Inaugural-Dissertation

zur

Erlangung der Doktorwürde

der

Gesamtfakultät für Mathematik, Ingenieur- und Naturwissenschaften

der

Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Silvia Seidlitz, M. Sc.

aus Hof (Saale)

Tag der mündlichen Prüfung: _____

Robust AI-Driven Spectral Imaging for Perioperative Care

Supervisor: Prof. Dr. Lena Maier-Hein

Abstract

Physicians face major challenges in perioperative decision-making, as they need to rely on clinical intuition and limited information for critical real-time judgments. Spectral imaging (SI) could support this process by rapidly and non-invasively revealing changes in tissue composition that alter spectral signatures. While such changes often remain invisible to the human eye or conventional RGB imaging, SI captures subtle variations in tissue reflectance spectra at each pixel. Combined with machine learning (ML), this high-dimensional data could efficiently yield clinically relevant insights.

Numerous proof-of-concept studies have demonstrated the potential of SI, particularly for estimating functional tissue parameters such as oxygenation, thereby enabling non-invasive distinction between perfused and ischemic tissue during surgery. However, several important clinical applications of SI remain underexplored:

Clinical Gap: Automated Surgical Scene Segmentation Visual discrimination of tissue types remains an important challenge for surgeons, and automated surgical scene segmentation is a key component of surgical data science applications such as surgical phase recognition and robot-assisted surgery. However, SI-based segmentation, particularly in open surgeries, has received little attention. Consequently, it remains unclear whether SI offers advantages over other imaging modalities (e.g., RGB imaging) for surgical scene segmentation and how to optimally represent the input data in terms of spatial granularity (e.g., pixels, entire images). Leveraging the largest semantically annotated SI database to date, we close this gap and demonstrate that SI consistently outperforms RGB across all spatial granularities. Our image-based SI segmentation reaches performance comparable to a second human expert.

Clinical Gap: Sepsis Diagnosis and Mortality Prediction in Critically Ill Patients Sepsis remains a leading cause of mortality and critical illness. Early detection is vital to reduce mortality risk, yet reliable biomarkers for timely diagnosis and outcome prediction are still lacking. Sepsis diagnosis in intensive care unit (ICU) patients is particularly challenging due to high baseline illness severity. SI could potentially close this gap by capturing early signs such as edema formation and microcirculatory dysfunction. However, prior studies compare sepsis patients to healthy volunteers or narrowly selected cohorts, introducing a substantial risk of shortcut learning from confounding factors such as age and treatment regimens. We address this critical gap through a prospective study in ICU patients, comprising the largest SI patient cohort to

date, in which we diagnose sepsis and predict mortality on the day of admission. Our SI-based ML models achieve high accuracy, particularly when combined with minimal clinical data, and outperform widely used biomarkers and scores, while enabling rapid, non-invasive, cost-effective and mobile assessments.

Technical Gap: Investigation of Domain Shifts A key challenge for SI analysis is its clinical translation. Numerous studies outside medical SI have shown that domain shifts between training and real-world application data can severely degrade algorithm performance, yet this issue has received little attention in medical SI. We are the first to investigate the impact of important real-world domain shifts: Illuminant and hardware-related shifts in functional tissue parameter estimation, geometric shifts (e.g., situs occlusions) in surgical scene segmentation, and population shifts in sepsis diagnosis and mortality prediction. Our results show that such shifts can substantially degrade downstream task performance.

Technical Gap: Mitigating performance degradation under domain shifts We propose methods to mitigate the performance degradation under domain shifts and improve algorithm robustness. To address drops in functional tissue parameter estimation due to illuminant changes, we introduce the first intraoperative, live illuminant estimation approach. Our method outperforms state-of-the-art illuminant estimation techniques from nonmedical domains, achieving accuracy close to the ideal scenario of a perfectly known illuminant. Additionally, we provide recommendations to mitigate hardware-related bias in SI study design. To enable robust surgical scene segmentation under geometric domain shifts, we introduce a surgery-inspired data augmentation strategy which restores in-distribution performance across diverse out-of-distribution scenarios.

In conclusion, this thesis contributes substantial advancements towards the robust and reliable application of ML-based SI analysis in real-world clinical settings. Specifically, it enables, for the first time, (1) intraoperative functional tissue parameter estimation under illuminant and hardware-related shifts, (2) automated surgical scene segmentation under geometric domain shifts, and (3) automated sepsis diagnosis and mortality prediction among ICU patients. Our findings are supported by extensive validation studies which are among the largest in the field of medical SI to date. To support the research community and facilitate the clinical translation of SI, we have publicly released datasets[1], as well as our code and pretrained models[2].

[1] https://spectralverse-heidelberg.org/
[2] https://github.com/IMSY-DKFZ/htc

Zusammenfassung

Ärzt:innen stehen im perioperativen Entscheidungsprozess vor großen Herausforderungen, da sie auf klinische Intuition und begrenzte Informationen angewiesen sind, um kritische Entscheidungen in Echtzeit zu treffen. Spektrale Bildgebung (SI) könnte diesen Prozess unterstützen, indem sie schnell und nicht-invasiv Veränderungen in der Gewebezusammensetzung aufzeigt, die spektrale Signaturen beeinflussen. Während solche Veränderungen für das menschliche Auge oder herkömmliche RGB-Bildgebung oft unsichtbar bleiben, erfasst SI subtile Unterschiede in den Gewebereflexionsspektren auf Pixelebene. In Kombination mit maschinellem Lernen (ML) könnten diese hochdimensionalen Daten effizient klinisch relevante Informationen liefern.

Zahlreiche Machbarkeitsstudien haben das Potenzial von SI insbesondere zur Schätzung funktioneller Gewebeparameter wie der Oxygenierung gezeigt. Dies ermöglicht beispielsweise eine nicht-invasive Unterscheidung zwischen durchblutetem und ischämischem Gewebe während chirurgischer Eingriffe. Andere wichtige klinische Anwendungen von SI sind jedoch bislang nicht ausreichend erforscht:

Klinische Lücke: Automatisierte Segmentierung chirurgischer Szenen Die visuelle Differenzierung von Gewebetypen stellt für Chirurg:innen weiterhin eine zentrale Herausforderung dar. AuSSerdem ist die automatisierte Segmentierung chirurgischer Szenen ist ein Schlüsselelement zahlreicher Anwendungen, beispielsweise in der robotergestützten Chirurgie. Die Segmentierung mithilfe von SI wurde bislang kaum untersucht, insbesondere in offenen Operationen. Folglich ist unklar, ob SI gegenüber anderen Bildgebungsmodalitäten (z. B. RGB) Vorteile bei der Segmentierung chirurgischer Szenen liefert und wie die Eingangsdaten optimal hinsichtlich ihrer räumlichen Granularität (z. B. Pixel, ganze Bilder) verarbeitet werden sollten. Aufbauend auf der bisher größten semantisch annotierten SI-Datenbank schließen wir diese Lücke und zeigen, dass SI über alle räumlichen Granularitäten hinweg besser abschneidet als RGB. Unsere bildbasierte SI-Segmentierung erreicht eine zu einem menschlichen Experten vergleichbare Leistung.

Klinische Lücke: Sepsisdiagnose und Mortalitätsprognose bei Intensivpatienten Sepsis bleibt eine führende Ursache für Mortalität und kritische Erkrankungen. Eine frühzeitige Erkennung ist entscheidend, um das Sterberisiko zu senken, jedoch fehlen bislang verlässliche Biomarker für eine rechtzeitige Diagnose und Prognose. Bei Patienten auf der Intensivstation ist die Sepsisdiagnose aufgrund der hohen zugrundlie-

genden Krankheitsschwere besonders herausfordernd. SI könnte diese Lücke schließen, indem sie frühe Anzeichen wie Ödembildung und mikrozirkulatorische Dysfunktion erfasst. Frühere Studien verglichen Sepsispatienten jedoch lediglich mit gesunden Probanden oder selektiven Kohorten, was ein erhebliches Risiko von „Shortcut Learning“ durch Störfaktoren wie Alter- oder Behandlungsregimes birgt. Wir adressieren diese kritische Lücke durch eine prospektive Studie an Intensivpatienten, der bislang größten SI-Patientenkohorte, in der wir am Tag der Aufnahme auf die Intensivstation Sepsis diagnostizieren und die 30-Tage-Mortalität vorhersagen. Unsere SI-basierten ML-Modelle erreichen hohe Genauigkeit, insbesondere in Kombination mit wenigen klinischen Daten, und übertreffen weit verbreitete Biomarker und Scores. Gleichzeitig ermöglichen sie schnelle, nicht-invasive, kosteneffiziente und mobile Messungen.

Technische Lücke: Untersuchung von Domänenverschiebungen Die klinische Translation SI-basierter Algorithmen stellt eine zentrale Herausforderung dar. Zahlreiche Studien aus anderen Bereichen haben gezeigt, dass Domänenverschiebungen zwischen Trainings- und Einsatzdaten die Leistungsfähigkeit von Algorithmen erheblich beeinträchtigen können. In der medizinischen SI-Analyse wurde dieses Problem jedoch bislang kaum berücksichtigt. In unserer Arbeit untersuchen wir erstmals relevante Domänenverschiebungen und deren Auswirkungen, wie Beleuchtungs- und hardwarebedingter Variationen auf die Schätzung funktioneller Gewebeparameter, geometrischer Veränderungen (z. B. Situsverdeckungen) auf die Segmentierung chirurgischer Szenen sowie populationsbedingte Unterschiede auf Sepsisdiagnose und Mortalitätsprognose. Unsere Ergebnisse belegen, dass solche Domänenverschiebungen die Leistungsfähigkeit SI-basierter Algorithmen deutlich reduzieren können.

Technische Lücke: Adressierung von Domänenverschiebungen Wir schlagen Methoden vor, um Leistungseinbußen bei Domänenverschiebungen zu mindern. Um Ungenauigkeiten bei der Schätzung funktioneller Gewebeparameter aufgrund von Beleuchtungsänderungen zu beheben, präsentieren wir den ersten intraoperativen Ansatz zur automatisierten Beleuchtungsschätzung. Unsere Methode übertrifft den Stand der Technik aus nicht-medizinischen Bereichen und erreicht eine Genauigkeit, die dem Idealfall einer bekannten Beleuchtung nahekommt. Darüber hinaus geben wir Empfehlungen zur Vorbeugung hardwarebedingter Variationen im Design von SI Studien. Für die Segmentierung chirurgischer Szenen bei geometrischen Domänenverschiebungen führen wir eine von der Chirurgie inspirierte Strategie zur Datenaugmentierung ein, die zur Trainingsdomäne vergleichbare Leistungen erzielt.

Zusammenfassend leistet diese Arbeit bedeutende Fortschritte für die robuste und zuverlässige Anwendung von ML-basierter SI-Analyse in realen klinischen Umgebungen. Sie ermöglicht erstmals (1) die intraoperative Schätzung funktioneller Gewebeparameter unter beleuchtungs- und hardwarebedingten Veränderungen, (2) die automatisierte Segmentierung chirurgischer Szenen unter geometrischen Domänenverschiebungen

und (3) die automatisierte Sepsisdiagnose und Mortalitätsprognose bei Intensivpatienten. Unsere Ergebnisse werden durch umfangreiche Validierungsstudien gestützt, die zu den größten im Bereich der medizinischen SI zählen. Zur Unterstützung der Forschungsgemeinschaft und Förderung der klinischen Translation von SI haben wir Daten[3] sowie unseren Code und trainierte Modelle[4] öffentlich zugänglich gemacht.

[3] https://spectralverse-heidelberg.org/
[4] https://github.com/IMSY-DKFZ/htc

Acknowledgements

Countless people have shaped and supported my Ph.D. in different ways, and I would like to sincerely thank them all, even if I cannot name each one explicitly here.

First and foremost, I would like to thank my supervisor, **Lena Maier-Hein**. You are an inspiring role model for women in tech, full of knowledge, ideas, and impressive networking skills. Despite your demanding workload, you always made time for discussions, advice, and support, both on my research projects and beyond. You provided me with an outstanding environment to conduct my research and gave me the opportunity to grow personally by entrusting me with the leadership of a group of spectral imaging researchers – a responsibility for which I am immensely grateful.

I would also like to thank the entire **Intelligent Systems in Spectral Imaging (ISSI)** team and the wider **Intelligent Medical Systems (IMSY)** group for their mutual support, positive spirit and the countless enjoyable moments we shared along the way. In particular, there are two people I could not have imagined this journey without. **Leonardo Ayala** – thank you for bringing so much joy to the team as our "fun minister" and for easing my group leadership responsibilities during intense phases. Our many discussions on hardware and experimental challenges were invaluable, and your support helped me stay motivated despite bureaucratic hurdles. **Jan Sellner** – working with you on our collaborative projects was both productive and fun. Your teamspirit, sharp mind, impressive programming skills and wonderful presonality had a lasting impact on my personal development and success. You were our "Computer Science Yoda", always ready to tackle tricky bugs and provide thoughtful code reviews. Your enthusiasm for continuously improving our practices and increasing efficiency, such as enhancing our coding infrastructure, was invaluable not only to the entire IMSY group, but also beyond.

My deep gratitude goes to all my clinical **collaborators**, especially **Maximilian Dietrich** and **Markus Weigand** from anesthesia, and **Alexander Studier-Fischer** and **Felix Nickel** from visceral surgery. I feel incredibly fortunate to have worked with such motivated and dedicated physicians who were always willing to go the extra mile to ensure high-quality, structured datasets. Your input was essential for identifying clinical challenges, shaping my research to be clinically meaningful, and patiently answering medical questions, no matter the hour. Together, we curated over 69 747 images from 1000 subjects – a substantial advancement in medical spectral imaging,

especially considering that, when I began my Ph.D., most publications were based on only a handful of patients.

