstadter through Fig. 8.1, a comic by Randall Munroe[1]:

Figure 8.1: xkcd comic about self-referential statements. The seemingly unfinished sentence of the second panel is completed by the acronym made up of all the initial letters of the sentence: 'IS META'. Published by Randall Munroe in June 2011 at `xkcd.com/917` under the CC-BY-NC license version 2.5.

[1]Noteworthy I have been proudly presenting xkcd comics to the *Intelligent Systems for Surgery and Endoscopy* (ISSE) team every week for more than three years now, so this ending did not come out of nowhere.

Disclosure of personal contributions

This appendix details my individual involvements for the research projects of this thesis. By transparently providing the respective context, I would like to acknowledge the contributions and roles of my colleagues in these projects. Given the opportunity I would also like to list some scientific contributions and achievements outside the projects mentioned in this thesis.

Chapter 4

The *Metrics Reloaded* initiative, which spanned 2.5 years of collaborative work through workshops, surveys, and expert discussions, culminated in two major publications: one addressing metric pitfalls [317], and another detailing problem fingerprints and recommendations [238]. Throughout this process, we shared interim results at various venues, including the Medical Imaging Meets International Conference on Neural Information Processing Systems (NeurIPS) workshop [315] and MIDL conference [316], while maintaining a regularly updated preprint on arXiv [237].

Metric Reloaded was initiated by Lena Maier-Hein and Annika Reinke, in the scope of Helmholtz Imaging, the Medical Open Network for Artificial Intelligence (MONAI) Working Group for Evaluation, Reproducibility and Benchmarks, and the Medical Image Computing and Computer Assisted Interventions (MICCAI) Special Interest Group for Challenges (formerly MICCAI board working group). Joined by Paul F. Jäger they later formed the Delphi core team. Together with Minu D. Tizabi, Evangelia Christodoulou, A. Emre Kavur, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, Tim Rädsch and Annette Kopp-Schneider, I was part of the 'extended Delphi core team' (see Sec. 4.1). Together with Michael A. Riegler, I led the ImLC expert group, coordinating, summarizing, and reporting on several meetings with the members M. Jorge Cardoso, Veronika Cheplygina, Michael M. Hoffman, Geert Litjens, Erik Meijering, Henning Müller, and Gaël Varoquaux. My main contributions include the systematic analysis of metric relationships (see Sec. 4.1.2), the identification of problem fingerprints for metric suitability (see Sec. 4.1.3), the compilation of the metric selection workflow (see Sec. 4.2.1), and writing ImLC related parts of the manuscript [238]. Although the scientific outreach was

predominantly performed by the core team, I had the honor to present the results to an external review board during the evaluation of the research focus E at the German Cancer Research Center in 2023. The following year, I presented and discussed the initiative at a workshop on 'Benchmarking and Evaluation' at the Chan Zuckerberg Initiative in San Francisco, that primarily focused on biological applications, opposed to the medical focus of this thesis. A white paper on the outcomes of this workshop is currently in preparation.

Chapter 5

The work on quantifiable task relations was largely my sole research project under the supervision of Lena Maier-Hein. Notable contributions were made by Akriti Srivastava and Tim Adler. I presented a first sketch of *task fingerprinting* as an abstract at the Medical Imaging Meets NeurIPS workshop in 2020 [342]. The following year, an extended paper was submitted to MICCAI, received a straight acceptance, and I was invited to give an oral presentation [136]. The year after, the same results were also presented at the German Conference on Medical Image Computing (BVM) [135]. The greatly extended results discussed in this thesis are currently under revision at nature *Communications Medicine*, but have already been published as a preprint [137]. I have open-sourced the general framework for knowledge transfer under the name *Medical Meta Learner*[1] on GitHub [132]. It is designed to be highly extensible in terms of problem types (e. g., tested for Semantic Segmentation (SemS) and regression tasks), novel learning routines (so-called 'schedulers', e. g., first order Meta Learning), modalities (e. g., bounding boxes and video clips), model architectures, and more. A set of initial plugins allows easy use of all the datasets described in Sec. 2.2, automated creation of task derivatives, parallelized hyperparameter optimization across cluster infrastructure, and more. The source code for replicating the experiments conducted in Sec. 5.2 has been released separately leveraging this plugin mechanism [133].

Chapter 6

The prevalence shift analysis was a close collaboration under the joint executive leadership of my colleague Piotr Kalinowski and myself. The core idea was developed in the context of *Metrics Reloaded* by Lena Maier-Hein, Paul Jäger and Luciana Ferrer, who initiated our research and supervised the project. We were joined by Evangelia Christodoulou, Annika Reinke, and Minu Tizabi in advisory roles. We submitted our results to MICCAI, received a straight acceptance, were invited to give an oral presentation, and were shortlisted for

[1] Abbreviated **MML**, which turned out to be prone for confusions as those two letters are frequently combined as abbreviations in computer science: Large Language Model (LLM), Linear Mixed Model (LMM), Masked Language Model (MLM), Multiplicative Linear Logic (MLL), Minimum Message Length (MML), Local Maximum Likelihood (LML), Metaverse Markup Language (MML), Mathematical Markup Language (MathML), Mathematics for Machine Learning (MML) or the Multimedia Laboratory (MMLab).

the best paper award [138]. After the conference we were invited to submit an extension of our work to a special issue on MICCAI 2023 of the journal *Medical Image Analysis*. After extending our experiments to include the quantification of prevalence shifts, this paper won the *Medical Image Analysis MICCAI 2023 Best Paper Award* [139]. The source code to replicate both, the original and the extended version, has been published on GitHub [133].

Bibliography

[1] Feb. 2024. URL: https://metrics-reloaded.dkfz.de/ (cit. on pp. 99, 119, 120, 152).

[2] Ernst Abbe. "Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung". In: *Archiv für mikroskopische Anatomie* 9.1 (1873), pp. 413–468 (cit. on p. 17).

[3] Nermeen Abou Baker and Uwe Handmann. "One size does not fit all in evaluating model selection scores for image classification". In: *Scientific Reports* 14.1 (2024), pp. 1–26 (cit. on pp. 172, 175, 191, 192).

[4] Alessandro Achille et al. "Task2vec: Task embedding for meta-learning". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6430–6439 (cit. on pp. 87, 165, 166, 186, 187, 223).

[5] Alessandro Achille et al. "The information complexity of learning tasks, their structure and their distance". In: *Information and Inference: A Journal of the IMA* 10.1 (2021), pp. 51–72 (cit. on pp. 87, 223).

[6] Adewole S Adamson and Avery Smith. "Machine learning and health care disparities in dermatology". In: *JAMA dermatology* 154.11 (2018), pp. 1247–1248 (cit. on p. 153).

[7] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. "On the surprising behavior of distance metrics in high dimensional space". In: *Database theory—ICDT 2001: 8th international conference London, UK, January 4–6, 2001 proceedings 8*. Springer. 2001, pp. 420–434 (cit. on p. 161).

[8] Ravi Aggarwal et al. "Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis". In: *NPJ digital medicine* 4.1 (2021), p. 65 (cit. on p. 85).

[9] Andrea Agostinelli et al. "How stable are transferability metrics evaluations?" In: *European Conference on Computer Vision*. Springer. 2022, pp. 303–321 (cit. on pp. 172, 175, 181, 189, 191, 192, 224).

[10] Nur Ahmed and Muntasir Wahed. "The De-democratization of AI: Deep learning and the compute divide in artificial intelligence research". In: *arXiv preprint arXiv:2010.15581* (2020) (cit. on p. 191).

[11] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. "Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 222–232 (cit. on pp. 89, 202).

[12] David Alvarez-Melis and Nicolo Fusi. "Geometric dataset distances via optimal transport". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 21428–21439 (cit. on pp. 87, 165).

[13] *Analysis of Images to Detect Abnormalities in Endoscopy (AIDA-E) challenge*. `https://aidasub-clebarrett.grand-challenge.org/home/` (cit. on pp. 25, 31).

[14] D Anderson and K Burnham. "Model selection and multi-model inference". In: *Second. NY: Springer-Verlag* 63.2020 (2004), p. 10 (cit. on p. 70).

[15] Jason Ansel et al. "PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation". In: *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, Apr. 2024. DOI: 10.1145/3620665.3640366. URL: `https://pytorch.org/assets/pytorch2-2.pdf` (cit. on pp. 73, 204, 205).

[16] Lynton Ardizzone et al. "Analyzing inverse problems with invertible neural networks". In: *arXiv preprint arXiv:1808.04730* (2018) (cit. on p. 161).

[17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR. 2017, pp. 214–223 (cit. on p. 163).

[18] Martin Arjovsky et al. "Invariant risk minimization". In: *arXiv preprint arXiv:1907.02893* (2019) (cit. on pp. 88, 218).

[19] Deborah Ashby and Adrian FM Smith. "Evidence-based medicine as Bayesian decision-making". In: *Statistics in medicine* 19.23 (2000), pp. 3291–3305 (cit. on p. 127).

[20] Murat Seckin Ayhan and Philipp Berens. "Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks". In: *Medical Imaging with Deep Learning*. 2018 (cit. on p. 217).

[21] Bobby Azad et al. "Foundational models in medical imaging: A comprehensive survey and future vision". In: *arXiv preprint arXiv:2310.18689* (2023) (cit. on p. 192).

[22] MA Badgeley et al. *Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med. 2019; 2: 31*. 2019 (cit. on p. 120).

[23] Monya Baker. "1,500 scientists lift the lid on reproducibility". In: *Nature* 533.7604 (2016), pp. 452–454 (cit. on p. 18).

[24] Christian F Baumgartner et al. “SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound”. In: *IEEE transactions on medical imaging* 36.11 (2017), pp. 2204–2215 (cit. on p. 145).

[25] Emma Beede et al. “A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy”. In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–12 (cit. on p. 18).

[26] Antonio Bella et al. “Quantification via probability estimators”. In: *2010 IEEE International Conference on Data Mining*. IEEE. 2010, pp. 737–742 (cit. on pp. 89, 206).

[27] Shai Ben-David et al. “Analysis of representations for domain adaptation”. In: *Advances in neural information processing systems* 19 (2006) (cit. on p. 87).

[28] James Bergstra and Yoshua Bengio. “Random search for hyper-parameter optimization.” In: *Journal of machine learning research* 13.2 (2012) (cit. on p. 76).

[29] Joseph Berkson. “Application of the logistic function to bio-assay”. In: *Journal of the American statistical association* 39.227 (1944), pp. 357–365 (cit. on p. 36).

[30] Olivier Bernard et al. “Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?” In: *IEEE transactions on medical imaging* 37.11 (2018), pp. 2514–2525 (cit. on p. 145).