This would not have been possible without the tremendous efforts of many medical students. Thank you, **Katharina Hölzl**, **Ayca von Garell**, **Berkin Özdemir**, and **Marita Klein**, for your dedication to our studies despite the heavy time commitment and frequent weekend and night shifts. Deep learning research thrives on accurate annotations, and I am sincerely grateful to all our **medical annotators**. Special thanks go to **Janne Heinecke** and **Jule Brandt** for taking over the management of the annotation process, training new annotators, and investing countless hours into revising and ensuring the quality of annotations.

Beyond the core research projects covered in this thesis, I was fortunate to contribute to many other collaborations. While I cannot list all my collaborators by name, I am deeply thankful for every opportunity to broaden my expertise and take part in exciting research endeavors.

Research on this scale is only possible with the right infrastructure. I am grateful to our **IMSY secretary and scientific coordinators** for their excellent handling of administrative tasks and for shielding me from bureaucratic chaos. My thanks also go to the entire **DKFZ cluster administration team** for providing reliable high-performance computing infrastructure.

Finally, I would like to thank my **family and friends** for standing by me through all the highs and lows. Your unwavering support made it possible for me to accomplish my goals. I could not have done this without you – you are truly awesome.

Contents

Part I

Introduction

1 Clinical Motivation and Open Challenges

1.1 Motivation

Advancements in medical imaging technologies have revolutionized healthcare by offering detailed insights into the human body without the need for invasive procedures. Since Wilhelm Conrad Röntgen captured the first X-ray in 1895, modalities such as computed tomography, magnetic resonance imaging, positron emission tomography, ultrasound, and various others have become essential for tasks like disease diagnosis, treatment planning, and imaging-guided interventions [36, 2]. However, widely used imaging techniques primarily provide morphological information and are limited in their ability to monitor changes in tissue chemical composition and function. Monitoring these changes, which can result from pathological alterations, is essential for the early detection and identification of tissue abnormalities [205].

Changes in tissue biochemical composition are often imperceptible to the human eye and conventional imaging methods that mimic it. However, these changes do alter the tissue's spectral signature. Spectral imaging (SI), a technique that combines imaging with spectroscopy and was initially developed for remote sensing [121], enables non-invasive and quantitative measurement of tissue composition along with its spatial distribution by capturing detailed spectral information for each image pixel [65] (cf. Figure 1.1). SI holds substantial promise for medical diagnostics and imaging-guided interventions, and could enhance our understanding of disease-related metabolic processes [205, 380]. For example, retinal SI has shown promise in the diagnosis of Alzheimer's disease [207]. Additionally, several studies have highlighted the potential of SI in accurately identifying tumors and delineating their boundaries across various cancers, including colon cancer [229], breast cancer [255], brain tumors [202], and skin cancer [203]. By reducing the risk of residual tumor tissue, SI could lead to improved prognosis and survival outcomes [228]. Moreover, functional tissue parameters such as oxygen saturation, perfusion, and water content can be estimated from SI data [367,

141, 188]. Identifying perfused and ischemic tissue is critical in most surgical interventions, such as verifying the successful transplantation of organs [332], ensuring proper anastomosis of blood vessels [326], and confirming effective vessel clamping to prevent excessive bleeding [21]. While typically injection of contrast agents is required to identify perfused and ischemic tissue, SI could provide this information non-invasively and repeatedly [21]. Furthermore, SI-enabled functional imaging has demonstrated potential for monitoring treatment response in diabetic foot ulcers [251] and shock therapy [327], as well as for guiding the optimization of surgical techniques [248].

Despite numerous proof-of-concept studies showcasing the potential of SI across various medical applications, its role in perioperative care remains underexplored. For example, it is still unclear whether SI offers benefits for automated surgical scene segmentation compared to conventional RGB imaging and how to optimally represent the SI data for deep learning (DL)-based analysis. Likewise, the use of SI for rapid, non-invasive sepsis diagnosis and mortality prediction in the intensive care unit (ICU) has not yet been investigated, despite the urgent need for such a tool. Beyond these clinical gaps, technical challenges concerning the robustness and generalizability of SI-based algorithms continue to limit their clinical translation. This thesis addresses 3 key open challenges in advancing the clinical adoption of SI in perioperative care.

1.2 Open Challenges

First, this thesis investigates the accuracy of functional tissue parameter estimation under real-world imaging conditions, focusing on challenges such as dynamic illuminant shifts during open surgery and hardware-related spectral variability (Section 1.2.1, Part II). Second, it explores the potential of SI for automated surgical scene segmentation, addressing geometric domain shifts arising from factors like situs occlusions and organ resections commonly encountered in real-world surgeries (Section 1.2.2, Part III). Third, it pioneers the use of SI for rapid, non-invasive sepsis diagnosis and mortality prediction in an ICU population, with a particular emphasis on evaluating algorithmic generalizability under population shifts (Section 1.2.3, Part IV). A graphical overview of the clinical and technical challenges addressed in this work is provided in Figure 1.1, and the resulting core research questions are outlined below.

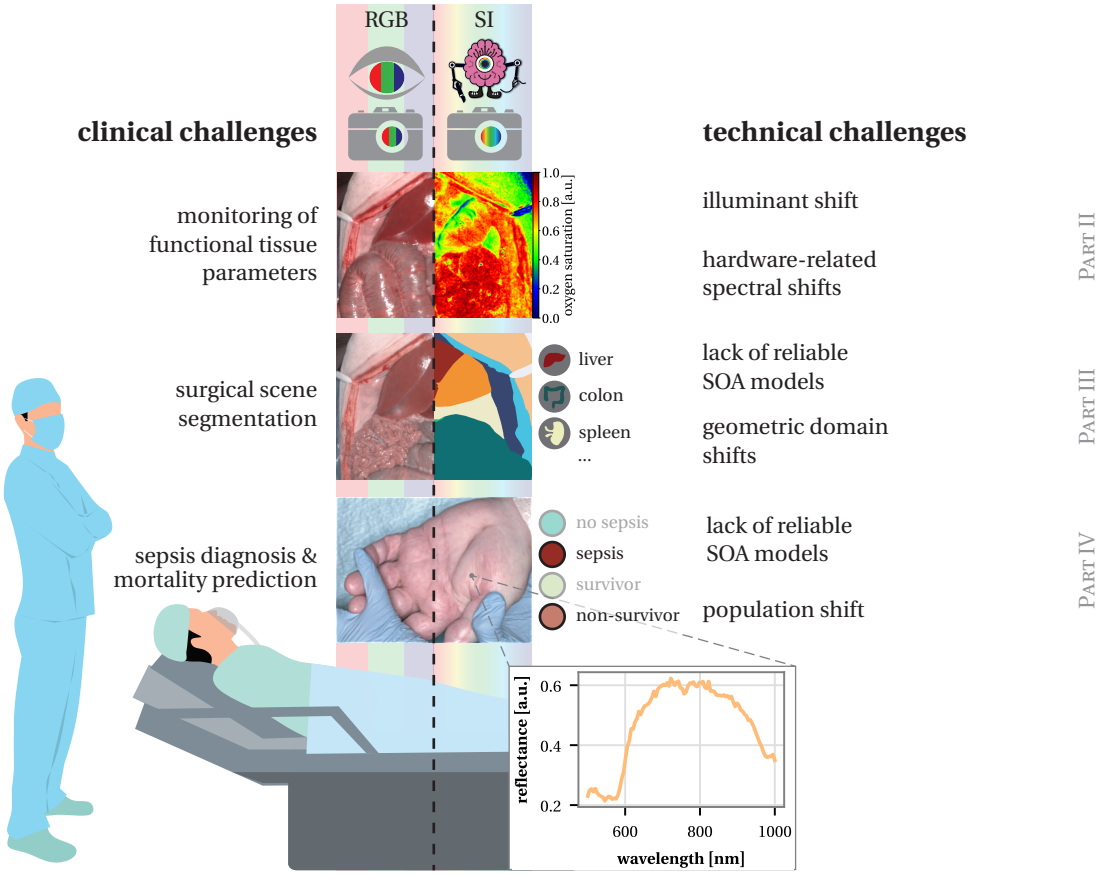


Figure 1.1: Potential benefits and challenges of spectral imaging (SI) in the perioperative workflow. Unlike conventional RGB imaging and human vision, SI captures detailed reflectance spectra for each pixel, illustrated here with an example skin pixel spectrum. This rich spectral information enables the estimation of functional tissue parameters, and holds potential for automated surgical scene segmentation, as well as rapid, non-invasive sepsis diagnosis and mortality prediction – critical challenges in perioperative care. However, clinical translation of SI faces several challenges. This thesis tackles inaccuracies in functional tissue parameter estimation caused by dynamic illuminant shifts during open surgery and hardware-related spectral variations (Part II). It further explores optimal SI data representations for automated surgical scene segmentation and addresses geometric domain shifts (Part III). Finally, it pioneers automated sepsis diagnosis and mortality prediction, evaluating the generalizability of algorithms trained on selectively chosen cohorts to a clinically more relevant intensive care unit population (Part IV).

1.2.1 RQ1: How can we achieve robust functional tissue parameter estimation with spectral imaging under real-world imaging conditions?

As outlined above, non-invasive, continuous, and quantitative functional parameter estimation from SI data holds substantial promise for a range of perioperative applications, including organ transplantation [332], anastomosis assessment [326], partial nephrectomy [21], the optimization of surgical techniques [248], as well as therapy monitoring [83, 327]. Realizing this potential in clinical practice, however, depends on obtaining accurate parameter estimates under real-world imaging conditions. As illustrated in Figure 1.2, this thesis focuses on two major technical challenges to achieving this goal: dynamic illuminant shifts during open surgery and hardware-induced spectral variability.

Illuminant Shifts During Open Surgery SI captures the intensity of light reflected from tissue across various spectral bands. The resulting spectra are influenced not only by the tissue's chemical composition and functional states but also by the illuminant spectrum (cf. Section 2.1 for a detailed review of spectral image formation). Therefore, accurate SI analysis requires precise knowledge of the illuminant spectrum to extract meaningful reflectance data and accurately estimate functional tissue parameters. This can be achieved either by controlling the lighting conditions to ensure that only a light source with a known spectrum illuminates the tissue or by capturing a reference image of a known reflectance standard under the same illumination conditions as the tissue image.

However, both approaches present challenges in clinical settings. In open surgeries, for instance, the surgical site is typically illuminated by multiple light sources, such as overhead lights, ceiling lights, and head torches, all of which may be moved during the procedure. This creates a dynamic and complex combined illuminant spectrum at the surgical site, which can change substantially over time. Requiring all lights except the known light source to be turned off during image acquisition is impractical in real-world open surgeries, as it would severely disrupt the surgical workflow and pose risks to patient safety. Similarly, calibrating the SI system with a reflectance standard is not feasible, as the standard would need to be placed in the surgical field, which is impossible due to sterility requirements.

To overcome this substantial barrier in the clinical translation of SI, Chapter 3 investigates the robust estimation of functional tissue parameters under real-world imaging conditions. It includes an analysis of how real-world illumination impacts the accuracy of functional parameter estimation, and introduces the first approach for live illuminant estimation that does not disrupt the surgical workflow. This work lays the

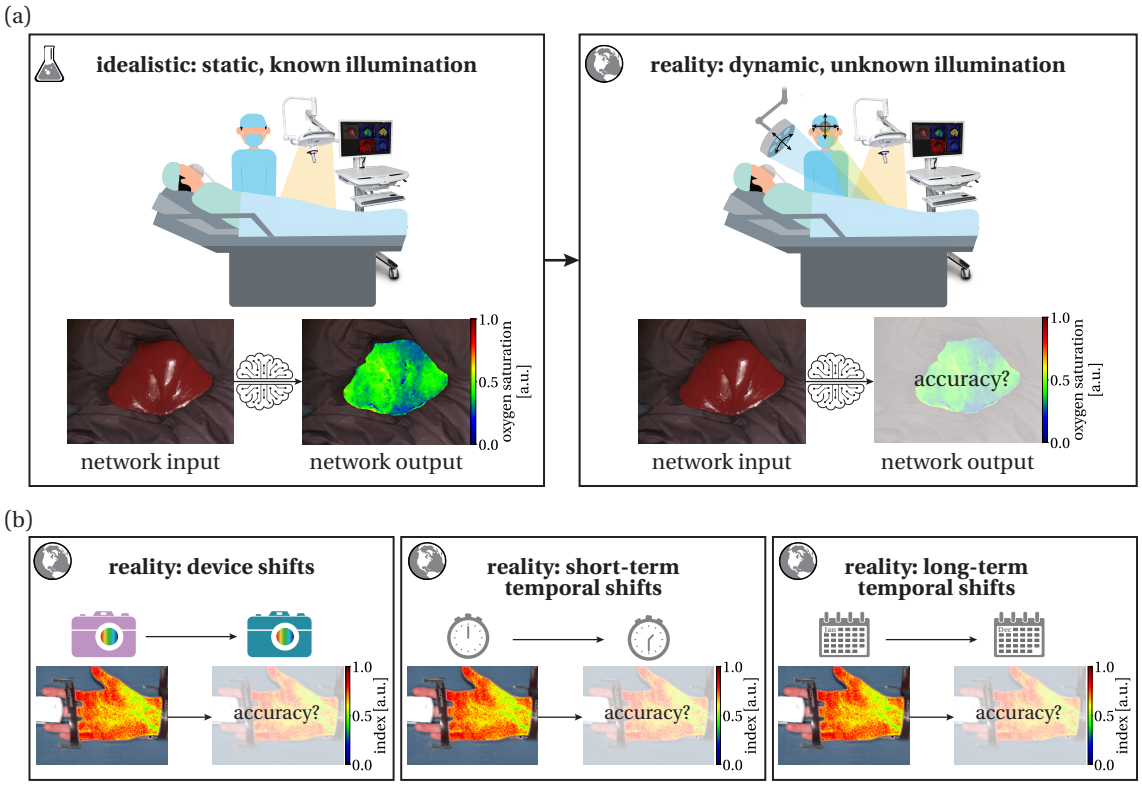


Figure 1.2: Research Question 1 (RQ1) investigates how to achieve accurate functional tissue parameter estimation from spectral imaging (SI) data under real-world imaging conditions. (a) In open surgeries, the illuminant spectrum shifts dynamically due to multiple light sources being switched on and off, or repositioned. Since conventional illuminant calibration methods cannot be safely integrated into the surgical workflow, this thesis analyzes how such real-world illumination shifts affect parameter accuracy and introduces the first approach for automated illuminant estimation. (b) Despite the growing adoption of SI devices, particularly the medically certified Tivita® systems (Diaspective Vision, Am Salzhaff, Germany), the impact of hardware-related spectral variability on the accuracy of functional parameter index estimation has not yet been studied. This thesis addresses this gap by analyzing spectral shifts across device generations and between devices of the same generation, as well as temporal variability over short-term (minutes to hours) and long-term (months to years) periods, and provides recommendations for a hardware bias-aware design of SI studies.

foundation for safely integrating functional parameter estimation with SI into clinical practice.

Hardware-Related Spectral Variability Beyond environmental influences such as illuminant shifts, variability in the SI hardware itself could affect the captured spectra and, consequently, the accuracy of functional tissue parameter estimation. This issue is particularly relevant given the growing body of studies using SI devices, most notably the medically certified Tivita® systems (Diaspective Vision, Am Salzhaff, Germany). These systems are widely used for assessing functional tissue parameter index images, involving different device instances and generations. Furthermore, many studies compare data acquired at different timepoints that can be up to months or even years apart (cf. Section 4.1). Such studies implicitly assume stability of measurements over time and comparability across devices, i.e., that neither short-term shifts (minutes to hours) nor long-term shifts (months to years), nor differences between devices or generations, affect measurement consistency. However, the extent of such hardware-related spectral variability and its impact on the accuracy of functional parameter estimation has not yet been investigated.

Chapter 4 closes this critical gap by presenting the first systematic analysis of hardware-related sources of variation in SI measurements using Tivita® systems and by quantifying their effect on functional tissue parameter indices. Based on these findings, it introduces a set of recommendations for mitigating hardware-related variation in SI study design. Adhering to these guidelines not only improves the reliability of functional tissue parameter estimation but also supports unbiased data acquisition, thereby laying a foundation for clinically robust SI applications beyond functional parameter estimation – such as automated surgical scene segmentation, as well as sepsis diagnosis and mortality prediction.

1.2.2 RQ2: How can we achieve robust surgical scene segmentation under geometric domain shifts?

Semantic segmentation of surgical scenes is a crucial foundation for numerous surgical data science applications, such as surgical phase recognition, and surgical robotics. Through an automated delineation of organs, tissues, and instruments in the surgical field, semantic segmentation enables intraoperative decision support and context-aware assistance, thereby improving the quality of interventional medicine [223, 221, 308].

While the state of the art in surgical scene segmentation has primarily focused on analyzing conventional RGB video data from minimally invasive surgeries [302, 125], and on binary segmentation tasks such as instrument segmentation [294], full semantic

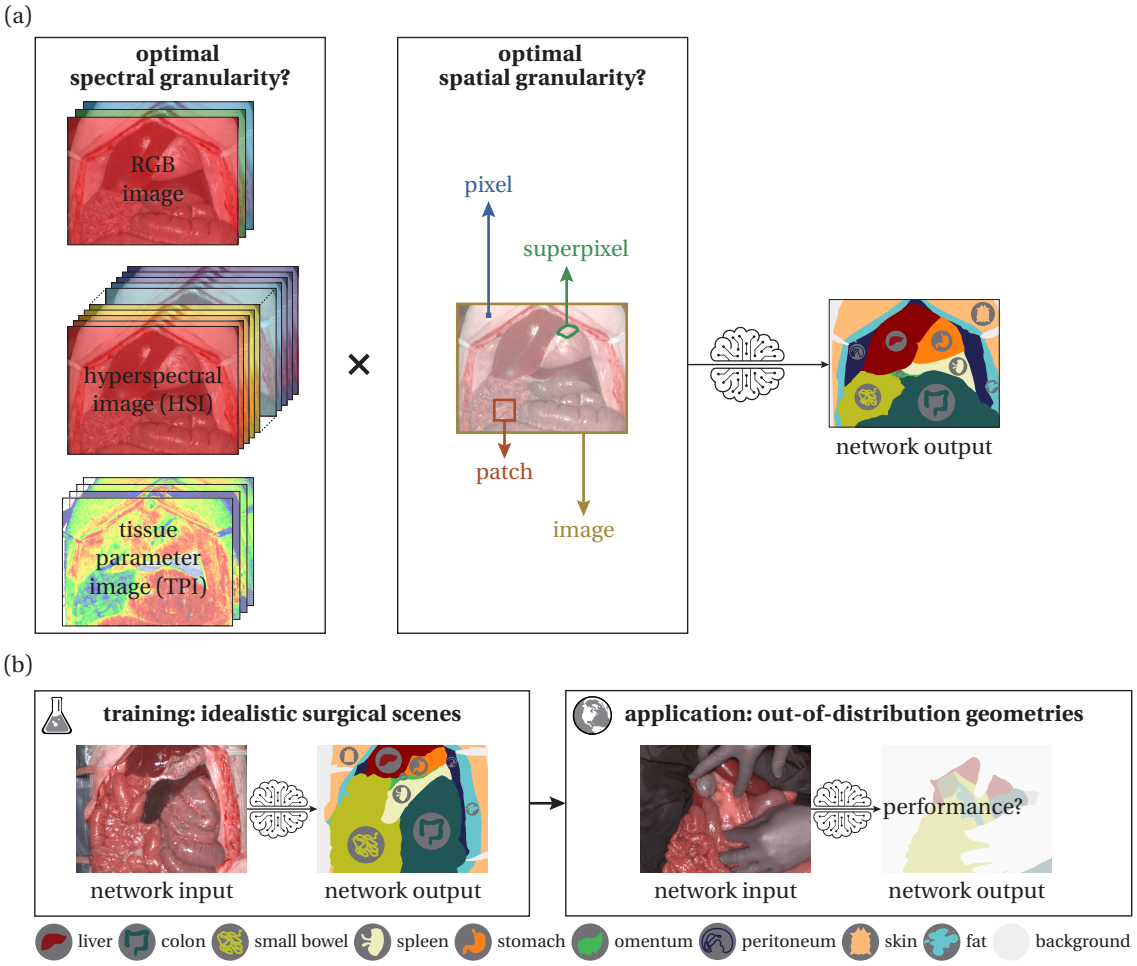


Figure 1.3: Research Question 2 (RQ2) investigates robust surgical scene segmentation under geometric domain shifts. (a) Surgical scene segmentation from spectral imaging (SI) data remains underexplored, with the optimal spectral granularity (conventional RGB imaging, full SI data or derived tissue parameter images (TPI)) and spatial granularity (pixels, superpixels, patches or full images) yet to be established. This thesis provides the first systematic analysis of these factors for deep learning-based surgical scene segmentation. (b) Real-world surgeries involve geometric domain shifts, such as situs occlusions and organ resections. This thesis offers the first evaluation of segmentation model performance under such shifts and introduces a novel approach to improve generalizability.

scene segmentation using SI data from open surgeries has received little attention. Consequently, it has not yet been determined whether SI data offers advantages over other modalities like RGB data or processed spectral data (e.g., functional tissue parameters) and how to most effectively represent SI data for DL-based segmentation (cf. Figure 1.3). These gaps in literature are addressed in Chapter 5.

Artificial intelligence (AI), and DL in particular, has driven substantial advancements in various disciplines, as demonstrated by recent breakthroughs in text-to-image generation [278], image synthesis and style transfer [290], and the success of large language models [348, 386, 274]. AI has also revolutionized the biomedical field, achieving important milestones in protein structure prediction [164], drug discovery [282], personalized medicine [116], and medical image analysis [206].

Despite the impressive performance of DL models on in-distribution data, generalization to real-world application data remains an important challenge. Numerous examples demonstrate that model accuracy can drop substantially due to shortcut learning and failing to extrapolate on out-of-distribution (OOD) data [288, 370, 378, 241]. This issue is particularly critical in healthcare, where understanding when algorithms might fail is essential to prevent patient harm. Nevertheless, the topic of generalizability remains largely underexplored in surgical scene segmentation.

To close this important gap, Chapter 6 presents the first investigation of the generalizability of surgical scene segmentation models under geometric domain shifts, such as situs occlusions and organ resections – common challenges in real-world surgeries. It introduces a novel approach to tackling these domain shifts, thereby advancing the clinical translation of DL-based surgical scene segmentation.

1.2.3 RQ3: Can we reliably diagnose sepsis and predict mortality in an intensive care unit population using skin spectral images?

Sepsis is a life-threatening syndrome that arises when a dysregulated host response to an infection causes organ dysfunction [320]. It is a leading cause of mortality, with an estimated 48.9 million cases resulting in 11 million deaths in 2017, representing 19.7 % of all global deaths [296]. A major challenge persists in the early and accurate diagnosis of sepsis, as the progression of irreversible organ damage increases the mortality risk with each hour of delayed treatment [97]. Conversely, the unnecessary treatment of patients misdiagnosed with sepsis using antibiotics contributes to the global surge in antibiotic resistance [350]. As the clinical diagnosis of sepsis typically relies on detecting organ dysfunction, sepsis is often recognized only in its more advanced stages [320]. The nonspecific symptoms and signs of the sepsis syndrome, as well as its complex, heterogeneous, and still not fully understood pathophysiology, further complicate an early diagnosis [136].

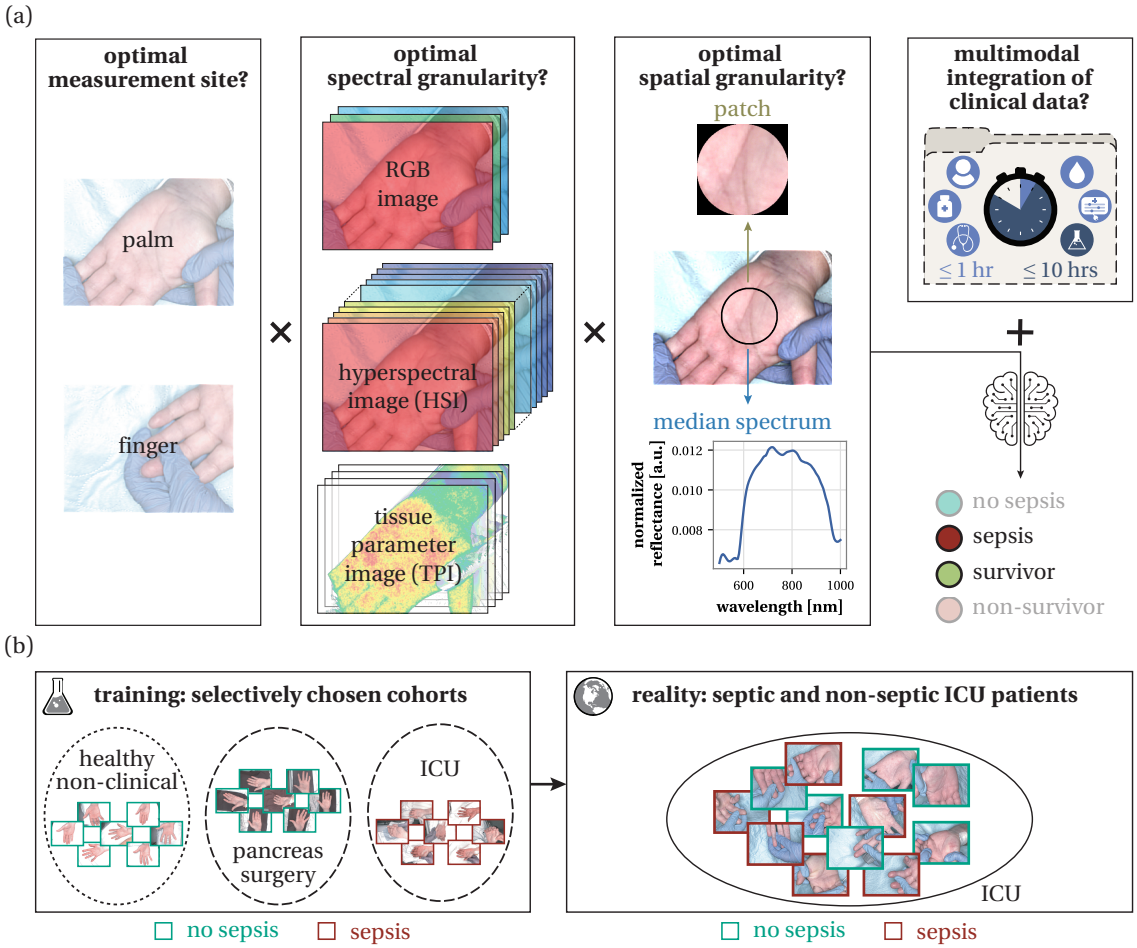


Figure 1.4: Research Question 3 (RQ3) investigates skin hyperspectral imaging (HSI) for rapid, non-invasive sepsis diagnosis and mortality prediction in intensive care unit (ICU) patients. (a) To date, automated sepsis diagnosis and mortality prediction from HSI has not been studied in ICU patients. This thesis addresses this gap by systematically evaluating the optimal measurement site (palm or finger), spectral granularity (RGB, full HSI, or derived tissue parameter images (TPI)), and spatial granularity (median spectra or patches). In light of recent advances in sepsis and mortality prediction from clinical data, the analysis further compares HSI-based models against clinical data baselines and examines the added value of multimodal integration. (b) Since previous studies compared septic patients with selectively chosen cohorts (e.g., healthy volunteers or patients undergoing pancreatic surgery), the resulting models are at risk of shortcut learning. This thesis therefore investigates their generalizability on a clinically relevant ICU population.