[31] Daniel Berrar. “Cross-Validation”. In: *Encyclopedia of Bioinformatics and Computational Biology* 1.April (2019), pp. 542–545 (cit. on p. 76).

[32] Bishwaranjan Bhattacharjee et al. “P2L: Predicting transfer learning for images and semantic relations”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, pp. 760–761 (cit. on pp. 167, 181, 186, 187, 191, 192).

[33] Christopher M Bishop. “Pattern recognition and machine learning”. In: *Springer google schola* 2 (2006), pp. 1122–1128 (cit. on pp. 36, 127, 166, 199, 202).

[34] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. “YOLOv4: Optimal Speed and Accuracy of Object Detection”. In: *ArXiv* abs/2004.10934 (2020) (cit. on p. 180).

[35] Vladimir I Bogachev and Aleksandr V Kolesnikov. “The Monge-Kantorovich problem: achievements, connections, and perspectives”. In: *Russian Mathematical Surveys* 67.5 (2012), p. 785 (cit. on p. 162).

[36] Jakesh Bohaju. *Brain Tumor*. 2020. DOI: 10.34740/KAGGLE/DSV/1370629. URL: https://doi.org/10.34740/KAGGLE/DSV/1370629 (cit. on pp. 23, 31).

[37] Rishi Bommasani et al. “On the opportunities and risks of foundation models”. In: *arXiv preprint arXiv:2108.07258* (2021) (cit. on pp. 5, 87, 192, 226).

[38] Hanna Borgli et al. "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy". In: *Scientific data* 7.1 (2020), p. 283 (cit. on pp. 25, 28, 77).

[39] Patrick M Bossuyt et al. "STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies". In: *Radiology* 277.3 (2015), pp. 826–832 (cit. on pp. 85, 152).

[40] Patrick M Bossuyt et al. "Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative". In: *Annals of internal medicine* 138.1 (2003), pp. 40–44 (cit. on p. 152).

[41] Y-Lan Boureau, Jean Ponce, and Yann LeCun. "A theoretical analysis of feature pooling in visual recognition". In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010, pp. 111–118 (cit. on p. 74).

[42] Stevo Bozinovski. "Reminder of the first paper on transfer learning in neural networks, 1976". In: *Informatica* 44.3 (2020) (cit. on p. 87).

[43] Glenn W Brier. "Verification of forecasts expressed in terms of probability". In: *Monthly weather review* 78.1 (1950), pp. 1–3 (cit. on p. 69).

[44] Jochen Bröcker. "Reliability, sufficiency, and the decomposition of proper scores". In: *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography* 135.643 (2009), pp. 1512–1519 (cit. on pp. 67, 68).

[45] Kay Henning Brodersen et al. "The balanced accuracy and its posterior distribution". In: *2010 20th international conference on pattern recognition*. IEEE. 2010, pp. 3121–3124 (cit. on p. 43).

[46] Allan G Bromley. "Charles Babbage's analytical engine, 1838". In: *Annals of the History of Computing* 4.3 (1982), pp. 196–217 (cit. on p. 4).

[47] Bernice B Brown. "Delphi process: a methodology used for the elicitation of opinions of experts". In: (1968) (cit. on p. 96).

[48] Tom B Brown. "Language models are few-shot learners". In: *arXiv preprint arXiv:2005.14165* (2020) (cit. on pp. 5, 8).

[49] Niko Brummer. "Measuring, refining and calibrating speaker and language information extracted from speech". PhD thesis. Stellenbosch: University of Stellenbosch, 2010 (cit. on p. 127).

[50] Alexander Buslaev et al. "Albumentations: fast and flexible image augmentations". In: *Information* 11.2 (2020), p. 125 (cit. on pp. 181, 205).

[51] Juan C Caicedo et al. "Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl". In: *Nature methods* 16.12 (2019), pp. 1247–1253 (cit. on p. 111).

[52] Luigi Carratino et al. “On mixup regularization”. In: *Journal of Machine Learning Research* 23.325 (2022), pp. 1–31 (cit. on p. 75).

[53] R Caruana et al. “Learning to learn: knowledge consolidation and transfer in inductive systems”. In: *Workshop held at NIPS-95, Vail, CO, see http://www. cs. cmu. edu/afs/user/caruana/pub/transfer. html.* 1995 (cit. on p. 6).

[54] Daniel C Castro, Ian Walker, and Ben Glocker. “Causality matters in medical imaging”. In: *Nature Communications* 11.1 (2020), p. 3673 (cit. on pp. 11, 20, 21, 88, 197–199, 218).

[55] *Cataract dataset*. `https://www.kaggle.com/datasets/jr2ngb/cataractdataset` (cit. on pp. 25, 31).

[56] Augustin Cauchy et al. “Méthode générale pour la résolution des systemes d’équations simultanées”. In: *Comp. Rend. Sci. Paris* 25.1847 (1847), pp. 536–538 (cit. on p. 71).

[57] Danton S Char, Nigam H Shah, and David Magnus. “Implementing machine learning in health care—addressing ethical challenges”. In: *New England Journal of Medicine* 378.11 (2018), pp. 981–983 (cit. on p. 154).

[58] Levy Chaves et al. “The performance of transferability metrics does not translate to medical tasks”. In: *MICCAI Workshop on Domain Adaptation and Representation Transfer*. Springer. 2023, pp. 105–114 (cit. on pp. 175, 179, 191, 192, 224).

[59] Lin Chen et al. “On Bias-Variance Alignment in Deep Models”. In: *The Twelfth International Conference on Learning Representations*. 2023 (cit. on pp. 62, 63).

[60] Zhiyuan Chen and Bing Liu. *Lifelong Machine Learning*. en. Synthesis Lectures on Artificial Intelligence and Machine Learning. Cham: Springer International Publishing, 2018. DOI: 10.1007/978-3-031-01581-6. URL: `https://link.springer.com/10.1007/978-3-031-01581-6` (visited on 10/30/2024) (cit. on pp. 6, 8, 81).

[61] Jun Cheng et al. “Enhanced performance of brain tumor classification via tumor region augmentation and partition”. In: *PloS one* 10.10 (2015), e0140381 (cit. on pp. 23, 31).

[62] V Cheplygina et al. “Exploring the similarity of medical imaging classification problems”. In: *2nd International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis,(LABELS 2017), 10-14 September 2017, Quebec City, Canada*. Springer. 2017, pp. 59–66 (cit. on pp. 87, 192).

[63] Davide Chicco and Giuseppe Jurman. “A statistical comparison between Matthews correlation coefficient (MCC), prevalence threshold, and Fowlkes–Mallows index”. In: *Journal of Biomedical Informatics* 144 (2023), p. 104426 (cit. on p. 84).

[64] Davide Chicco and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC genomics* 21 (2020), pp. 1–13 (cit. on pp. 84, 132).

[65] Davide Chicco and Giuseppe Jurman. "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification". In: *BioData Mining* 16.1 (2023), p. 4 (cit. on p. 84).

[66] Davide Chicco, Valery Starovoitov, and Giuseppe Jurman. "The benefits of the Matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment". In: *Ieee Access* 9 (2021), pp. 47112–47124 (cit. on p. 84).

[67] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation". In: *BioData mining* 14 (2021), pp. 1–22 (cit. on p. 84).

[68] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. "The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment". In: *Ieee Access* 9 (2021), pp. 78368–78381 (cit. on pp. 84, 127).

[69] Peter Christen, David J Hand, and Nishadi Kirielle. "A review of the F-measure: its history, properties, criticism, and alternatives". In: *ACM Computing Surveys* 56.3 (2023), pp. 1–24 (cit. on p. 84).

[70] Margaret Chustecki et al. "Benefits and risks of AI in health care: Narrative review". In: *Interactive Journal of Medical Research* 13.1 (2024), e53616 (cit. on p. 19).

[71] Neil T Clancy et al. "Surgical spectral imaging". In: *Medical image analysis* 63 (2020), p. 101699 (cit. on p. 33).

[72] Noel Codella et al. "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)". In: *arXiv preprint arXiv:1902.03368* (2019) (cit. on p. 145).

[73] Gregory Cohen et al. "EMNIST: Extending MNIST to handwritten letters". In: *2017 international joint conference on neural networks (IJCNN)*. IEEE. 2017, pp. 2921–2926 (cit. on pp. 27, 29).

[74] Gregory Cohen et al. "EMNIST: Extending MNIST to handwritten letters". In: *2017 International Joint Conference on Neural Networks (IJCNN)* (2017), pp. 2921–2926 (cit. on p. 75).

[75] Jacob Cohen. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46 (cit. on p. 46).

[76] Jacob Cohen. “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.” In: *Psychological bulletin* 70.4 (1968), p. 213 (cit. on p. 46).

[77] International Skin Imaging Collaboration. *SIIM-ISIC 2020 Challenge Dataset*. 2020. URL: https://doi.org/10.34970/2020-ds01 (cit. on p. 32).

[78] Olivier Commowick et al. “Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure”. In: *Scientific reports* 8.1 (2018), p. 13650 (cit. on p. 145).

[79] Marc-Alexandre Côté et al. “Tractometer: towards validation of tractography pipelines”. eng. In: *Medical Image Analysis* 17.7 (Oct. 2013), pp. 844–857. ISSN: 1361-8423. DOI: 10.1016/j.media.2013.03.009 (cit. on p. 154).

[80] Michael Crawshaw. “Multi-task learning with deep neural networks: A survey”. In: *arXiv preprint arXiv:2009.09796* (2020) (cit. on p. 77).

[81] Florinel-Alin Croitoru et al. “Diffusion models in vision: A survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.9 (2023), pp. 10850–10869 (cit. on p. 18).

[82] Imre Csiszár. “I-divergence geometry of probability distributions and minimization problems”. In: *The annals of probability* (1975), pp. 146–158 (cit. on p. 167).

[83] Ekin D Cubuk et al. “Autoaugment: Learning augmentation policies from data”. In: *arXiv preprint arXiv:1805.09501* (2018) (cit. on p. 181).

[84] Yin Cui et al. “Large scale fine-grained categorization and domain-specific transfer learning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4109–4118 (cit. on p. 165).

[85] Alexander D’Amour et al. “Underspecification presents challenges for credibility in modern machine learning”. In: *Journal of Machine Learning Research* 23.226 (2022), pp. 1–61 (cit. on p. 153).

[86] Roxana Daneshjou et al. “Disparities in dermatology AI performance on a diverse, curated clinical image set”. In: *Science advances* 8.31 (2022), eabq6147 (cit. on p. 18).

[87] Jesse Davis and Mark Goadrich. “The relationship between Precision-Recall and ROC curves”. In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 233–240 (cit. on pp. 59, 138, 139).