In addition to the early identification of septic patients, accurately and swiftly recognizing those at high risk of mortality is crucial. Timely identification enables the prompt implementation of appropriate interventions, which can substantially enhance patient outcomes. It also contributes to the overall efficiency and effectiveness of patient care by optimizing resource allocation, informing decisions about palliative care, and offering greater insight into the factors influencing patient outcomes [155, 190].

Despite extensive research suggesting over 250 potential diagnostic or prognostic biomarkers for sepsis and mortality, no robust and reliable biomarker has yet been identified [265]. Consequently, there is a critical need for rapid, non-invasive diagnostic tools that can accurately identify sepsis and predict patient mortality. In ICU patients, the diagnosis of sepsis is particularly challenging due to disease complexity, high baseline illness severity, and the difficulty of distinguishing sepsis from non-infectious systemic inflammation [42, 212].

Building on the hypothesis that hyperspectral imaging (HSI) enables automated sepsis diagnosis and mortality prediction by capturing early pathophysiological changes such as microcirculatory dysfunction and edema formation, Chapter 7 presents the first analysis of SI for rapid, non-invasive sepsis diagnosis and mortality prediction in ICU patients. As illustrated in Figure 1.4, the study investigates the optimal measurement site (palm or finger), spectral granularity (conventional RGB imaging, full SI data, or derived tissue parameter images (TPI) data), and spatial granularity (median spectra or patches) for DL-based prediction. Furthermore, given recent advances in sepsis and mortality prediction from high-dimensional clinical data, the analysis also compares SI-based models to clinical data baselines and explores the added value of combining SI with clinical data.

Previous HSI-based studies compared septic patients with selectively chosen cohorts, such as healthy volunteers or patients undergoing pancreatic surgery (cf. Section 7.1). Such designs are prone to shortcut learning, as potential confounders like age, comorbidity and therapy regimens may bias the models [85], thereby limiting their generalizability to unseen data [113]. To address this risk, Chapter 7 investigates the generalizability of such algorithms to a clinically relevant ICU population.

1.3 Outline

This thesis is structured into 5 parts. Part I provides an introduction by outlining the motivation behind the research, formulating the research questions, and providing an outline in the present Chapter 1. It also presents the necessary background in medicine, biophotonics, and machine learning in Chapter 2. Part II, Part III and Part IV detail the research conducted to address the respective research questions RQ1, RQ2, and RQ3 (cf. Section 1.2). These parts are organized into chapters, each focusing on a specific set of

sub-research questions and comprising a related work section, a materials and methods section, an experiments and results section, and a discussion and conclusion section. This thesis closes with a high-level summary of the main findings and contributions, followed by an outlook on remaining open questions in Part V.

Fundamentals 2

2.1 Biophotonics Background

Biophotonics, which exploits the interaction of light with biological tissue for advanced imaging, diagnostics and therapy, is a rapidly evolving field poised to significantly impact current and future healthcare [253]. The basic functional principle of biophotonics imaging is that light-tissue interaction depends on the tissue composition. Since changes in light-tissue interaction change the spectral signature of the tissue, changes in tissue composition can be measured [40]. While such changes are imperceptible to the human eye, SI techniques exploit this principle to monitor and measure tissue function and chemical composition by capturing tissue reflectance. The fundamentals of light-tissue interaction are presented in Section 2.1.1. The subsequent sections focus on SI, the biophotonics imaging method employed in this work. These sections cover SI hardware (Section 2.1.2), processing techniques (Section 2.1.3), and methods for estimating functional tissue parameters from SI data (Section 2.1.4).

2.1.1 Light-Tissue Interaction

The interaction of light with biological tissue encompasses several physical processes, as illustrated in Figure 2.1 for light interaction in human skin tissue. First, light from a light source strikes the biological tissue. The light may either be reflected at the tissue surface, referred to as *specular reflection*, or undergoes refraction, scattering, and absorption within the tissue. The likelihood of scattering and absorption events depends on both the tissue composition and the wavelength of the light. Some of the incident light eventually emerges from the tissue surface upon multiple scattering events, referred to as *diffuse reflection*. SI devices capture this diffusely reflected light to extract the encoded information about tissue composition and function.

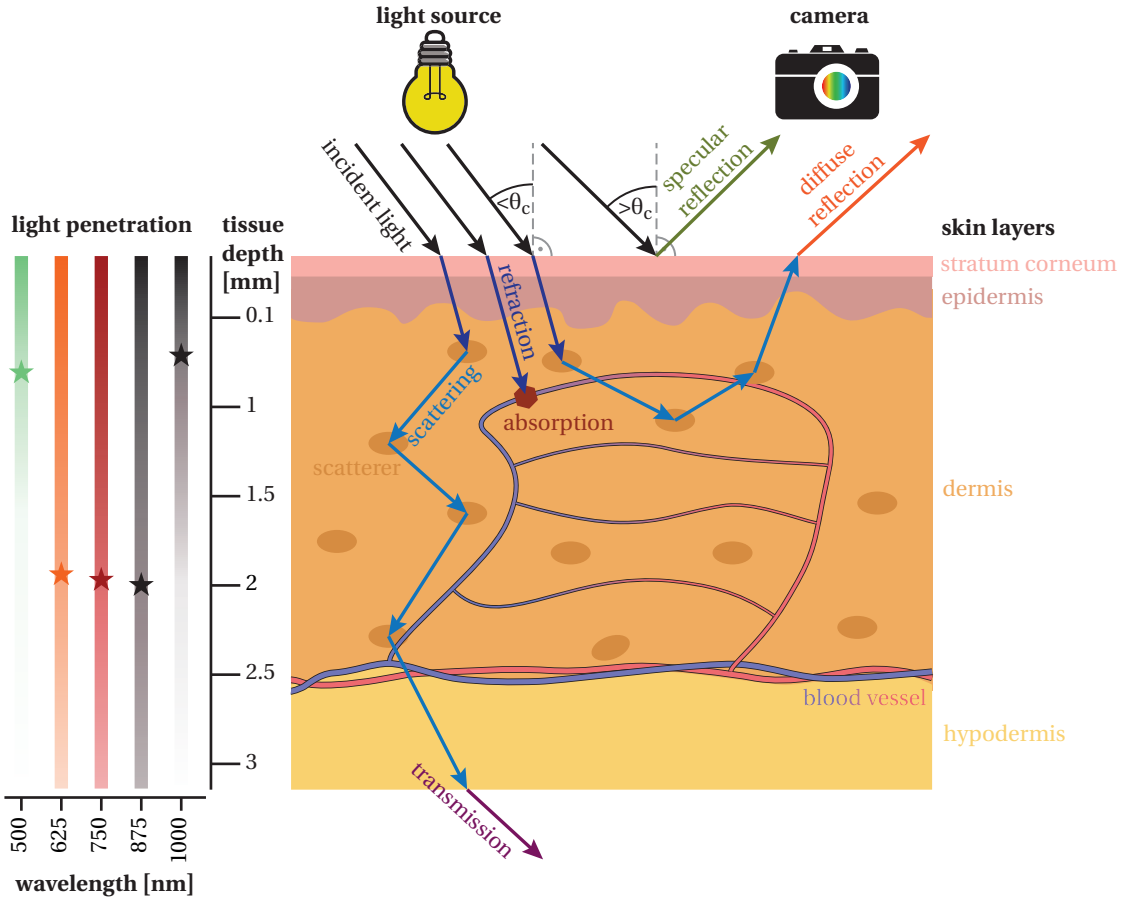


Figure 2.1: Schematic overview of light-tissue interactions in biological tissue. When light interacts with human tissue (here: skin model, according to [264]), it can either undergo specular reflection at the surface or experience refraction, scattering, and absorption within the tissue. The likelihood of scattering and absorption events depends on both the wavelength and tissue composition. Consequently, the diffusely reflected light, captured by the spectral imaging device, carries valuable information about the tissue's structure and composition. Additionally, the depth of light penetration in biological tissue depends on the wavelength, with the depth where light intensity for a certain wavelength is halved denoted by stars. For visible and near-infrared (NIR) light, this depth ranges from several hundred micrometers to a few millimeters. Therefore, diffusely reflected light only carries information about superficial tissue layers. Figure inspired from [311].

Specular Reflection and Refraction When light encounters the boundary between two media with different refractive indices, such as air and biological tissue, its behavior depends on the angle of incidence, θ, relative to the surface normal (cf. Figure 2.1). If θ exceeds the critical angle, θ_c, total internal reflection can occur, also called specular reflection. Specular reflectance, which describes the ratio of specularly reflected photons to incident photons, is approximately wavelength-independent, as the wavelength dependence of the refractive index is minimal [160]. Thus, aside from a multiplicative factor, the spectrum of specularly reflected light is nearly identical to that of the incident light source. If θ is below θ_c, the light is refracted into the tissue. The refracted light can undergo scattering and absorption inside the tissue.

Absorption Absorption refers to the transfer of photon energy to the tissue, leading to the destruction of the photon. The primary absorbers in biological tissues include the chromophores deoxyhemoglobin (dHb), oxyhemoglobin (HbO_2), water, lipid, and melanin, with the latter predominantly found in skin and eye tissues [343, 157]. Their absorption coefficients μ_a, which quantify on average how far a photon can travel in a medium with a given chromophore concentration before being absorbed, are depicted in Figure 2.2[1]. Key to functional parameter estimation using SI is the fact that absorption coefficients vary between different chromophores and are also wavelength-dependent. Additionally, the total absorption coefficient in biological tissues is a linear combination of the individual absorption coefficients, weighted by the concentrations of the respective chromophores [358]. Consequently, diffuse reflectance spectra, which carry information about tissue absorption, can be utilized to estimate the concentrations c of tissue chromophores, such as dHb (c_{dHb}) and HbO_2 (c_{HbO2}), as well as functional tissue parameters like tissue oxygen saturation (StO_2):

$$\mathrm{StO2} = \frac{c_{\mathrm{HbO2}}}{c_{\mathrm{dHb}} + c_{\mathrm{HbO2}}} \tag{2.1}$$

Scattering Scattering refers to the alteration of a photon's trajectory and/or wavelength. Biological tissues are highly scattering, with the primary scatterers being cellular structures such as membranes, nuclei, lysosomes, mitochondria, and entire cells [243, 48]. Elastic scattering occurs when a photon excites a molecule into a transient virtual state, which subsequently re-emits the photon in a different direction upon relaxation, without changing the photon's energy [358]. This process, together with absorption, dominates light-tissue interactions in biological tissues [48]. In contrast, inelastic scattering (e.g., Raman scattering), where the photon energy changes, occurs with a probability several orders of magnitude smaller than that of elastic scattering [48]. The

[1] More specifically, the absorption coefficients measure the likelihood of a photon to be absorbed per unit path length travelled.

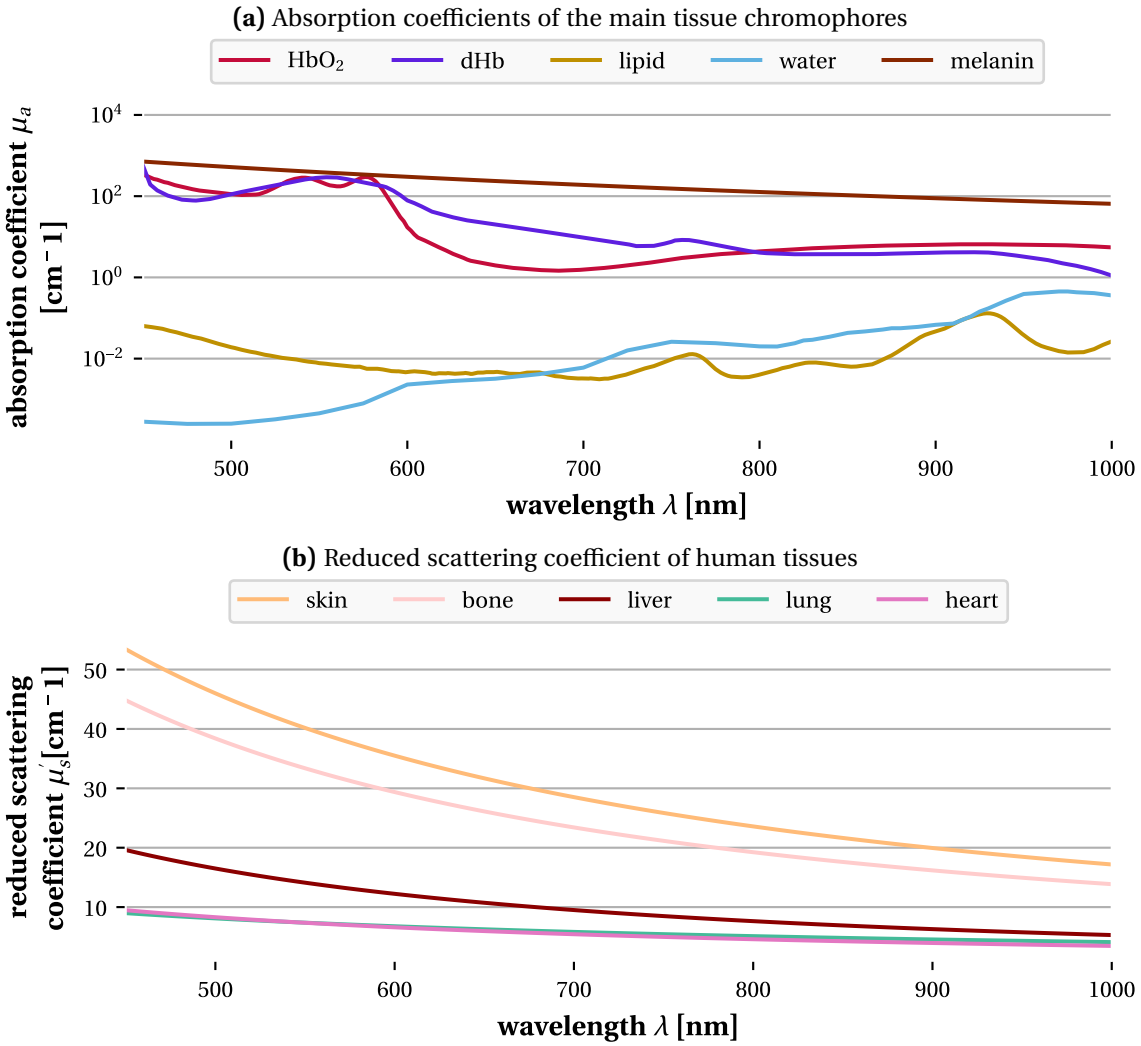


Figure 2.2: Absorption and scattering in biological tissues. (a) Absorption coefficients are shown for key chromophores in human tissue, within the 450–1000 nm spectral range, including deoxyhemoglobin (dHb), oxyhemoglobin (HbO_2), lipids, water, and melanin. The underlying data was sourced from the website [269], based on a comprehensive body of experimental research summarized in [157]. For dHb and HbO_2, the absorption coefficients were calculated from their molar extinction coefficients, assuming a typical blood concentration of 150 g/L. (b) The reduced scattering coefficient of human tissues is depicted, using averaged values compiled from multiple studies as presented in [157].