[88] Matthias De Lange et al. “A continual learning survey: Defying forgetting in classification tasks”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.7 (2021), pp. 3366–3385 (cit. on p. 79).

[89] Daswin De Silva and Damminda Alahakoon. “An artificial intelligence life cycle: From conception to production”. In: *Patterns* 3.6 (2022) (cit. on p. 7).

[90] Morris H DeGroot and Stephen E Fienberg. "Assessing probability assessors: calibration and refinement". In: (1981) (cit. on p. 68).

[91] Morris H DeGroot and Stephen E Fienberg. "The comparison and evaluation of forecasters". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 32.1-2 (1983), pp. 12–22 (cit. on pp. 69, 125).

[92] Rosario Delgado and Xavier-Andoni Tibau. "Why Cohen's Kappa should be avoided as performance measure in classification". In: *PloS one* 14.9 (2019), e0222916 (cit. on pp. 46, 127).

[93] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on pp. 5, 9, 27, 29, 33, 42, 87, 159, 168, 178, 181, 205).

[94] Nicki Skafte Detlefsen et al. "Torchmetrics-measuring reproducibility in pytorch". In: *Journal of Open Source Software* 7.70 (2022), p. 4101 (cit. on pp. 85, 209).

[95] Walid Al-Dhabyani et al. "Dataset of breast ultrasound images". In: *Data in brief* 28 (2020), p. 104863 (cit. on pp. 22, 31).

[96] Lucas V Dias et al. "ImageDataset2Vec: An image dataset embedding for algorithm selection". In: *Expert Systems with Applications* 180 (2021), p. 115053 (cit. on pp. 87, 179, 181).

[97] Yifan Ding, Liqiang Wang, and Boqing Gong. "Analyzing Deep Neural Network's Transferability via Fréchet Distance". In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2021), pp. 3931–3940 (cit. on pp. 165, 186, 187).

[98] Yuhe Ding et al. "Which model to transfer? a survey on transferability estimation". In: *arXiv preprint arXiv:2402.15231* (2024) (cit. on pp. 87, 179, 191, 192).

[99] Jérôme Dockès, Gaël Varoquaux, and Jean-Baptiste Poline. "Preventing dataset shift from breaking machine-learning biomarkers". In: *GigaScience* 10.9 (2021), giab055 (cit. on pp. 89, 198).

[100] George R Doddington et al. "The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective". In: *Speech communication* 31.2-3 (2000), pp. 225–254 (cit. on p. 127).

[101] James M Dolezal et al. "Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology". In: *Nature communications* 13.1 (2022), p. 6572 (cit. on p. 113).

[102] DC Dowson and BV666017 Landau. "The Fréchet distance between multivariate normal distributions". In: *Journal of multivariate analysis* 12.3 (1982), pp. 450–455 (cit. on p. 164).

[103] Kshitij Dwivedi et al. “Duality diagram similarity: a generic framework for initialization selection in task transfer learning”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 497–513 (cit. on p. 172).

[104] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. “Training generative neural networks via maximum mean discrepancy optimization”. In: *arXiv preprint arXiv:1505.03906* (2015) (cit. on p. 161).

[105] Thomas Eche et al. “Toward generalizability in the deployment of artificial intelligence in radiology: role of computation stress testing to overcome underspecification”. In: *Radiology: Artificial Intelligence* 3.6 (2021), e210097 (cit. on p. 153).

[106] Matthias Eisenmann et al. “Why is the winner the best?” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 19955–19966 (cit. on p. 174).

[107] David G Ellis, Carlos M Alvarez, and Michele R Aizenberg. “Qualitative criteria for feasible cranial implant designs”. In: *Cranial Implant Design Challenge*. Springer. 2021, pp. 8–18 (cit. on p. 154).

[108] Mark Everingham et al. “The pascal visual object classes challenge: A retrospective”. In: *International journal of computer vision* 111 (2015), pp. 98–136 (cit. on p. 59).

[109] William Falcon and The PyTorch Lightning team. *PyTorch Lightning*. Version 1.4. Mar. 2019. DOI: 10.5281/zenodo.3828935. URL: https://github.com/Lightning-AI/lightning (cit. on pp. 178, 204, 205).

[110] Sergio MM de Faria et al. “Light field image dataset of skin lesions”. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2019, pp. 3905–3908 (cit. on pp. 27, 29).

[111] Tom Fawcett. “An introduction to ROC analysis”. In: *Pattern recognition letters* 27.8 (2006), pp. 861–874 (cit. on pp. 58, 117).

[112] Li Fei-Fei, Rob Fergus, and Pietro Perona. “Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories”. In: *2004 conference on computer vision and pattern recognition workshop*. IEEE. 2004, pp. 178–178 (cit. on pp. 27, 29).

[113] Luciana Ferrer. “Analysis and comparison of classification metrics”. In: *arXiv preprint arXiv:2209.05355* (2022) (cit. on pp. 43, 45, 84, 112, 127, 130, 131, 134, 144, 199, 201, 202, 219).

[114] John Field. *Lifelong learning and the new educational order*. ERIC, 2000 (cit. on p. 6).

[115] Chris Fifty et al. “Efficiently identifying task groupings for multi-task learning”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 27503–27516 (cit. on pp. 87, 182).

[116] Samuel G Finlayson et al. "The clinician and dataset shift in artificial intelligence". In: *New England Journal of Medicine* 385.3 (2021), pp. 283–286 (cit. on pp. 9, 203).

[117] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 1126–1135 (cit. on pp. 80, 223).

[118] George Forman. "Quantifying counts and costs via classification". In: *Data Mining and Knowledge Discovery* 17 (2008), pp. 164–206 (cit. on pp. 201, 206).

[119] Maurice Fréchet. "Sur la distance de deux lois de probabilité". In: *Annales de l'ISUP*. Vol. 6. 3. 1957, pp. 183–198 (cit. on p. 162).

[120] Karoline Freeman et al. "Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy". In: *bmj* 374 (2021) (cit. on p. 18).

[121] Tony Freeth et al. "Decoding the ancient Greek astronomical calculator known as the Antikythera Mechanism". In: *Nature* 444.7119 (2006), pp. 587–591 (cit. on p. 3).

[122] João Gama et al. "A survey on concept drift adaptation". In: *ACM computing surveys (CSUR)* 46.4 (2014), pp. 1–37 (cit. on p. 88).

[123] Shanghua Gao et al. "Res2Net: A New Multi-Scale Backbone Architecture". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2021), pp. 652–662 (cit. on p. 180).

[124] Saurabh Garg et al. "A unified view of label shift estimation". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3290–3300 (cit. on p. 89).

[125] Yonatan Geifman and Ran El-Yaniv. "Selective classification for deep neural networks". In: *Advances in neural information processing systems* 30 (2017) (cit. on pp. 38, 153).

[126] Robert Geirhos et al. "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673 (cit. on p. 153).

[127] Negin Ghamsarian et al. "Relevance-based compression of cataract surgery videos using convolutional neural networks". In: *Proceedings of the 28th ACM international conference on multimedia*. 2020, pp. 3577–3585 (cit. on pp. 26, 27, 31).

[128] Ioannis Giotis et al. "MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images". In: *Expert systems with applications* 42.19 (2015), pp. 6578–6585 (cit. on pp. 27, 31).

[129] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. "Deep sparse rectifier neural networks". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 315–323 (cit. on p. 73).

[130] M Maria Glymour and Sander Greenland. "Causal diagrams". In: *Modern epidemiology* 3 (2008), pp. 183–209 (cit. on p. 20).

[131] Tilmann Gneiting and Adrian E Raftery. "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American statistical Association* 102.477 (2007), pp. 359–378 (cit. on pp. 68, 69).

[132] Patrick Godau. *Medical Meta Learner*. Version 1.0.0. Dec. 2024. URL: https://github.com/IMSY-DKFZ/mml (cit. on pp. 224, 238).

[133] Patrick Godau. *task-fingerprinting*. Version 0.1.0. Oct. 2024. URL: https://github.com/IMSY-DKFZ/task-fingerprinting (cit. on pp. 192, 224, 238, 239).

[134] Patrick Godau and Piotr Kalinowski. *prevalence-shifts*. Version 0.2.0. Aug. 2024. URL: https://github.com/IMSY-DKFZ/prevalence-shifts (cit. on pp. 203, 216, 224).

[135] Patrick Godau and Lena Maier-Hein. "Task Fingerprinting for Meta Learning in Biomedical Image Analysis". In: *Bildverarbeitung für die Medizin 2022: Proceedings, German Workshop on Medical Image Computing, Heidelberg, June 26-28, 2022*. Springer. 2022, pp. 260–260 (cit. on p. 238).

[136] Patrick Godau and Lena Maier-Hein. "Task Fingerprinting for Meta Learning inBiomedical Image Analysis". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 436–446 (cit. on pp. 157, 163, 166, 167, 176, 179, 186, 187, 192, 223, 232, 238).

[137] Patrick Godau et al. "Beyond Knowledge Silos: Task Fingerprinting for Democratization of Medical Imaging AI". In: *arXiv preprint arXiv:2412.08763* (2024) (cit. on pp. 160, 168, 177, 180, 184–190, 238).

[138] Patrick Godau et al. "Deployment of image analysis algorithms under prevalence shifts". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 389–399 (cit. on pp. 195, 233, 239).

[139] Patrick Godau et al. "Navigating prevalence shifts in image analysis algorithm deployment". In: *Medical Image Analysis* (2025), p. 103504 (cit. on pp. 195, 208, 210–213, 215, 216, 233, 239).

[140] Kurt Gödel. "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I". In: *Monatshefte für mathematik und physik* 38 (1931), pp. 173–198 (cit. on p. 3).

[141] Pablo González et al. "A review on quantification learning". In: *ACM Computing Surveys (CSUR)* 50.5 (2017), pp. 1–40 (cit. on p. 201).

[142] Víctor González-Castro, Rocío Alaiz-Rodríguez, and Enrique Alegre. "Class distribution estimation based on the Hellinger distance". In: *Information Sciences* 218 (2013), pp. 146–164 (cit. on pp. 201, 206).

[143] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016 (cit. on pp. 70, 71, 73).

[144] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014) (cit. on pp. 86, 161).

[145] Mark J Gooding et al. "Comparative evaluation of autocontouring in clinical practice: A practical method using the Turing test". In: *Medical physics* 45.11 (2018), pp. 5105–5115 (cit. on p. 83).

[146] Jan Gorodkin. "Comparing two K-category assignments by a K-category correlation coefficient". In: *Computational biology and chemistry* 28.5-6 (2004), pp. 367–374 (cit. on p. 54).

[147] Margherita Grandini, Enrico Bagli, and Giorgio Visani. "Metrics for multi-class classification: an overview". In: *arXiv preprint arXiv:2008.05756* (2020) (cit. on pp. 84, 101).

[148] Arthur Gretton et al. "A kernel two-sample test". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773 (cit. on p. 161).

[149] Gregory Griffin, Alex Holub, Pietro Perona, et al. *Caltech-256 object category dataset*. Tech. rep. Technical Report 7694, California Institute of Technology Pasadena, 2007 (cit. on pp. 27, 29).

[150] Sebastian Gruber and Florian Buettner. "Better uncertainty calibration via proper scores for classification and beyond". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 8618–8632 (cit. on pp. 62, 65–67, 70, 144, 145).

[151] Chuan Guo et al. "On calibration of modern neural networks". In: *International conference on machine learning*. PMLR. 2017, pp. 1321–1330 (cit. on pp. 65, 66, 199).

[152] Kartik Gupta et al. "Calibration of neural networks using splines". In: *arXiv preprint arXiv:2006.12800* (2020) (cit. on p. 142).

[153] Niv Haim et al. "Reconstructing training data from trained neural networks". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 22911–22924 (cit. on p. 161).

[154] Dongyoon Han et al. "Rethinking Channel Dimensions for Efficient Model Design". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 732–741 (cit. on p. 180).

[155] Zeyu Han et al. "Parameter-efficient fine-tuning for large models: A comprehensive survey". In: *arXiv preprint arXiv:2403.14608* (2024) (cit. on p. 226).

[156] James A Hanley and Barbara J McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." In: *Radiology* 143.1 (1982), pp. 29–36 (cit. on p. 57).

[157] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009 (cit. on pp. 127, 199, 202).

[158] Tomoyuki Hatakeyama, Xueting Wang, and Toshihiko Yamasaki. "Transferability prediction among classification and regression tasks using optimal transport". In: *Multimedia Tools and Applications* 83.9 (2024), pp. 25105–25119 (cit. on pp. 10, 87, 172).

[159] Ryuichiro Hataya et al. "Faster autoaugment: Learning augmentation strategies using backpropagation". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer. 2020, pp. 1–16 (cit. on p. 181).

[160] Trine B Haugen et al. "Visem: A multimodal video dataset of human spermatozoa". In: *Proceedings of the 10th ACM Multimedia Systems Conference*. 2019, pp. 261–266 (cit. on p. 145).

[161] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on pp. 74, 168, 180, 205).

[162] Xin He, Kaiyong Zhao, and Xiaowen Chu. "AutoML: A survey of the state-of-the-art". In: *Knowledge-based systems* 212 (2021), p. 106622 (cit. on p. 87).

[163] Paul Heidke. "Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst". In: *Geografiska Annaler* 8.4 (1926), pp. 301–349 (cit. on p. 46).

[164] Aluru V. N. M Hemateja. *Covid19 X-Ray classification dataset on kaggle*. `https://www.kaggle.com/ahemateja19bec1025/covid-xray-dataset`. Accessed: 2022-01-13. 2021 (cit. on pp. 22, 31).

[165] Dan Hendrycks and Kevin Gimpel. "Gaussian error linear units (gelus)". In: *arXiv preprint arXiv:1606.08415* (2016) (cit. on p. 73).

[166] Martin Heusel et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017 (cit. on pp. 165, 186, 187).

[167] Steven A Hicks et al. "On evaluation metrics for medical applications of artificial intelligence". In: *Scientific reports* 12.1 (2022), p. 5979 (cit. on p. 85).

[168] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network". In: *arXiv preprint arXiv:1503.02531* (2015) (cit. on p. 226).

[169] S Hochreiter. "Long Short-term Memory". In: *Neural Computation MIT-Press* (1997) (cit. on p. 4).

[170] Douglas R Hofstadter. *Gödel, Escher, Bach: an eternal golden braid*. Basic books, 1999 (cit. on pp. 3–5).

[171] Daniel J Hopkins and Gary King. "A method of automated nonparametric content analysis for social science". In: *American Journal of Political Science* 54.1 (2010), pp. 229–247 (cit. on pp. 201, 206).

[172] Timothy Hospedales et al. "Meta-learning in neural networks: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 44.9 (2021), pp. 5149–5169 (cit. on pp. 8, 80).

[173] Mohammad Hossin and Md Nasir Sulaiman. "A review on evaluation metrics for data classification evaluations". In: *International journal of data mining & knowledge management process* 5.2 (2015), p. 1 (cit. on p. 84).

[174] Godfrey N Hounsfield. "Computed medical imaging". In: *Science* 210.4465 (1980), pp. 22–28 (cit. on p. 17).

[175] Andrew G. Howard et al. "Searching for MobileNetV3". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 1314–1324 (cit. on p. 180).

[176] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. "What makes ImageNet good for transfer learning?" In: *arXiv preprint arXiv:1608.08614* (2016) (cit. on p. 9).

[177] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019 (cit. on p. 80).

[178] Hussein Ibrahim et al. "Health data poverty: an assailable barrier to equitable digital health care". In: *The Lancet Digital Health* 3.4 (2021), e260–e265 (cit. on p. 153).

[179] John PA Ioannidis. "Why most published research findings are false". In: *PLoS medicine* 2.8 (2005), e124 (cit. on p. 18).

[180] Sergey Ioffe. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167* (2015) (cit. on p. 74).

[181] Jeremy Irvin et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597 (cit. on pp. 22, 23, 30, 77, 178).

[182] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature methods* 18.2 (2021), pp. 203–211 (cit. on pp. 75, 87).

[183] Paul Jaccard. "The distribution of the flora in the alpine zone. 1". In: *New phytologist* 11.2 (1912), pp. 37–50 (cit. on p. 53).

[184] Paul F Jaeger et al. "A Call to Reflect on Evaluation Practices for Failure Detection in Image Classification". In: *International Conference on Learning Representations* (2023) (cit. on p. 153).

[185] Stefan Jaeger et al. “Two public chest X-ray datasets for computer-aided screening of pulmonary diseases”. In: *Quantitative imaging in medicine and surgery* 4.6 (2014), p. 475 (cit. on pp. 22, 31).

[186] Pierre Jannin. “Towards responsible research in digital technology for health care”. In: *arXiv preprint arXiv:2110.09255* (2021) (cit. on p. 153).

[187] Justin M Johnson and Taghi M Khoshgoftaar. “Survey on deep learning with class imbalance”. In: *Journal of big data* 6.1 (2019), pp. 1–54 (cit. on pp. 199, 203, 205).

[188] Giuseppe Jurman, Samantha Riccadonna, and Cesare Furlanello. “A comparison of MCC and CEN error measures in multi-class prediction”. In: (2012) (cit. on p. 84).

[189] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. “Reinforcement learning: A survey”. In: *Journal of artificial intelligence research* 4 (1996), pp. 237–285 (cit. on p. 5).

[190] Feng Kang, Rong Jin, and Rahul Sukthankar. “Correlated label propagation with application to multi-label learning”. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*. Vol. 2. IEEE. 2006, pp. 1719–1726 (cit. on p. 120).

[191] Karthik, Maggie, and Sohier Dane. *APTOS 2019 Blindness Detection*. `https://kaggle.com/competitions/aptos2019-blindness-detection`. Kaggle. 2019 (cit. on pp. 25, 32).

[192] Jeremy Kawahara et al. “Seven-point checklist and skin lesion classification using multitask multimodal neural nets”. In: *IEEE journal of biomedical and health informatics* 23.2 (2018), pp. 538–546 (cit. on pp. 27, 29).

[193] Christopher J Kelly et al. “Key challenges for delivering clinical impact with artificial intelligence”. In: *BMC medicine* 17 (2019), pp. 1–9 (cit. on pp. 8, 9, 19).

[194] Maurice G Kendall. “A new measure of rank correlation”. In: *Biometrika* 30.1-2 (1938), pp. 81–93 (cit. on p. 173).

[195] Daniel S Kermany et al. “Identifying medical diagnoses and treatable diseases by image-based deep learning”. In: *cell* 172.5 (2018), pp. 1122–1131 (cit. on pp. 22, 30).

[196] Bangul Khan et al. “Drawbacks of artificial intelligence and their potential solutions in the healthcare sector”. In: *Biomedical Materials & Devices* 1.2 (2023), pp. 731–738 (cit. on pp. 19, 84).

[197] Faiza Khan Khattak et al. “MLHOps: machine learning for healthcare operations”. In: *arXiv preprint arXiv:2305.02474* (2023) (cit. on p. 203).

[198] Aditya Khosla et al. “Novel dataset for fine-grained image categorization: Stanford dogs”. In: *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*. Vol. 2. 1. 2011 (cit. on pp. 28, 29).

[199] Hee E Kim et al. "Transfer learning for medical image classification: a literature review". In: *BMC medical imaging* 22.1 (2022), p. 69 (cit. on p. 79).

[200] Diederik P Kingma. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on pp. 71, 206).

[201] James Kirkpatrick et al. "Overcoming catastrophic forgetting in neural networks". In: *Proceedings of the national academy of sciences* 114.13 (2017), pp. 3521–3526 (cit. on p. 80).

[202] Stephen Cole Kleene. "Introduction to metamathematics". In: (1952) (cit. on p. 4).

[203] Florian Kofler et al. "Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient". In: *arXiv preprint arXiv:2103.06205* (2021) (cit. on p. 83).

[204] Florian Kofler et al. "Blob loss: Instance imbalance aware loss functions for semantic segmentation". In: *International Conference on Information Processing in Medical Imaging*. Springer. 2023, pp. 755–767 (cit. on p. 145).

[205] Alexander Kolesnikov et al. "Big transfer (bit): General visual representation learning". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer. 2020, pp. 491–507 (cit. on pp. 191, 193).

[206] John Kornak and Ying Lu. "Bayesian decision analysis for choosing between diagnostic/prognostic prediction procedures". In: *Statistics and its interface* 4.1 (2011), p. 27 (cit. on p. 127).

[207] Jan Kottner et al. "Guidelines for reporting reliability and agreement studies (GRRAS) were proposed". In: *International journal of nursing studies* 48.6 (2011), pp. 661–671 (cit. on p. 120).

[208] Alex Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. 2009. URL: `https://www.cs.toronto.edu/~kriz/cifar.html` (cit. on p. 75).

[209] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images". In: (2009) (cit. on pp. 27, 29).

[210] 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 pp. 5, 18, 72).

[211] Meelis Kull et al. "Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration". In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 66).

[212] 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. 69).

[213] Ananya Kumar, Percy S Liang, and Tengyu Ma. "Verified uncertainty calibration". In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on pp. 64–66).