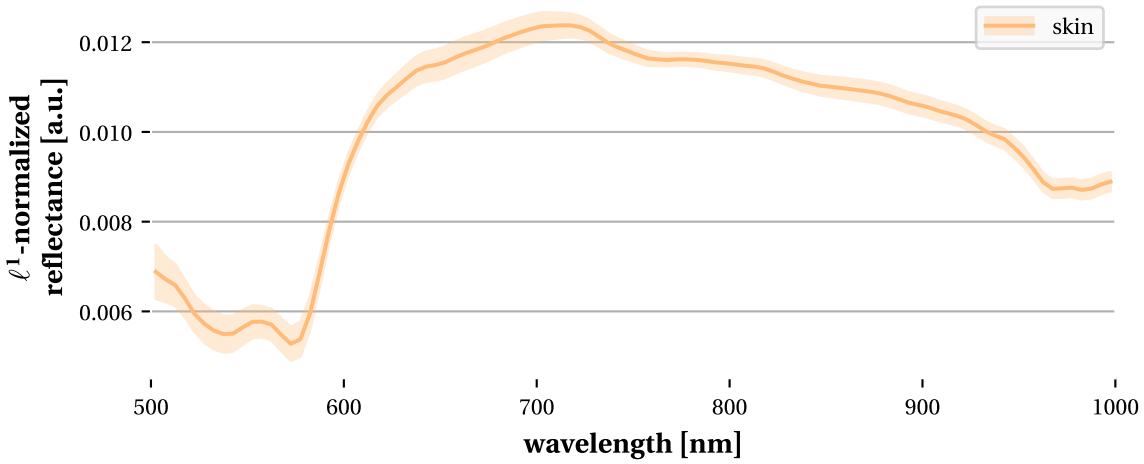


Figure 2.3: Human skin reflectance spectrum. The ℓ^1-normalized reflectance spectrum of skin from healthy volunteers is displayed, with the mean spectrum across 25 individuals shown as solid line and standard deviation represented by a shaded area. Details of the underlying dataset are provided in Section 7.2.2.

anisotropy factor, denoted as g, represents the expected value of the deflection angle α of a photon from its initial trajectory upon scattering:

$$g = \langle \cos(\alpha) \rangle \tag{2.2}$$

In biological tissue, g typically approximates 0.9, indicating a high degree of forward scattering [376]. The likelihood of photon scattering per unit path length is described by the scattering coefficient, μ_s, while the reduced scattering coefficient μ_s' further accounts for the direction of scattering:

$$\mu_s' = \mu_s \cdot (1 - g) \tag{2.3}$$

As illustrated in Figure 2.2, the reduced scattering coefficient is influenced by both the wavelength of light and the quantity and distribution of scatterers within the tissue, which differ across tissue types.

Multiple scattering events can result in a photon exiting the tissue from the same side as the incident illumination, contributing to diffuse reflection, or from the opposite side, contributing to transmission. In SI, it is the diffusely reflected light that is typically captured by the imaging device. It is used to determine the tissue *diffuse reflectance*, measuring the ratio of diffusely backscattered photons to incident photons. An exemplary diffuse reflectance spectrum for skin is shown in Figure 2.3.

The average number of photons penetrating to a given tissue depth decreases progressively due to absorption and scattering, a process called *attenuation*. Since absorption

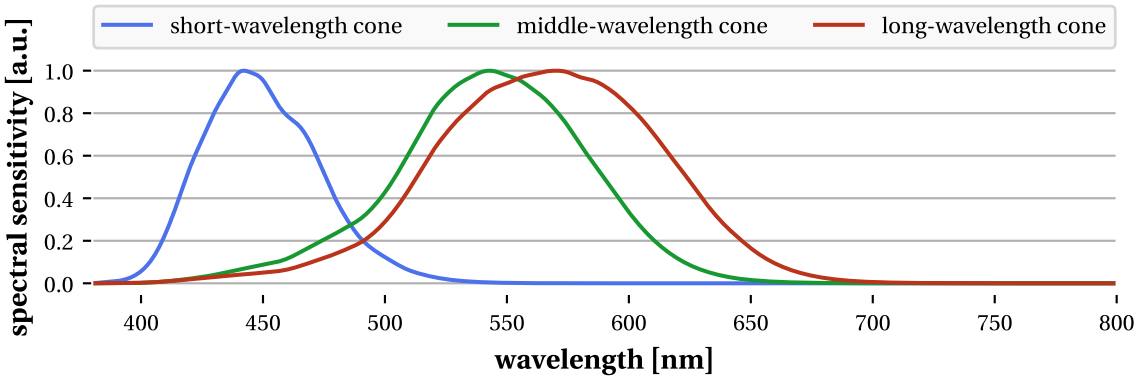


Figure 2.4: Human vision spectral sensitivity. Human vision is limited by the presence of 3 types of cone cells, the short-, middle- and long-wavelength cones. Their spectral sensitivities are broad, peaking around 420 nm, 530 nm, and 560 nm, respectively. Spectral sensitivity data for human cone cells is sourced from [325].

and scattering depend on both tissue composition and photon wavelength, the light penetration depth is also wavelength-dependent. In skin, the penetration depth at which the intensity of the incident light is halved ranges from several hundred micrometers to a few millimeters for visible and NIR light, with light in the wavelength range 600–930 nm penetrating deepest [20, 101]. Consequently, diffusely reflected light primarily provides information about superficial tissue layers, making SI particularly useful for analyzing tissue surfaces, such as skin or the exposed surface of internal organs during surgery.

2.1.2 Spectral Imaging Hardware

SI devices capture the diffusely reflected light of biological tissue across multiple wavelengths, typically spanning the visible to NIR range [270]. Unlike traditional spectroscopy, which measures the reflectance spectrum at a single point, these devices capture spectral data for every pixel in an image, producing 3-dimensional imaging cubes with two spatial and one spectral dimension. This allows for the analysis of spatial variations in tissue reflectance. Compared to human vision and conventional RGB imaging, which mimics human vision (see Figure 2.4 for the spectral sensitivity of human vision), SI devices record the reflected light in a greater number of narrower spectral bands, often extending beyond visible light. Depending on the number and bandwidth of these spectral channels, the imaging modality is either referred to as multispectral imaging (MSI) or HSI. MSI captures up to tens of relatively broad, non-contiguous spectral bands, whereas HSI captures up to hundreds of narrow spectral bands [65].

There are several approaches to MSI and HSI, which can be broadly divided into two categories: scanning techniques, which capture data by sequentially scanning either the spatial or spectral dimension, and snapshot techniques, which acquire both spectral and spatial information simultaneously. Figure 2.5 illustrates the functional principles of cameras from both categories that are used in this thesis.

The medical device-graded HSI systems TIVITA® Tissue and TIVITA® Surgery (Diaspective Vision GmbH, Am Salzhaff, Germany), utilized in this thesis, employ a push-broom scanning technique. In push-broom scanning, a broad-spectrum light source illuminates the entire field of view across the full wavelength range. As depicted in Figure 2.5, an entrance slit blocks all reflected light except for a single line of the image, which is then spectrally dispersed onto the camera sensor. By moving the slit across the scene, the system captures the complete HSI cube line by line.

This technique provides detailed spectral information, with the TIVITA® systems capturing 100 spectral channels across the visible and NIR range from 500–1000 nm, each with a bandwidth of approximately 5 nm (cf. Figure 2.6 for spectral sensitivities of the channels) [141, 188]. However, it suffers from poor temporal resolution: capturing a single HSI cube of dimensions $640 \times 480 \times 100$ (width $\times$ height $\times$ number of spectral channels) takes an acquisition time of approximately 7 s. This leads to several limitations in clinical settings. First, motion artifacts from patient movement, such as breathing, can distort the images, and the system is not suitable for handheld use. Second, highly dynamic processes, such as rapid perfusion changes, cannot be adequately monitored at low temporal resolution [360].

Snapshot imaging offers an alternative technique for video-rate SI, capturing the entire MSI cube in a single exposure. This is achieved using mosaic sensors, where a repeating pattern of bandpass filter arrays is placed over the sensor, with each sensor pixel assigned to a specific bandpass filter (cf. Figure 2.5). Since each pixel in the array captures a different spectral channel, the simultaneous acquisition of all spectral channels is possible. The recorded data is then *demosaiced* to reconstruct the MSI cube by stacking the pixel values corresponding to the same filter array. The MSI snapshot camera MQ022HG-IM-SM4x4-VIS (XIMEA GmbH, Münster, Germany) used in this thesis captures 16 spectral channels at a rate of 25 Hz [25]. While this technique supports high-speed imaging, it comes with trade-offs in both spectral and spatial resolution. The spectral resolution is reduced, with fewer channels and broader spectral bandwidths compared to HSI devices, as illustrated in Figure 2.7. Most of the spectral channels exhibit secondary peaks in spectral sensitivity due to second-order interferences in the bandpass filters[2]. Additionally, the spatial resolution is lower, generating MSI cubes with spatial dimensions of $272\,\text{px} \times 512\,\text{px}$.

For a comprehensive review of additional imaging techniques, see [65, 73].

[2]This problem has been solved in newer generations of the camera.

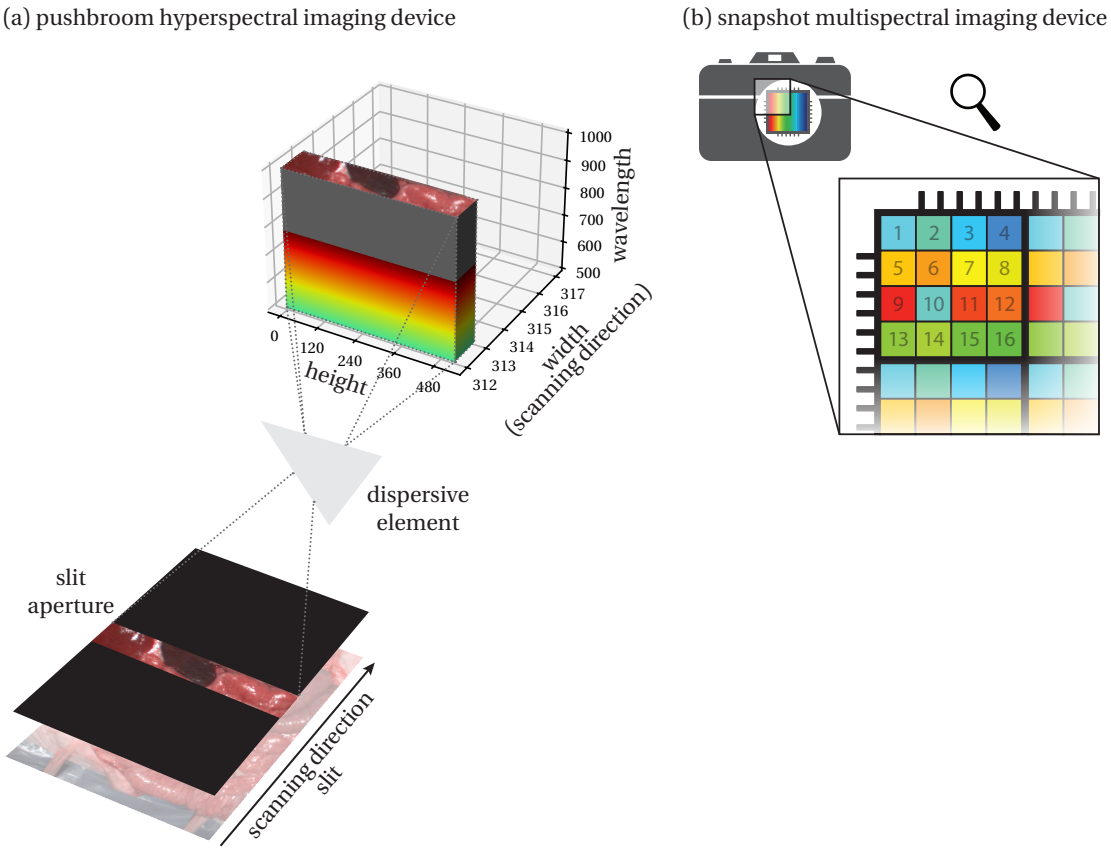


Figure 2.5: Functional principles of scanning and snapshot spectral imaging devices. (a) In the push-broom scanning device TIVITA® (Diaspective Vision GmbH, Am Salzhaff, Germany) used in our work, a slit aperture selects a line of the image, which is then spectrally dispersed onto a camera sensor. As the slit moves across the scene, the hyperspectral imaging cube is gradually built. (b) In the snapshot device MQ022HG-IM-SM4x4-VIS (XIMEA GmbH, Münster, Germany) used in our work, spectral data is captured simultaneously for all pixels in the image. This is achieved through a mosaic sensor, where each pixel is subdivided into an array of subpixels, each with its own bandpass filter. This subfigure is adapted from [21].