[214] Makerere AI Lab. *Bean disease dataset*. Jan. 2020. URL: https://github.com/AI-Lab-Makerere/ibean/ (cit. on pp. 28, 31).

[215] Alexandre Lacoste et al. "Quantifying the carbon emissions of machine learning". In: *arXiv preprint arXiv:1910.09700* (2019) (cit. on p. 155).

[216] Paul C Lauterbur. "Image formation by induced local interactions: examples employing nuclear magnetic resonance". In: *nature* 242.5394 (1973), pp. 190–191 (cit. on p. 18).

[217] Alexander Lavin et al. "Technology readiness levels for machine learning systems". In: *Nature Communications* 13.1 (2022), p. 6039 (cit. on p. 152).

[218] EPV Le et al. "Artificial intelligence in breast imaging". In: *Clinical radiology* 74.5 (2019), pp. 357–366 (cit. on p. 145).

[219] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324 (cit. on pp. 27, 29, 42).

[220] Youngwan Lee et al. "An Energy and GPU-Computation Efficient Backbone Network for Real-Time Object Detection". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2019), pp. 752–760 (cit. on p. 180).

[221] Constance D Lehman et al. "Diagnostic accuracy of digital screening mammography with and without computer-aided detection". In: *JAMA internal medicine* 175.11 (2015), pp. 1828–1837 (cit. on p. 18).

[222] Andreas Leibetseder et al. "Lapgyn4: a dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology". In: *Proceedings of the 9th ACM multimedia systems conference*. 2018, pp. 357–362 (cit. on pp. 24, 28).

[223] Jochen K Lennerz et al. "A unifying force for the realization of medical AI". In: *NPJ Digital Medicine* 5.1 (2022), p. 172 (cit. on p. 152).

[224] Hengduo Li et al. "An analysis of pre-training on object detection". In: *arXiv preprint arXiv:1904.05871* (2019) (cit. on p. 225).

[225] Qinbin Li et al. "A survey on federated learning systems: Vision, hype and reality for data privacy and protection". In: *IEEE Transactions on Knowledge and Data Engineering* 35.4 (2021), pp. 3347–3366 (cit. on p. 86).

[226] Xiang Li et al. "Selective Kernel Networks". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 510–519 (cit. on p. 180).

[227] Kung-Yee Liang and Scott L Zeger. "Longitudinal data analysis using generalized linear models". In: *Biometrika* 73.1 (1986), pp. 13–22 (cit. on p. 120).

[228] Tsung-Yi Lin et al. "Microsoft coco: Common objects in context". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755 (cit. on p. 58).

[229] Seppo Linnainmaa. "The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors". PhD thesis. Master's Thesis (in Finnish), Univ. Helsinki, 1970 (cit. on p. 4).

[230] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. "Detecting and correcting for label shift with black box predictors". In: *International conference on machine learning*. PMLR. 2018, pp. 3122–3130 (cit. on pp. 89, 198, 200, 201, 206).

[231] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88 (cit. on pp. 9, 18–20).

[232] Ruhan Liu et al. "Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge". In: *Patterns* 3.6 (2022) (cit. on pp. 24, 32, 178).

[233] Xinran Liu et al. "Wasserstein task embedding for measuring task similarities". In: *Neural Networks* 181 (2025), p. 106796 (cit. on pp. 87, 172).

[234] Guolan Lu and Baowei Fei. "Medical hyperspectral imaging: a review". In: *Journal of biomedical optics* 19.1 (2014), pp. 010901–010901 (cit. on p. 33).

[235] Carsten Lüth et al. "Navigating the pitfalls of active learning evaluation: A systematic framework for meaningful performance assessment". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 9789–9836 (cit. on p. 86).

[236] Lena Maier-Hein et al. "Can masses of non-experts train highly accurate image classifiers? A crowdsourcing approach to instrument segmentation in laparoscopic images". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II 17*. Springer. 2014, pp. 438–445 (cit. on p. 86).

[237] Lena Maier-Hein et al. "Metrics reloaded: Recommendations for image analysis validation". In: *arXiv preprint arXiv:2206.01653* (2022) (cit. on pp. 99, 237).

[238] Lena Maier-Hein et al. "Metrics reloaded: recommendations for image analysis validation". In: *Nature methods* 21.2 (2024), pp. 195–212 (cit. on pp. 95, 100, 102, 103, 105, 108–110, 114, 115, 120–126, 128, 129, 132, 135–139, 141, 143, 147–151, 232, 237).

[239] Lena Maier-Hein et al. "Surgical data science–from concepts toward clinical translation". In: *Medical image analysis* 76 (2022), p. 102306 (cit. on pp. 8–10).

[240] Lena Maier-Hein et al. "Why rankings of biomedical image analysis competitions should be interpreted with care". In: *Nature communications* 9.1 (2018), p. 5217 (cit. on pp. 9, 42, 58, 83, 85, 101, 107, 127, 202, 209, 224).

[241] JI Marcum. "A Statistical Theory of Target Detection by Pulsed Radar". In: (1947) (cit. on p. 58).

[242] Florian Markowetz. "All models are wrong and yours are useless: making clinical prediction models impactful for patients". In: *npj Precision Oncology* 8.1 (2024), p. 54 (cit. on pp. 9, 19).

[243] Christos Matsoukas et al. "What makes transfer learning work for medical images: Feature reuse & other factors". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 9225–9234 (cit. on p. 9).

[244] Brian W Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405.2 (1975), pp. 442–451 (cit. on p. 54).

[245] Melissa D McCradden et al. "A research ethics framework for the clinical translation of healthcare machine learning". In: *The American Journal of Bioethics* 22.5 (2022), pp. 8–22 (cit. on p. 153).

[246] Warren S McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5 (1943), pp. 115–133 (cit. on p. 4).

[247] Teresa Mendonça et al. "Ph2: A public database for the analysis of dermoscopic images". In: *Dermoscopy image analysis* 2 (2015) (cit. on pp. 27, 32).

[248] Thomas Mensink et al. "Factors of influence for transfer learning across diverse appearance domains and task types". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.12 (2021), pp. 9298–9314 (cit. on p. 191).

[249] George A Miller. "WordNet: a lexical database for English". In: *Communications of the ACM* 38.11 (1995), pp. 39–41 (cit. on p. 28).

[250] Bonan Min et al. "Recent advances in natural language processing via large pre-trained language models: A survey". In: *ACM Computing Surveys* 56.2 (2023), pp. 1–40 (cit. on p. 18).

[251] Matthias Minderer et al. "Revisiting the calibration of modern neural networks". In: *Advances in neural information processing systems* 34 (2021), pp. 15682–15694 (cit. on p. 226).

[252] Marvin Minsky and Seymour Papert. "An introduction to computational geometry". In: *Cambridge tiass., HIT* 479.480 (1969), p. 104 (cit. on p. 4).

[253] Diganta Misra. "Mish: A self regularized non-monotonic activation function". In: *arXiv preprint arXiv:1908.08681* (2019) (cit. on p. 73).

[254] Tom M Mitchell and Tom M Mitchell. *Machine learning*. Vol. 1. 9. McGraw-hill New York, 1997 (cit. on p. 70).

[255] Sara Moccia, Elena De Momi, and Leonardo S. Mattos. *Laryngeal dataset*. Oct. 2017. DOI: 10.5281/zenodo.1003200. URL: https://doi.org/10.5281/zenodo.1003200 (cit. on pp. 26, 29).

[256] Sara Moccia et al. *NBI-InfFrames*. Jan. 2018. DOI: 10.5281/zenodo.1162784. URL: https://doi.org/10.5281/zenodo.1162784 (cit. on pp. 26, 29).

[257] Miguel Molina-Moreno et al. "Automated Style-Aware Selection of Annotated Pre-Training Databases in Biomedical Imaging". In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2023, pp. 1–5 (cit. on p. 87).

[258] Gaspard Monge. "Mémoire sur la théorie des déblais et des remblais". In: *Mem. Math. Phys. Acad. Royale Sci.* (1781), pp. 666–704 (cit. on p. 162).

[259] Université de Montréal. *The Declaration - Montreal Responsible AI*. en. 2017. URL: https://www.montrealdeclaration-responsibleai.com/the-declaration (visited on 09/09/2022) (cit. on p. 153).

[260] Karel GM Moons et al. "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration". In: *Annals of internal medicine* 162.1 (2015), W1–W73 (cit. on pp. 85, 111, 152).

[261] Jose G Moreno-Torres et al. "A unifying view on dataset shift in classification". In: *Pattern recognition* 45.1 (2012), pp. 521–530 (cit. on pp. 11, 88, 197, 198).

[262] Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. "QuaPy: A Python-based framework for quantification". In: *Proceedings of the 30th ACM international conference on information & knowledge management*. 2021, pp. 4534–4543 (cit. on p. 206).

[263] Alejandro Moreo, Pablo González, and Juan José del Coz. "Kernel density estimation for multiclass quantification". In: *Machine Learning* 114.4 (2025), p. 92 (cit. on pp. 89, 201, 206, 209, 216, 217).

[264] Allan H Murphy. "A new vector partition of the probability score". In: *Journal of Applied Meteorology and Climatology* 12.4 (1973), pp. 595–600 (cit. on p. 69).

[265] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. "Obtaining well calibrated probabilities using bayesian binning". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 29. 1. 2015 (cit. on p. 65).

[266] Yukiko Nagao et al. "Robust classification of cell cycle phase and biological feature extraction by image-based deep learning". In: *Molecular biology of the cell* 31.13 (2020), pp. 1346–1354 (cit. on p. 145).

[267] Prashant Nasa, Ravi Jain, and Deven Juneja. "Delphi methodology in healthcare research: how to decide its appropriateness". In: *World journal of methodology* 11.4 (2021), p. 116 (cit. on p. 96).

[268] Yuval Netzer et al. "Reading digits in natural images with unsupervised feature learning". In: *NIPS workshop on deep learning and unsupervised feature learning*. Vol. 2011. 2. Granada. 2011, p. 4 (cit. on pp. 27, 29).

[269] Cuong Nguyen, Thanh-Toan Do, and Gustavo Carneiro. "Similarity of classification tasks". In: *arXiv preprint arXiv:2101.11201* (2021) (cit. on p. 87).

[270] Cuong Nguyen et al. “Leep: A new measure to evaluate transferability of learned representations”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7294–7305 (cit. on p. 87).

[271] Beau Norgeot et al. “Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist”. In: *Nature medicine* 26.9 (2020), pp. 1320–1324 (cit. on p. 85).

[272] Robert J O’Shea et al. “Systematic review of research design and reporting of imaging studies applying convolutional neural networks for radiological cancer diagnosis”. In: *European Radiology* 31 (2021), pp. 7969–7983 (cit. on p. 85).

[273] Luke Oakden-Rayner et al. “Hidden stratification causes clinically meaningful failures in machine learning for medical imaging”. In: *Proceedings of the ACM conference on health, inference, and learning*. 2020, pp. 151–159 (cit. on p. 154).

[274] Ziad Obermeyer et al. “Dissecting racial bias in an algorithm used to manage the health of populations”. In: *Science* 366.6464 (2019), pp. 447–453 (cit. on p. 153).

[275] and others. “Sources of performance variability in deep learning-based polyp detection”. In: *International Journal of Computer Assisted Radiology and Surgery* 18.7 (2023), pp. 1311–1322 (cit. on p. 83).

[276] Frédéric Ouimet and Raimon Tolosana-Delgado. “Asymptotic properties of Dirichlet kernel density estimators”. In: *Journal of Multivariate Analysis* 187 (2022), p. 104832 (cit. on p. 67).

[277] Yaniv Ovadia et al. “Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift”. In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 89).

[278] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359 (cit. on pp. 6, 10, 78, 79).

[279] Victor M Panaretos and Yoav Zemel. “Statistical aspects of Wasserstein distances”. In: *Annual review of statistics and its application* 6.1 (2019), pp. 405–431 (cit. on pp. 163, 164).

[280] Trishan Panch, Heather Mattie, and Leo Anthony Celi. “The “inconvenient truth” about AI in healthcare”. In: *NPJ digital medicine* 2.1 (2019), pp. 1–3 (cit. on pp. 8, 9, 11, 20).

[281] Michael Panchenko, Anes Benmerzoug, and Miguel de Benito Delgado. “Class-wise and reduced calibration methods”. In: *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2022, pp. 1093–1100 (cit. on pp. 64, 66).

[282] Michal Pándy et al. "Transferability estimation using bhattacharyya class separability". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 9172–9182 (cit. on pp. 172, 192).

[283] German I Parisi et al. "Continual lifelong learning with neural networks: A review". In: *Neural networks* 113 (2019), pp. 54–71 (cit. on p. 81).

[284] Seong Ho Park and Kyunghwa Han. "Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction". In: *Radiology* 286.3 (2018), pp. 800–809 (cit. on p. 85).

[285] Seong Ho Park et al. "Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis". In: *Radiology* 306.1 (2023), pp. 20–31 (cit. on p. 153).

[286] Emanuel Parzen. "On estimation of a probability density function and mode". In: *The annals of mathematical statistics* 33.3 (1962), pp. 1065–1076 (cit. on p. 67).

[287] Blaise Pascal. *La Machine d'arithmétique. Lettre dédicatoire à Monseigneur le Chancelier*. 1645 (cit. on p. 3).

[288] David Patterson et al. "Carbon emissions and large neural network training". In: *arXiv preprint arXiv:2104.10350* (2021) (cit. on p. 155).

[289] Stephen G Pauker and Jerome P Kassirer. "Therapeutic decision making: a cost-benefit analysis". In: *New England Journal of Medicine* 293.5 (1975), pp. 229–234 (cit. on p. 133).

[290] Judea Pearl et al. "Models, reasoning and inference". In: *Cambridge, UK: Cambridge University Press* 19.2 (2000), p. 3 (cit. on p. 197).

[291] Judea Pearl and Elias Bareinboim. "External validity: From do-calculus to transportability across populations". In: *Probabilistic and causal inference: The works of Judea Pearl*. 2022, pp. 451–482 (cit. on p. 20).

[292] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830 (cit. on pp. 85, 209).

[293] Charles S Peirce. "The numerical measure of the success of predictions". In: *Science* 93 (1884), pp. 453–454 (cit. on p. 48).

[294] Xingchao Peng, Yichen Li, and Kate Saenko. "Domain2vec: Domain embedding for unsupervised domain adaptation". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer. 2020, pp. 756–774 (cit. on pp. 87, 172).

[295] Alexandre Perez-Lebel, Marine Le Morvan, and Gaël Varoquaux. "Beyond calibration: estimating the grouping loss of modern neural networks". In: *International Conference on Learning Representations* (2023) (cit. on p. 152).

[296] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017 (cit. on p. 197).

[297] Allan Pinkus. "Approximation theory of the MLP model in neural networks". In: *Acta numerica* 8 (1999), pp. 143–195 (cit. on p. 72).

[298] John Platt et al. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in large margin classifiers* 10.3 (1999), pp. 61–74 (cit. on p. 201).

[299] Konstantin Pogorelov et al. "Nerthus: A bowel preparation quality video dataset". In: *Proceedings of the 8th ACM on Multimedia Systems Conference*. 2017, pp. 170–174 (cit. on pp. 25, 28).

[300] Russell A Poldrack, Grace Huckins, and Gael Varoquaux. "Establishment of best practices for evidence for prediction: a review". In: *JAMA psychiatry* 77.5 (2020), pp. 534–540 (cit. on p. 85).

[301] Phillip Pope et al. "The intrinsic dimension of images and its impact on learning". In: *arXiv preprint arXiv:2104.08894* (2021) (cit. on p. 177).

[302] Teodora Popordanoska, Raphael Sayer, and Matthew Blaschko. "A consistent and differentiable lp canonical calibration error estimator". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 7933–7946 (cit. on pp. 67, 140, 144).

[303] David Powers. "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". In: *Journal of Machine Learning Technologies* 2.1 (2011), pp. 37–63 (cit. on pp. 48, 51, 54).

[304] David MW Powers. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". In: *arXiv preprint arXiv:2010.16061* (2020) (cit. on p. 48).

[305] Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. "Tunability: Importance of hyperparameters of machine learning algorithms". In: *Journal of Machine Learning Research* 20.53 (2019), pp. 1–32 (cit. on p. 76).

[306] Joaquin Quinonero-Candela et al. *Dataset Shift in Machine Learning*. MIT Press, 2008 (cit. on pp. 88, 197).

[307] Joaquin Quinonero-Candela et al. "Evaluating predictive uncertainty challenge". In: *Machine Learning Challenges Workshop*. Springer. 2005, pp. 1–27 (cit. on p. 143).

[308] Ilija Radosavovic et al. "Designing Network Design Spaces". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 10425–10433 (cit. on p. 180).

[309] Tim Rädsch et al. "Labelling instructions matter in biomedical image analysis". In: *Nature Machine Intelligence* 5.3 (2023), pp. 273–283 (cit. on p. 86).

[310] Maithra Raghu et al. "Transfusion: Understanding transfer learning for medical imaging". In: *Advances in neural information processing systems* 32 (2019) (cit. on pp. 9, 87, 165, 176, 178, 192).

[311] Pranav Rajpurkar et al. "Mura dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs". In: *Medical imaging with deep learning*. 2017 (cit. on pp. 22, 31).

[312] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. "On wasserstein two-sample testing and related families of nonparametric tests". In: *Entropy* 19.2 (2017), p. 47 (cit. on p. 163).

[313] Benjamin Ramtoula et al. "Visual dna: Representing and comparing images using distributions of neuron activations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 11113–11123 (cit. on pp. 87, 164, 186, 187).

[314] Khalid Raza and Nripendra K Singh. "A tour of unsupervised deep learning for medical image analysis". In: *Current Medical Imaging Reviews* 17.9 (2021), pp. 1059–1077 (cit. on p. 86).

[315] Annika Reinke and et al. *Metrics reloaded*. Dec. 2022. URL: `http://www.cse.cuhk.edu.hk/~qdou/public/medneurips2022/52.pdf` (cit. on pp. 95, 237).

[316] Annika Reinke et al. "Metrics Reloaded-A new recommendation framework for biomedical image analysis validation". In: *Medical Imaging with Deep Learning*. 2022 (cit. on pp. 95, 237).

[317] Annika Reinke et al. "Understanding metric-related pitfalls in image analysis validation". In: *Nature methods* 21.2 (2024), pp. 182–194 (cit. on pp. 98, 104, 119, 120, 192, 199, 231, 237).

[318] Cedric Renggli et al. "Which model to transfer? finding the needle in the growing haystack". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 9205–9214 (cit. on pp. 171, 176).

[319] Nicola Rieke et al. "The future of digital health with federated learning". In: *NPJ digital medicine* 3.1 (2020), p. 119 (cit. on p. 86).

[320] Richard D Riley et al. "External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges". In: *bmj* 353 (2016) (cit. on p. 125).

[321] Torsten Rohlfing. "Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable". In: *IEEE transactions on medical imaging* 31.2 (2011), pp. 153–163 (cit. on p. 83).

[322] Wilhelm Conrad Röntgen. *Ueber eine neue Art von Strahlen*. Phys.-med. Gesellschaft, 1895 (cit. on p. 17).

[323] Mélanie Roschewitz et al. “Automatic correction of performance drift under acquisition shift in medical image classification”. In: *Nature Communications* 14.1 (2023), p. 6608 (cit. on p. 218).

[324] F Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. 1961 (cit. on p. 4).

[325] Frank Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), p. 386 (cit. on pp. 4, 71, 72).

[326] Murray Rosenblatt. “Remarks on Some Nonparametric Estimates of a Density Function”. In: *The Annals of Mathematical Statistics* (1956), pp. 832–837 (cit. on p. 67).

[327] Tobias Roß et al. “Beyond rankings: learning (more) from algorithm validation”. In: *Medical image analysis* 86 (2023), p. 102765 (cit. on p. 154).

[328] Veronica Rotemberg et al. “A patient-centric dataset of images and metadata for identifying melanomas using clinical context”. In: *Scientific data* 8.1 (2021), p. 34 (cit. on pp. 27, 32, 77).

[329] S Ruder. “An Overview of Multi-Task Learning in Deep Neural Networks”. In: *arXiv preprint arXiv:1706.05098* (2017) (cit. on pp. 77, 78).

[330] 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 pp. 4, 71).

[331] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115 (2015), pp. 211–252 (cit. on p. 77).

[332] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016 (cit. on p. 70).

[333] Marco Saerens, Patrice Latinne, and Christine Decaestecker. “Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure”. In: *Neural computation* 14.1 (2002), pp. 21–41 (cit. on pp. 89, 201, 206).

[334] Pedro Sanchez et al. “Causal machine learning for healthcare and precision medicine”. In: *Royal Society Open Science* 9.8 (2022), p. 220638 (cit. on p. 88).

[335] Suchi Saria and Adarsh Subbaswamy. “Tutorial: safe and reliable machine learning”. In: *arXiv preprint arXiv:1904.07204* (2019) (cit. on p. 203).

[336] Yutaka Sasaki. “The truth of the F-measure”. In: (2007) (cit. on pp. 52, 137).

[337] Leonard J Savage. “On Rereading R. A. Fisher”. In: *The Annals of Statistics* (1976), pp. 441–500 (cit. on p. 165).