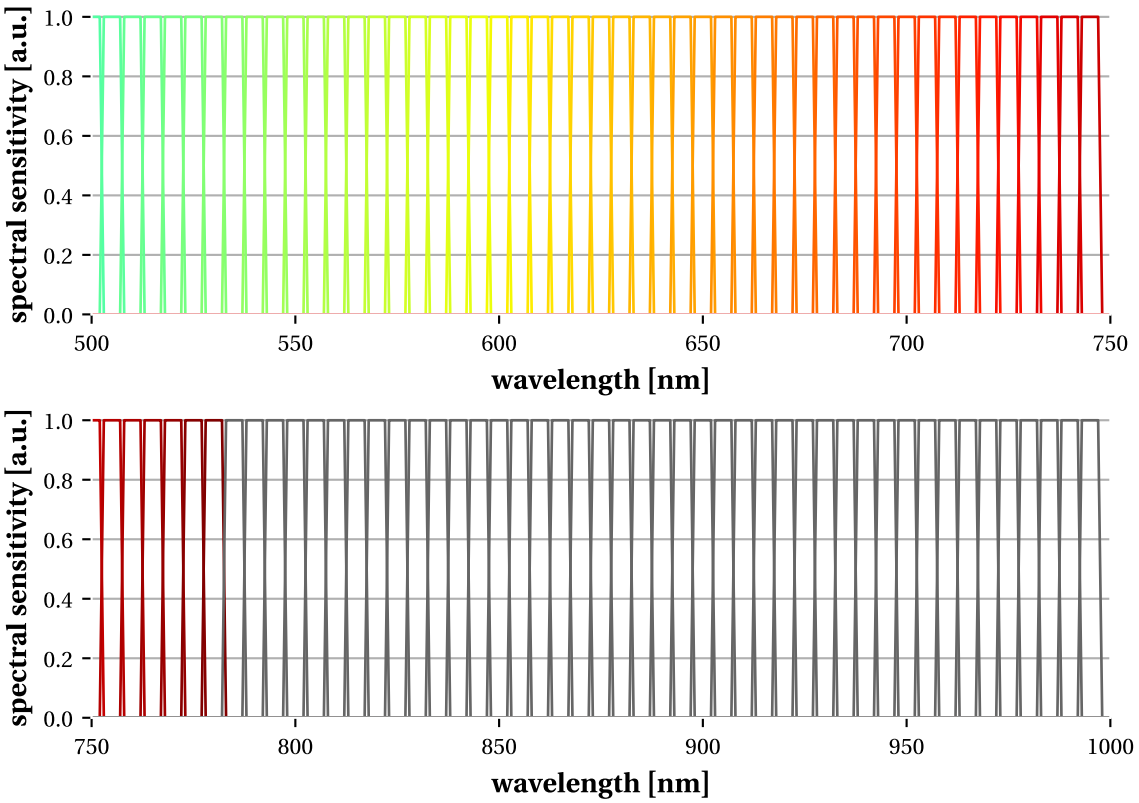


Figure 2.6: Estimated spectral sensitivities of our hyperspectral imaging device. Unlike human RGB vision (see Figure 2.4), which is limited to 3 broad spectral channels, the TIVITA® system (Diaspective Vision GmbH, Am Salzhaff, Germany), used in our work, captures 100 narrow, contiguous spectral channels in the range 500–1000 nm at a bandwidth of 5 nm. Figure adapted from [311].

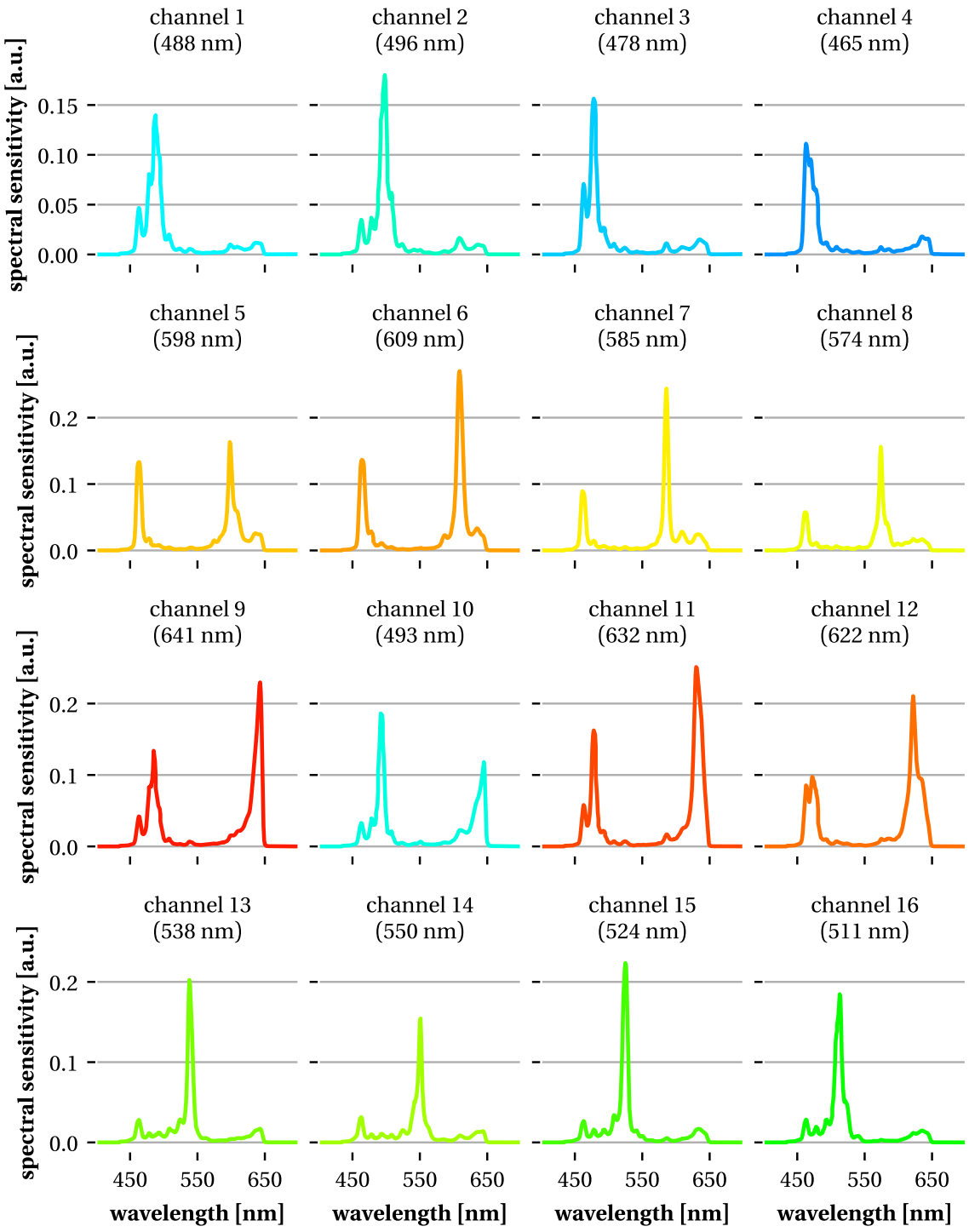


Figure 2.7: Spectral sensitivities of our multispectral imaging device. The MQ022HG-IM-SM4x4-VIS (XIMEA GmbH, Münster, Germany), used in our work, captures 16 spectral channels within the range 465–641 nm. In comparison to our hyperspectral imaging device (cf. Figure 2.6), the spectral channels are broader and often exhibit secondary peaks in spectral sensitivity due to second-order interferences in the bandpass filters. Figure adapted from [21].

2.1.3 Spectral Image Processing

Raw SI measurements are influenced by various factors, such as the illumination source, the optical properties of the tissue, and the characteristics of optical components and sensor. To extract meaningful diffuse reflectance spectra from SI data, several preprocessing steps are typically applied, including dark and white calibration, and normalization.

Dark and White Calibration Sensors in SI devices are subject to dark current, which is the signal generated by the sensor in the absence of light. It is a function of the sensor temperature and proportional to the exposure time. To correct for this noise source, a dark image is taken by keeping the camera shutter closed [215].

White calibration is performed to correct for variations in the illumination and sensor response. It involves capturing an image of a white reference standard, a material with known reflectance properties, under the same lighting conditions as the tissue [91]. Typically, a National Institute of Standards and Technology-certified Spectralon® (Labsphere Inc., North Sutton, United States of America) diffuse reflectance target is used, offering nearly uniform reflectance above 99 % across the 400–1500 nm wavelength range [95].

The tissue reflectance image R is obtained from the raw SI measurement I, the dark image D and white reference image W as:

$$R = \frac{I - D}{W - D} \tag{2.4}$$

White calibration can be challenging under dynamically changing illumination, as a new white reference image must be captured each time the lighting changes. This is particularly problematic in open surgeries, where lighting conditions can vary substantially due to multiple light sources like overhead lamps, ceiling lights, and head torches, which are frequently adjusted or switched on and off during the procedure [24]. Recalibrating with a diffuse reflectance target at the surgical site is generally infeasible due to sterility concerns. As a result, up to date, all additional light sources must be turned off during SI acquisition to maintain the static lighting conditions used during calibration [141, 188]. We address this important issue in Chapter 3.

Normalization Variations in illumination intensity, such as changes in the distance between the light source, tissue, and camera, can cause multiplicative shifts in the measured spectral reflectance R [66]. To correct for this, the reflectance image is normalized by dividing it by the mean reflectance across all spectral channels, effectively

performing ℓ^1-normalization along the spectral dimension [367]. This normalization is essential for ensuring consistent comparison of reflectance spectra across different measurements.

Additional processing steps may be required depending on the specific application and camera setup, such as compensating for a non-linear camera response or temperature-dependent dark current. Techniques to handle these challenges have been explored in prior work [134, 126, 230, 170].

2.1.4 Functional Parameter Estimation from Spectral Images

A key application of medical SI is estimating functional tissue parameters at each image pixel, as illustrated in Figure 1.2 with an example parametric map. This information can offer valuable guidance for perioperative decision-making. However, several challenges need to be addressed to enable functional tissue parameter estimation based on SI data.

While light-tissue interactions in biological tissue can be analytically described under mild modeling assumptions (e.g., only elastic scattering being present) using the radiative transfer equation, this differential equation is difficult to solve without introducing simplifications based on strong modeling assumptions [384]. One common simplification leads to the Beer-Lambert law, the most widely used method for estimating functional parameters from SI data [77, 300, 29, 368, 256]. Although this approach is computationally efficient, it relies on several assumptions that are violated in clinical SI applications. These include the assumption that photons travel the same path length through tissue (ignoring wavelength-dependency and tissue inhomogeneities), that chromophores do not interact with each other (invalid, for instance, with fluorescent molecules), that scattering remains constant (ignoring, for instance, changes in scattering coefficients during neuronal or muscle activation [171]), and that chromophore concentrations are homogeneous across the tissue (which is rarely the case due to the presence of multiple tissue types) [138]. A comprehensive overview of the extensions of the Beer-Lambert law and their limitations can be found in [256]. Overall, methods derived from the Beer-Lambert law are inadequate for estimating functional parameters in complex biological tissues [366].

Another substantial challenge in validating functional parameter estimation methods is the absence of ground truth data. Currently, there is no established reference technique that can provide functional parameters, such as StO_2, across the entire field of view of a camera.

To overcome these challenges and the limitations of Beer-Lambert-based methods, several alternative approaches have been proposed to estimate functional tissue pa-

rameters from HSI and MSI data. These include techniques such as spectral derivatives and Monte Carlo simulations.

Spectral Derivatives Holmer et al. proposed using spectral derivatives to estimate functional tissue properties from HSI data acquired with the TIVITA® systems. Since changes in the molecular composition of biological tissue cause intensity variations in the reflectance spectrum without shifting the positions of its peaks, they suggested quantifying chromophore concentrations using the first and second derivatives of the spectrum in wavelength regions most sensitive to concentration changes [141]. For example, hemoglobin displays characteristic absorption peaks between 570 nm and 590 nm, with intensity variations being a function of its oxygen content. While this method is straightforward to implement, it requires a high number of narrow spectral bands to accurately compute the spectral derivatives. To this end, this approach is used throughout this thesis in the estimation of functional parameters from the HSI cubes acquired with the TIVITA® cameras, while a different approach is needed for functional parameter estimation from MSI data. Further details on the spectral derivative approach are provided in [141].

Monte Carlo Approach To enable a data-driven approach for estimating functional tissue parameters from MSI data in the absence of labeled real data, machine learning (ML) methods have been proposed to regress these parameters using Monte Carlo simulations of diffuse reflectance spectra [367, 368]. This method involves generating a digital representation of the tissue, incorporating all necessary optical and physiological parameters. Subsequently, the probabilistic path of photons through the tissue is simulated and the fraction of diffusely reflected photons collected to determine the reflectance spectrum. We employ this approach in Chapter 3 to estimate the StO_2 from MSI data. Specifically, our simulations utilize the GPU-accelerated version of the Monte Carlo Multi-Layered framework [356, 357], developed by Alerstam et al. [11], with functional parameter regression performed using a random forest model (cf. Section 2.3.2 for an introduction of random forests). Further implementation details can be found in [366, 367, 21] and in Chapter 3.

2.2 Medical Background

Having covered the physical foundation of SI and the functional parameter estimation thereof, this section provides the medical background necessary to understand the clinical motivation, challenges and hypothesis related to the exploitation of SI for functional parameter estimation and fully semantic scene segmentation in surgery

(Section 2.2.1), as well as for automated sepsis diagnosis and mortality prediction in perioperative care (Section 2.2.2).