[338] J Scherer et al. *Die Joint Imaging Platform (JIP) des Deutschen Konsortiums für Translationale Krebsforschung (DKTK)*. © Georg Thieme Verlag KG, 2020 (cit. on p. 193).

[339] Enrique F Schisterman et al. "Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples". In: *Epidemiology* 16.1 (2005), pp. 73–81 (cit. on p. 58).

[340] Jürgen Schmidhuber. "Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook". PhD thesis. Technische Universität München, 1987 (cit. on p. 6).

[341] Patrick Schober, Christa Boer, and Lothar A Schwarte. "Correlation coefficients: appropriate use and interpretation". In: *Anesthesia & analgesia* 126.5 (2018), pp. 1763–1768 (cit. on p. 172).

[342] Patrick Scholz and Lena Maier-Hein. *Quantification of Task Similarity for Efficient Knowledge Transfer in Biomedical Image Analysis*. Dec. 2020. URL: http://www.cse.cuhk.edu.hk/~qdou/public/medneurips2020/Quantification%20of%20task%20similarity%20for%20efficientknowledge%20transfer%20in%20biomedical%20image%20analysis.pdf (cit. on pp. 157, 232, 238).

[343] Peter Schulam and Suchi Saria. "Can you trust this prediction? Auditing pointwise reliability after learning". In: *The 22nd international conference on artificial intelligence and statistics*. PMLR. 2019, pp. 1022–1031 (cit. on p. 153).

[344] Kenneth F Schulz et al. "CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials". In: *Annals of internal medicine* 152.11 (2010), pp. 726–732 (cit. on p. 152).

[345] Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. "A comparative evaluation of quantification methods". In: *arXiv preprint arXiv:2103.03223* (2021) (cit. on p. 89).

[346] Fabrizio Sebastiani. "An axiomatically derived measure for the evaluation of classification algorithms". In: *Proceedings of the 2015 international conference on the theory of information retrieval*. 2015, pp. 11–20 (cit. on p. 84).

[347] Fabrizio Sebastiani. "Evaluation measures for quantification: An axiomatic approach". In: *Information Retrieval Journal* 23.3 (2020), pp. 255–288 (cit. on p. 209).

[348] Nigam H Shah, Arnold Milstein, and Steven C Bagley. "Making machine learning models clinically useful". In: *Jama* 322.14 (2019), pp. 1351–1352 (cit. on p. 8).

[349] Bobak Shahriari et al. "Taking the human out of the loop: A review of Bayesian optimization". In: *Proceedings of the IEEE* 104.1 (2015), pp. 148–175 (cit. on p. 76).

[350] Shai Shalev-Shwartz et al. "Online learning and online convex optimization". In: *Foundations and Trends® in Machine Learning* 4.2 (2012), pp. 107–194 (cit. on p. 80).

[351] Grace S Shieh. "A weighted Kendall's tau statistic". In: *Statistics & probability letters* 39.1 (1998), pp. 17–24 (cit. on p. 173).

[352] Hidetoshi Shimodaira. "Improving predictive inference under covariate shift by weighting the log-likelihood function". In: *Journal of statistical planning and inference* 90.2 (2000), pp. 227–244 (cit. on p. 201).

[353] Connor Shorten and Taghi M Khoshgoftaar. "A survey on image data augmentation for deep learning". In: *Journal of big data* 6.1 (2019), pp. 1–48 (cit. on p. 181).

[354] David Silver et al. "Mastering chess and shogi by self-play with a general reinforcement learning algorithm". In: *arXiv preprint arXiv:1712.01815* (2017) (cit. on p. 5).

[355] Iveta Simera et al. "Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network". In: *BMC medicine* 8 (2010), pp. 1–6 (cit. on p. 97).

[356] Ana-Maria Šimundić. "Measures of diagnostic accuracy: basic definitions". In: *ejifcc* 19.4 (2009), p. 203 (cit. on p. 48).

[357] Amit Singhal et al. "Modern information retrieval: A brief overview". In: *IEEE Data Eng. Bull.* 24.4 (2001), pp. 35–43 (cit. on p. 166).

[358] Pia H Smedsrud et al. "Kvasir-Capsule, a video capsule endoscopy dataset". In: *Scientific Data* 8.1 (2021), p. 142 (cit. on pp. 26, 32).

[359] Leslie N Smith. "Cyclical learning rates for training neural networks". In: *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2017, pp. 464–472 (cit. on pp. 178, 205).

[360] Linda B. Smith and Michael Gasser. "The Development of Embodied Cognition: Six Lessons from Babies". In: *Artificial Life* 11 (2005), pp. 13–29. URL: https://api.semanticscholar.org/CorpusID:7107473 (cit. on p. 7).

[361] Jake Snell, Kevin Swersky, and Richard Zemel. "Prototypical networks for few-shot learning". In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 80).

[362] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. "Practical bayesian optimization of machine learning algorithms". In: *Advances in neural information processing systems* 25 (2012) (cit. on p. 76).

[363] Marina Sokolova and Guy Lapalme. "A systematic analysis of performance measures for classification tasks". In: *Information processing & management* 45.4 (2009), pp. 427–437 (cit. on pp. 84, 101).

[364] Andrea Soltoggio et al. "A collective AI via lifelong learning and sharing at the edge". In: *Nature Machine Intelligence* 6.3 (2024), pp. 251–264 (cit. on pp. 90, 91).

[365] Le Song, Kenji Fukumizu, and Arthur Gretton. "Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models". In: *IEEE Signal Processing Magazine* 30.4 (2013), pp. 98–111 (cit. on p. 161).

[366] Nitish Srivastava et al. "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. 74).

[367] Trevor Standley et al. "Which tasks should be learned together in multi-task learning?" In: *International conference on machine learning*. PMLR. 2020, pp. 9120–9132 (cit. on pp. 78, 182, 191).

[368] Ewout W Steyerberg et al. "Assessing the performance of prediction models: a framework for traditional and novel measures". In: *Epidemiology* 21.1 (2010), pp. 128–138 (cit. on pp. 84, 125).

[369] Petre Stoica and Prabhu Babu. "Pearson–Matthews correlation coefficients for binary and multinary classification". In: *Signal Processing* 222 (2024), p. 109511 (cit. on p. 54).

[370] Marilyn Strathern. "'Improving ratings': audit in the British University system". In: *European review* 5.3 (1997), pp. 305–321 (cit. on p. 83).

[371] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and policy considerations for modern deep learning research". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 09. 2020, pp. 13693–13696 (cit. on p. 155).

[372] Adarsh Subbaswamy and Suchi Saria. "From development to deployment: dataset shift, causality, and shift-stable models in health AI". In: *Biostatistics* 21.2 (2020), pp. 345–352 (cit. on pp. 9, 11, 88).

[373] Carole Sudre and Paul Smith. *MetricsReloaded*. Version 0.1.0. 2022. URL: `https://github.com/Project-MONAI/MetricsReloaded` (cit. on pp. 119, 120, 152).

[374] Cecilia Summers and Michael J Dinneen. "Nondeterminism and instability in neural network optimization". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 9913–9922 (cit. on p. 120).

[375] Ke Sun et al. "High-Resolution Representations for Labeling Pixels and Regions". In: *ArXiv* abs/1904.04514 (2019) (cit. on p. 180).

[376] John A Swets. "The Relative Operating Characteristic in Psychology: A technique for isolating effects of response bias finds wide use in the study of perception and cognition." In: *Science* 182.4116 (1973), pp. 990–1000 (cit. on p. 47).

[377] Abdel Aziz Taha and Allan Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool". In: *BMC medical imaging* 15 (2015), pp. 1–28 (cit. on pp. 84, 101, 154).

[378] Chuanqi Tan et al. “A survey on deep transfer learning”. In: *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*. Springer. 2018, pp. 270–279 (cit. on p. 78).

[379] Mingxing Tan and Quoc V. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *ArXiv* abs/1905.11946 (2019) (cit. on p. 180).

[380] Mingxing Tan and Quoc V. Le. “MixConv: Mixed Depthwise Convolutional Kernels”. In: *ArXiv* abs/1907.09595 (2019) (cit. on p. 180).

[381] Yang Tan, Yang Li, and Shao-Lun Huang. “Otce: A transferability metric for cross-domain cross-task representations”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 15779–15788 (cit. on pp. 87, 166, 172).

[382] Rohan Taori et al. “Measuring robustness to natural distribution shifts in image classification”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 18583–18599 (cit. on p. 197).

[383] The Institute for Ethical Ai and Machine Learning. *The Institute for Ethical AI & Machine Learning*. `https://ethical.institute/principles.html`. Accessed: 2022-5-21. 2018 (cit. on p. 153).

[384] Rachel L Thomas and David Uminsky. “Reliance on metrics is a fundamental challenge for AI”. In: *Patterns* 3.5 (2022) (cit. on pp. 83, 84).

[385] Sir W Thomson. “The tide gauge, tidal harmonic analyser, and tide predicter”. In: *Minutes of the Proceedings of the Institution of Civil Engineers*. Vol. 65. 1881. Thomas Telford-ICE Virtual Library. 1881, pp. 2–25 (cit. on p. 3).

[386] Sebastian Thrun and Lorien Pratt. “Learning to learn: Introduction and overview”. In: *Learning to learn*. Springer, 1998, pp. 3–17 (cit. on p. 7).

[387] Christian Tomani et al. “Post-hoc uncertainty calibration for domain drift scenarios”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 10124–10132 (cit. on p. 140).

[388] Anh T Tran, Cuong V Nguyen, and Tal Hassner. “Transferability and hardness of supervised classification tasks”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1395–1405 (cit. on p. 87).

[389] A Turing. “On computable numbers, with an application to the Entscheidungs problem”. In: *Proceedings of the London Mathematical Society Series/2 (42)* (1936), pp. 230–42 (cit. on p. 4).

[390] Andru P Twinanda et al. “Endonet: a deep architecture for recognition tasks on laparoscopic videos”. In: *IEEE transactions on medical imaging* 36.1 (2016), pp. 86–97 (cit. on pp. 23, 28, 29, 178).

[391] Richard Usatine and Rachel Manci. *Dermoscopedia*. `https://dermoscopedia.org/File:DF_chinese_dms.JPG`. 2021 (cit. on p. 145).

[392] Femke Vaassen et al. "Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy". In: *Physics and Imaging in Radiation Oncology* 13 (2020), pp. 1–6 (cit. on p. 83).

[393] Juozas Vaicenavicius et al. "Evaluating model calibration in classification". In: *The 22nd international conference on artificial intelligence and statistics*. PMLR. 2019, pp. 3459–3467 (cit. on pp. 61–65, 144).

[394] Ben Van Calster et al. "Calibration: the Achilles heel of predictive analytics". In: *BMC medicine* 17.1 (2019), p. 230 (cit. on p. 113).

[395] David A Van Leeuwen and Niko Brümmer. *An introduction to application-independent evaluation of speaker recognition systems*. Springer, 2007 (cit. on p. 127).

[396] Cornelius Joost Van Rijsbergen. *Information retrieval. 2nd. newton, ma*. 1979 (cit. on p. 52).

[397] Simon Vandenhende et al. "Multi-task learning for dense prediction tasks: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 44.7 (2021), pp. 3614–3633 (cit. on pp. 77, 78).

[398] J Vanschoren. "Meta-Learning: A Survey". In: *arXiv preprint arXiv:1810.03548* (2018) (cit. on p. 80).

[399] Gaël Varoquaux and Veronika Cheplygina. "Machine learning for medical imaging: methodological failures and recommendations for the future". In: *NPJ digital medicine* 5.1 (2022), p. 48 (cit. on pp. 19, 20, 85).

[400] Gaël Varoquaux and Olivier Colliot. "Evaluating machine learning models and their diagnostic value". In: *Machine learning for brain disorders* (2023), pp. 601–630 (cit. on pp. 69, 84).

[401] Leonid Nisonovich Vaserstein. "Markov processes over denumerable products of spaces, describing large systems of automata". In: *Problemy Peredachi Informatsii* 5.3 (1969), pp. 64–72 (cit. on p. 162).

[402] Andrew J Vickers and Elena B Elkin. "Decision curve analysis: a novel method for evaluating prediction models". In: *Medical Decision Making* 26.6 (2006), pp. 565–574 (cit. on p. 59).

[403] Andrew J Vickers, Ben Van Calster, and Ewout W Steyerberg. "Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests". In: *bmj* 352 (2016) (cit. on pp. 59, 111).

[404] Sebastiano Vigna. "A weighted correlation index for rankings with ties". In: *Proceedings of the 24th international conference on World Wide Web*. 2015, pp. 1166–1176 (cit. on p. 173).

[405] Cédric Villani et al. *Optimal transport: old and new*. Vol. 338. Springer, 2009 (cit. on p. 163).

[406] John Von Neumann. "First Draft of a Report on the EDVAC". In: *IEEE Annals of the History of Computing* 15.4 (1993), pp. 27–75 (cit. on p. 4).

[407] John Von Neumann, Arthur Walter Burks, et al. "Theory of self-reproducing automata". In: (1966) (cit. on p. 4).

[408] Ž Vujović et al. "Classification model evaluation metrics". In: *International Journal of Advanced Computer Science and Applications* 12.6 (2021), pp. 599–606 (cit. on p. 83).

[409] Chien-Yao Wang et al. "CSPNet: A New Backbone that can Enhance Learning Capability of CNN". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2020), pp. 1571–1580 (cit. on p. 180).

[410] Liyuan Wang et al. "A comprehensive survey of continual learning: theory, method and application". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024) (cit. on p. 79).

[411] Mei Wang and Weihong Deng. "Deep visual domain adaptation: A survey". In: *Neurocomputing* 312 (2018), pp. 135–153 (cit. on p. 79).

[412] Qilong Wang et al. "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 11531–11539 (cit. on p. 180).

[413] Xin Wang, Yudong Chen, and Wenwu Zhu. "A survey on curriculum learning". In: *IEEE transactions on pattern analysis and machine intelligence* 44.9 (2021), pp. 4555–4576 (cit. on p. 80).

[414] Matthijs J Warrens. "Some paradoxical results for the quadratically weighted kappa". In: *Psychometrika* 77 (2012), pp. 315–323 (cit. on pp. 46, 117, 127).

[415] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. "A survey of transfer learning". In: *Journal of Big data* 3 (2016), pp. 1–40 (cit. on p. 78).

[416] Jonathan Wenger, Hedvig Kjellström, and Rudolph Triebel. "Non-parametric calibration for classification". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 178–190 (cit. on p. 142).

[417] Paul J Werbos. "Applications of advances in nonlinear sensitivity analysis". In: *System Modeling and Optimization: Proceedings of the 10th IFIP Conference New York City, USA, August 31–September 4, 1981*. Springer. 2005, pp. 762–770 (cit. on p. 4).

[418] Stephen A White. "Introduction to BPMN". In: *Ibm Cooperation* 2 (2004) (cit. on p. 121).

[419] David Widmann, Fredrik Lindsten, and Dave Zachariah. "Calibration tests in multi-class classification: A unifying framework". In: *Advances in neural information processing systems* 32 (2019) (cit. on pp. 66, 67).

[420] Manuel Wiesenfarth et al. "Methods and open-source toolkit for analyzing and visualizing challenge results". In: *Scientific reports* 11.1 (2021), pp. 1–15 (cit. on pp. 85, 119, 120, 175, 176, 186, 188, 192).

[421] Ross Wightman. *PyTorch Image Models*. Version 1.0.11. DOI: 10.5281/zenodo.4414861. URL: https://github.com/huggingface/pytorch-image-models (cit. on pp. 180, 181, 205).

[422] Ross Wightman, Hugo Touvron, and Hervé Jégou. "Resnet strikes back: An improved training procedure in timm". In: *arXiv preprint arXiv:2110.00476* (2021) (cit. on pp. 74, 178, 181).

[423] Martin J Willemink et al. "Preparing medical imaging data for machine learning". In: *Radiology* 295.1 (2020), pp. 4–15 (cit. on p. 19).

[424] Laure Wynants et al. "Three myths about risk thresholds for prediction models". In: *BMC medicine* 17 (2019), pp. 1–7 (cit. on p. 111).

[425] Cihang Xie et al. "Adversarial Examples Improve Image Recognition". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 816–825 (cit. on p. 180).

[426] Qizhe Xie et al. "Self-Training With Noisy Student Improves ImageNet Classification". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 10684–10695 (cit. on p. 180).

[427] Xiaohan Xu et al. "A survey on knowledge distillation of large language models". In: *arXiv preprint arXiv:2402.13116* (2024) (cit. on p. 226).

[428] Ismet Zeki Yalniz et al. "Billion-scale semi-supervised learning for image classification". In: *ArXiv* abs/1905.00546 (2019) (cit. on p. 180).

[429] Brandon Yang et al. "CondConv: Conditionally Parameterized Convolutions for Efficient Inference". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019 (cit. on p. 180).

[430] Suorong Yang et al. "Image data augmentation for deep learning: A survey". In: *arXiv preprint arXiv:2204.08610* (2022) (cit. on p. 181).

[431] Xingyi Yang et al. "Covid-ct-dataset: a ct scan dataset about covid-19". In: *arXiv preprint arXiv:2003.13865* (2020) (cit. on pp. 23, 31).

[432] Yinchong Yang and Florian Buettner. "Multi-output gaussian processes for uncertainty-aware recommender systems". In: *Uncertainty in Artificial Intelligence*. PMLR. 2021, pp. 1505–1514 (cit. on p. 113).

[433] Kaichao You et al. “Logme: Practical assessment of pre-trained models for transfer learning”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12133–12143 (cit. on pp. 87, 172, 173, 192).

[434] William J Youden. “Index for rating diagnostic tests”. In: *Cancer* 3.1 (1950), pp. 32–35 (cit. on p. 48).

[435] Tianhe Yu et al. “Gradient surgery for multi-task learning”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 5824–5836 (cit. on pp. 77, 78).

[436] G Udny Yule. “On the methods of measuring association between two attributes”. In: *Journal of the Royal Statistical Society* 75.6 (1912), pp. 579–652 (cit. on p. 54).

[437] Sangdoo Yun et al. “Cutmix: Regularization strategy to train strong classifiers with localizable features”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6023–6032 (cit. on p. 75).

[438] Amir R Zamir et al. “Taskonomy: Disentangling task transfer learning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3712–3722 (cit. on pp. 87, 170, 181, 185).

[439] John R Zech et al. “Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study”. In: *PLoS medicine* 15.11 (2018), e1002683 (cit. on pp. 88, 198).

[440] Angela Zhang et al. “Shifting machine learning for healthcare from development to deployment and from models to data”. In: *Nature biomedical engineering* 6.12 (2022), pp. 1330–1345 (cit. on pp. 88, 198).

[441] Hang Zhang et al. “ResNeSt: Split-Attention Networks”. In: *ArXiv* abs/2004.08955 (2020) (cit. on p. 180).

[442] Hongyi Zhang. “mixup: Beyond empirical risk minimization”. In: *arXiv preprint arXiv:1710.09412* (2017) (cit. on p. 75).

[443] Kun Zhang et al. “Domain adaptation under target and conditional shift”. In: *International conference on machine learning*. Pmlr. 2013, pp. 819–827 (cit. on pp. 89, 198, 200, 201).

[444] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595 (cit. on p. 159).

[445] Wen Zhang et al. “A survey on negative transfer”. In: *IEEE/CAA Journal of Automatica Sinica* 10.2 (2022), pp. 305–329 (cit. on p. 87).

[446] Ying Zhang et al. “DeepPhagy: a deep learning framework for quantitatively measuring autophagy activity in Saccharomyces cerevisiae”. In: *Autophagy* 16.4 (2020), pp. 626–640 (cit. on p. 145).

[447] Yu Zhang and Qiang Yang. "A survey on multi-task learning". In: *IEEE transactions on knowledge and data engineering* 34.12 (2021), pp. 5586–5609 (cit. on pp. 77, 86).

[448] Fan Zhou et al. "Task similarity estimation through adversarial multitask neural network". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.2 (2020), pp. 466–480 (cit. on p. 182).

[449] Zhi-Hua Zhou. "Ensemble Methods: Foundations and Algorithms". In: (2012) (cit. on p. 217).

[450] Qiuming Zhu. "On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset". In: *Pattern Recognition Letters* 136 (2020), pp. 71–80 (cit. on pp. 54, 85).

[451] Fuzhen Zhuang et al. "A comprehensive survey on transfer learning". In: *Proceedings of the IEEE* 109.1 (2020), pp. 43–76 (cit. on pp. 78, 79, 87).

Disclosure

During the writing of this work I used the search engines *Google Scholar*, *Semantic Scholar* and *Perplexity AI* as research tools in order to identify related existing literature and assess their relevance. I further used the LLMs *Google Translate*, *Claude* and *DeepL Write* for linguistic and grammatical improvements. Any generated content by these services has been carefully reviewed and verified.

Lifelong Machine Learning
for Biomedical Image Classification
Ph. D. Thesis

Supervised by Prof. Dr. Lena Maier-Hein

This work has been set using LaTeX and KOMA-Script.