2.2.1 Challenges and Future of Surgical Interventions

Surgical interventions are increasingly complex procedures that require a high level of skill, precision, and decision-making [299, 112]. For instance, complications following visceral surgery, also known as abdominal surgery, remain a major concern, affecting nearly half of all patients undergoing major abdominal procedures. Postoperative complications lead to a considerable increase in length of hospital stay, healthcare costs, and patient morbidity and mortality rates [198]. In Germany, 2 % of patients undergoing visceral surgery die within the hospital [33]. At a global scale, death within 30 days from surgery is the third leading cause of mortality, accounting for 7.7 % of all global deaths [247].

Over the past decades, surgical techniques have evolved significantly with the rise of laparoscopic surgery. Unlike traditional open surgery, laparoscopic surgery is performed through small incisions, offering benefits such as reduced postoperative pain, shorter recovery times, and fewer postoperative complications [238]. As a result, laparoscopic procedures are increasingly replacing open surgeries [323, 305].

However, despite these advantages, laparoscopic surgery presents new challenges for surgeons, such as reduced tactile feedback, limited dexterity and field of view due to the confined space, and a two-dimensional camera view of the surgical field that hampers depth perception [45]. Robot-assisted surgery, introduced in 1984 to address these limitations [114], enhances surgical precision and reduces complication rates by providing improved dexterity, tremor reduction, and 3-dimensional visualization of the surgical field. This leads to better outcomes and a more ergonomic working environment for surgeons [244]. As a result, robot-assisted surgery has gained popularity across various surgical specialties and has become the gold standard in several minimally invasive procedures such as tumor nephrectomy, renal tumor excision, and prostatectomy [49].

Despite these substantial advancements in surgical techniques and technology, surgical interventions remain challenging and still carry a high risk of complications [191]. Approximately 30 % of surgical complications are attributed to human error, particularly misrecognition [333]. Patient outcomes are linked to the technical skills of the surgeon [330]. Inexperienced surgeons often lack sufficient anatomical knowledge, and reduction of cognitive abilities due to fatigue deteriorates the surgeon's performance [166, 129].

To this end, further technological innovations are needed to enhance surgical vision, provide real-time guidance, and support decision-making, ultimately reducing com-

plications and improving patient outcomes [60]. We propose that functional tissue monitoring and automated surgical scene segmentation based on SI data could significantly enhance the quality and outcomes of surgical care. The following sections discuss the importance of functional tissue information (Section 2.2.1.1) and fully semantic scene segmentation (Section 2.2.1.2) in surgery.

2.2.1.1 Functional Tissue Information in Surgery

Real-time functional tissue information, such as StO_2, perfusion, hemoglobin and water content (cf. Figure 1.2 for exemplary parametric maps), is crucial for a variety of surgical procedures.

In resective procedures such as tumornephrectomy or hemicolectomy, for instance, it is crucial to interrupt blood flow to a specific tissue region through vascular clamping, a process referred to as ischemia induction [338]. Verifying successful ischemia is essential, as selective clamping of segmental arteries can be challenging due to substantial inter-patient variability in vascular anatomy. Failure to achieve proper ischemia can result in excessive bleeding during resection [234]. The current gold standard involves injecting a fluorescent dye, leading to several limitations – the test cannot be easily repeated if clamping is unsuccessful, and there is a risk of severe complications, such as anaphylactic shock [108, 64]. MSI has shown promise for real-time ischemia monitoring without the need for an invasive application of contrast agents [21].

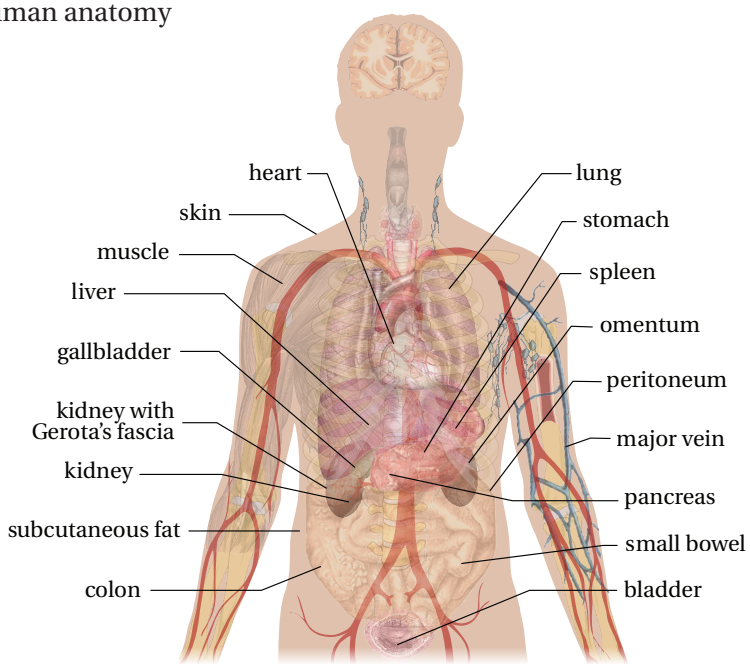
Inversely, ensuring adequate blood and oxygen supply to tissues during reperfusion is critical, especially in anastomotic areas and transplanted organs. Inadequate tissue perfusion can lead to ischemia, necrosis, organ dysfunction, and anastomotic leakage, all of which contribute to higher morbidity and mortality rates [228, 326]. Several studies have highlighted the potential of SI for monitoring tissue function during colorectal resection [159], liver resection [331], esophagectomy [181], and transplantation [332, 316].

Furthermore, functional tissue information obtained from SI data also holds potential for guiding therapy, such as optimizing the treatment of intraoperative hemorrhagic shock [327]. Additionally, SI has been investigated for improving surgical techniques, such as determining optimal resection margins in oncological surgeries [159], and improving anastomotic techniques in esophagectomy [248].

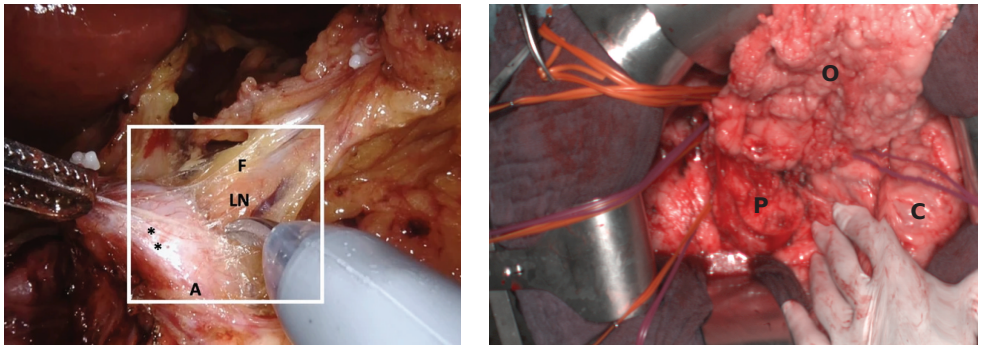
2.2.1.2 Automated Surgical Scene Segmentation

Abdominal organs vary substantially among patients in terms of size, shape, position, and vascular supply, making tissue differentiation in visceral surgery particularly challenging, as depicted in Figure 2.8 [285, 150]. Surgeons must meticulously identify

(a) idealistic human anatomy



(b) intraoperative reality



∗ nerve **A** artery **LN** lymph node **F** fat tissue **P** pancreas **O** omentum **C** colon

Figure 2.8: Human anatomy. (a) The anatomical sketch illustrates the 18 organ classes subject to surgical scene segmentation in Chapter 5 and Chapter 6. (b) In opposite to the clear tissue discriminability conveyed in the anatomical sketch, intraoperative tissue discrimination is challenging. Sample images are shown for a minimally invasive gastrectomy (left) and an open pancreatectomy (right). Anatomical sketch adapted from [311], based on an image by Mikael Häggström via Wikimedia Commons, Public Domain [132]. Intraoperative sample figure from [191].

structures in each case, with anatomical variations further complicated by the presence of connective tissue covering organs and vulnerable structures like nerves, vessels, and ducts. This increases the risk of unintended tissue damage and surgical complications. For example, surgically-induced neuropathic pain affects 10–50 % of patients undergoing routine surgeries due to nerve transection, contusion, stretching, or inflammation [173, 324]. Bile duct injury, primarily caused by misidentification of biliary anatomy by the surgeon, is a severe complication occuring in 0.4–1.5 % of cholecystectomies. It results in increased postoperative morbidity and mortality rates and often leads to a substantial reduction in quality of life [262, 18, 63]. Tissue discrimination becomes even more difficult in the presence of pathologies such as tumors, which can distort anatomy and appearance. Delineating pathological from healthy tissue is critical, as removing healthy tissue may cause complications and functional impairments, while incomplete tumor removal increases the risk of recurrence and reoperation [254, 228].

Multiple studies have demonstrated the potential of SI in accurately identifying specific structures, such as tumors, and delineating their boundaries [203, 229, 255, 202]. However, the potential of SI for automated, fully semantic scene segmentation in visceral surgery remains underexplored. Surgical scene segmentation, which involves the precise identification and delineation of all relevant anatomical structures in the surgical field (e.g., organs, tissues, instruments, pathologies), is critical for advancing computer-assisted surgery. Providing this information in real-time, for instance through augmented reality overlays, could help reduce human misinterpretations and ensure safer, high-quality surgeries, that are less dependent the surgeon's experience level [184]. Beyond improving intraoperative tissue discrimination, automated surgical scene segmentation could serve as foundation for numerous other applications, such as surgical education, tracking of target structures, and the development of navigation and decision support systems that are aware of the surgical context, offering warnings against potential complications and providing actionable recommendations [183, 128]. Furthermore, it is an important prerequisite for autonomous robotic surgery, where the robot must fully understand the surgical scene to perform procedures independently [308, 184].

While surgical scene segmentation is an active area of research, most existing methods rely on RGB images, as this is the standard imaging modality used in minimally invasive surgeries. However, RGB images provide limited information about the underlying tissue properties and composition. In contrast, SI provides rich spectral data that captures detailed biochemical information, offering the potential for more precise tissue differentiation. This suggests that integrating SI could lead to a more comprehensive and accurate segmentation of the surgical scene compared to conventional RGB imaging. To explore this hypothesis, we investigate the potential of SI for automated surgical scene segmentation in Chapter 5 and Chapter 6.

2.2.2 Sepsis and Mortality in Intensive Care

Sepsis is a severe clinical syndrome in which infection induces a dysregulated immune response leading to life-threatening organ dysfunction [320]. It is associated with high mortality and morbidity and requires immediate recognition and treatment to prevent further organ damage and death. The following sections provide an overview of the epidemiology (Section 2.2.2.1), pathophysiology (Section 2.2.2.2), and diagnosis and therapy of sepsis (Section 2.2.2.3).

2.2.2.1 Epidemiology of Sepsis

The Global Picture Despite intensive research efforts, sepsis remains a major cause of morbidity and mortality worldwide [362]. Collecting population-level data on sepsis is challenging, particularly in low-income countries where sepsis often goes unrecognized or unreported [103]. As a result, estimating the true global burden of sepsis is complex. The 2020 IHME Global Burden of Sepsis study [296] estimated that in 2017, sepsis affected approximately 48.9 million people worldwide (with a 95 % confidence interval (CI) of 38.9 to 62.9 million). These cases led to 11 million sepsis-related deaths (with a 95 % CI of 10 to 12 million). Notably, sepsis contributed to approximately 19.7 % (95 % CI of 18.2 to 21.4 %) of all deaths in 2017. The distribution of sepsis cases is uneven across countries (cf. Figure 2.9), with approximately 85 % (95 % CI: 82.2 to 87.4 %) occurring in lower-middle-income countries.

A study conducted across 1072 US hospitals in 2014 found that sepsis was the most common cause of in-hospital deaths, with one in every two to 3 deaths occurring in patients with sepsis. In most cases, sepsis was already present on admission [210]. Furthermore, physical disability, cognitive impairment, and hospital readmission are common outcomes for sepsis survivors, requiring ongoing medical treatment and support and highlighting the substantial economic burden of sepsis [271]. Every third sepsis survivor dies within one year, and every sixth survivor develops persistent cognitive impairment [271, 242]. According to a retrospective study on about 2.5 million sepsis cases in the US between 2010 and 2016, the average cost per hospitalized sepsis patient ranges from about $18 000 to over $50 000. In 2013, the total cost of sepsis in the US was estimated to be over $24 billion, representing 13 % of all US hospital costs [260].

Epidemiology in Germany In Germany, an average of 169 patients die from sepsis daily. In 2017, there were approximately 90 000 sepsis cases, with around 26 000 involving septic shock. The overall in-hospital mortality rate for septic patients was 38.4 %, rising to 56.7 % for those with septic shock [362]. The number of sepsis cases is continuously increasing at a rate of 5.7 % per year based on data from 2013 to 2017 [102]. According to

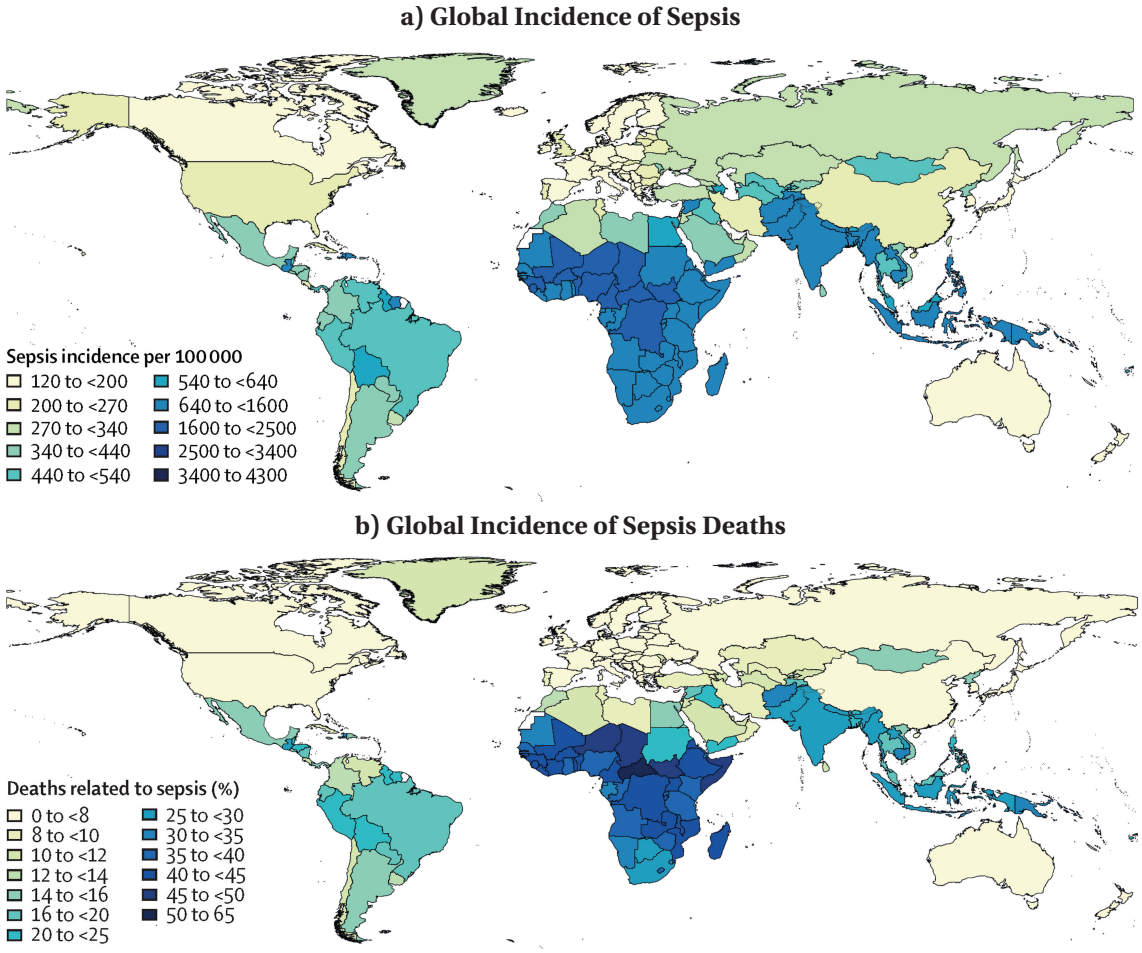


Figure 2.9: Global burden of sepsis in 2017. Map of the estimated incidence of sepsis (a) and deaths related to sepsis (b) for all countries worldwide in 2017. Figure modified from [296].

a study by Rose et al., the sepsis incidence in Germany was 178 per 100 000 inhabitants in 2016. Substantial regional differences in sepsis prevalence and mortality rates exist and could be associated with the socioeconomic status and pharmacy density [293]. Among 11 883 patients in 133 ICUs, a sepsis rate of 17.9 % could be observed, with in-hospital mortality ranging from 40.4 % to 55.2 % [88, 127].

2.2.2.2 Pathophysiology of Sepsis

Sepsis is a syndrome which is both influenced by pathogen (e.g., kind of pathogen, site of infection) and host factors (e.g., genetics, age, comorbidities, environmental factors) and evolves over time. The pathogenesis of sepsis therefore remains heterogeneous and not fully understood [90, 362, 55]. Distinguished from infections, the key characteristics of sepsis are (1) a dysregulated host response to infection and (2) the presence of organ dysfunction [320]. The goal of this section is to provide a high-level summary of the main drivers of sepsis pathophysiology, namely a dysregulated inflammation response that triggers a complex interplay between endothelial and immune system with associated coagulation abnormalities (cf. Figure 2.10) [19].

Initiation through Infection Sepsis arises from an initial infection, with the lung and abdomen being the most common sites of infection [200, 351]. In patients that are immunocompetent at the time of infection, bacterial infections, particularly with pathogens like *Escherichia coli* and *Staphylococcus aureus*, predominate. Conversely, viral and fungal infections are more common in immunodeficient patients [362].

Immune Dysregulation In immune homeostasis, there is a finely tuned balance between pro-inflammatory immune response, which is essential for the elimination of pathogens, and anti-inflammatory immune response, which is essential for the regulation of inflammation and tissue repair. The pathophysiology of sepsis is characterized by a disruption of this balance. The resulting immune dysregulation varies across patients and over the course of sepsis, ranging from *hyperinflammation* with potential collateral organ damage, to *immunosuppression* (also referred to as *immuno-paralysis* or *immune exhaustion*) associated with increased susceptibility to secondary infections and reactivation of dormant viruses due to apoptotic depletion and functional unresponsiveness (“exhaustion”) of immune cells [118]. The timeline of immune dysregulation in sepsis is not yet well understood [19]. While previous studies suggested that sepsis progresses from an initial hyperinflammation phase to a later immunosuppression phase, more recent studies lead to the hypothesis that phases of hyperinflammation and immunosuppression can alternate or coexist dynamically throughout the course of sepsis [143]. Both pro-inflammatory and anti-inflammatory immune

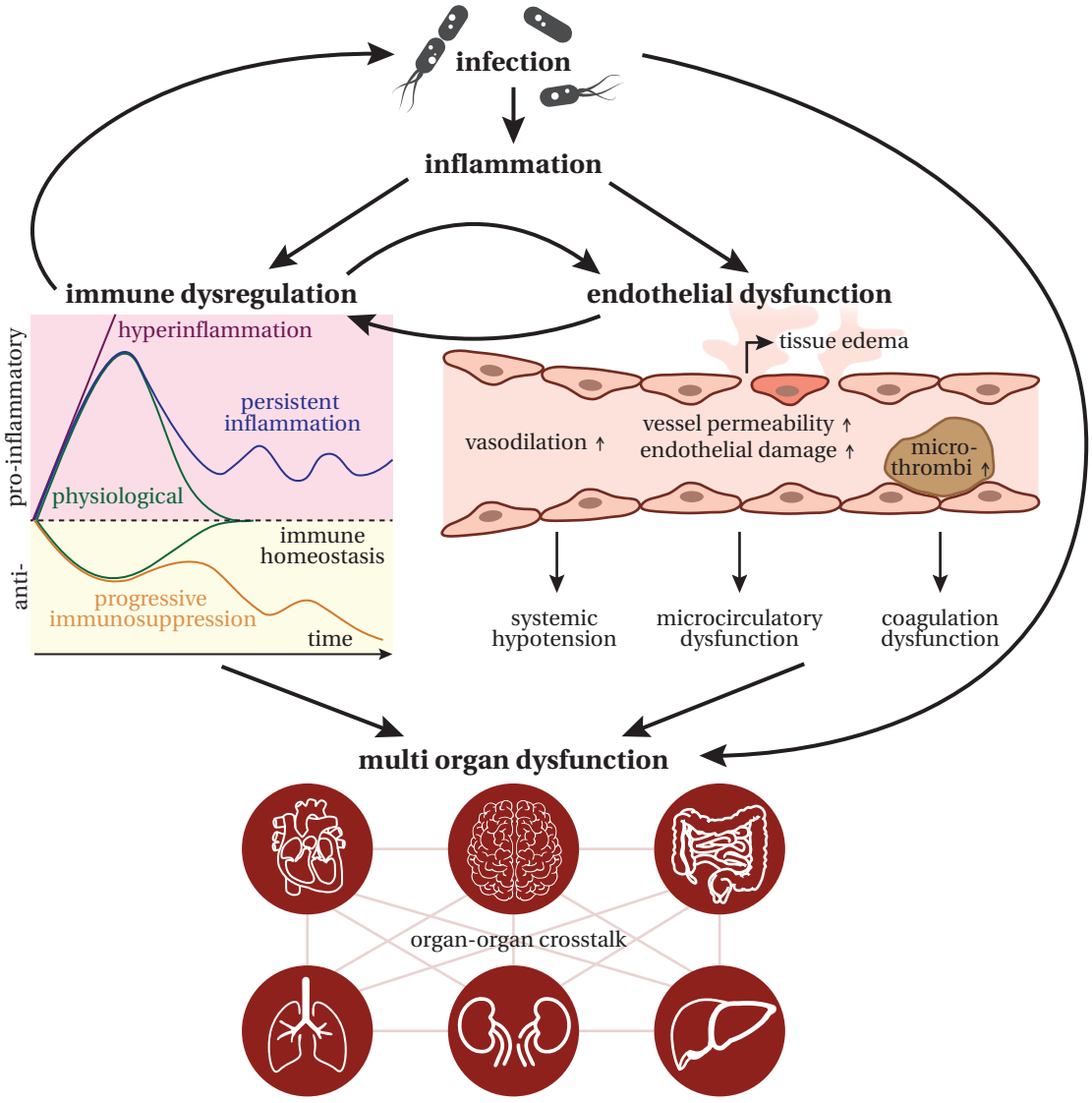


Figure 2.10: Pathophysiology of sepsis. Sepsis is initiated through a local infection. Normally, the immune system balances pro- and anti-inflammatory responses to maintain homeostasis. However, sepsis involves immune dysregulation, ranging from hyperinflammation to immunosuppression. The exaggerated immune response collaterally damages the endothelium and microvasculature, causing edema, systemic hypotension, and microcirculatory and coagulation dysfunction. These processes lead to organ dysfunction, and due to organ-organ crosstalk, multiple organ systems often fail simultaneously in sepsis.

responses are accompanied by the release of cytokines, chemokines, and other inflammatory mediators. An excessive secretion of cytokines, referred to as *cytokine storm*, is associated with sepsis. It leads to an auto-amplification cascade of the immune response causing fever, shock, respiratory failure and early death due to multiple organ failure [246, 19].

Endothelial, Microcirculatory and Coagulation Dysfunction Endothelial cells, which line the interior surface of vessels, contribute to the pro-inflammatory immune response by sensing pathogens, recruiting immune cells, and producing inflammatory mediators [16]. Early in sepsis, a maladaptive endothelial cell activation and endothelial damage (e.g., due to the cytokine storm or bacterial endotoxins) lead to an increased vessel permeability. The increased vessel permeability allows the leakage of plasma proteins and fluids into the interstitial space. This leads to the formation of interstitial edema, which provoke complications such as respiratory failure (referred to as acute respiratory distress syndrome) due to the accumulation of fluid in the alveoli [199] or septic encephalopathy caused by damage in the blood-brain barrier [217]. Furthermore, the formation of interstitial edema promotes increased venous pressure, resulting in areas of microvascular stasis and tissue hypoperfusion [90]. The decreased local blood flow velocity might also contribute to the amplification of the inflammatory response by increasing the contact time between immune cells and endothelial cells [267].

The activation of endothelial cells leads them to release more nitric oxide, which causes vasodilation and disrupts calcium homeostasis and compensatory reflexes. The heterogeneous distribution of nitric oxide expression in the vascular bed causes uneven vasodilation, leading to a characteristic redistribution of blood flow in sepsis. Under physiological conditions, perfusion and oxygen delivery are coupled to metabolic demand, but in sepsis, this redistribution disrupts efficient oxygen and nutrient delivery. As a result, areas of tissue become either hypo- or hyperperfused, contributing to organ dysfunction [267]. Extensive vasodilation and loss of intravascular fluid volume due to increased vessel permeability are key drivers in the development of systemic hypotension, compromising blood flow and consequently oxygen supply to vital organs such as the heart, brain, and kidneys, and ultimately causing organ damage. Patients with persistent hypotension that does not respond to resuscitation therapy are diagnosed with septic shock. Septic shock is associated with a hospital mortality rate of above 40 %, that is twice as high as the in-hospital mortality of sepsis alone [320, 296].

Under physiological conditions, the endothelium regulates the balance between coagulation and fibrinolysis to achieve hemostasis and prevent both systemic bleeding and clotting [139]. Endothelial damage due to the inflammation in sepsis disrupts this balance, leading to a pro-coagulant state [19]. As a consequence of the maladaptive endothelial cell activation, the release of coagulation factors by the endothelial cells is amplified, which promotes platelet aggregation and the formation of microthrombi

[362]. The coagulation dysfunction in sepsis varies from mild to severe hematological abnormalities with thrombi forming in small- and medium-sized vessels. The latter condition, referred to as sepsis-induced coagulopathy (SIC), evolves in up to one third of septic patients [4, 304]. Severe SIC leads to simultaneous widespread microvascular thrombosis and excessive bleeding and is a key contributor to multiorgan failure [204].

Overall, the septic failure of the microcirculation resulting from endothelial dysfunction promotes tissue hypoxia and organ dysfunction, particularly in the lungs, kidneys, and liver [362].

Multiple Organ Failure Sepsis can impact any organ system, with dysfunction varying from mild impairment to total organ failure. Single-organ dysfunction is uncommon in sepsis. Because of organ-organ crosstalk, the dysfunction of one organ typically results in the dysfunction of another, leading to the simultaneous dysfunction of multiple organ systems. For example, kidney failure may lead to fluid overlead, affecting heart and lung function, and the impaired elimination of toxins and metabolites from the blood might further impair the functioning of other organs [201]. Sepsis mortality is influenced by both the pattern and the number of co-occurring organ dysfunctions, with rates increasing as the number of organ failures rises [298]. Studies report heterogeneous mortality rates of 14–40 % with a single organ failure, 20–76 % with two organ failures, 30–90 % with 3 organ failures, and up to 100 % with 4 or more organ failures [41].

The development of organ dysfunction across various organ systems involves a combination of several mechanisms. An excessive immune response contributes to tissue damage through the toxicity of mediators, such as reactive oxygen species, which harm the endothelium and mitochondria. Circulatory alterations, including systemic hypotension, microcirculatory dysfunction, coagulopathy, and cardiovascular impairment, result in tissue edema and tissue hypoxia. In recent years, cellular metabolic alterations in septic patients are increasingly studied as they may also contribute to the development of organ dysfunction [68, 201]. Many cell death pathways are dysregulated in sepsis, either due to direct interaction with pathogens or as a result of the immune response, leading to for example increased apoptosis of endothelial cells, respiratory and gut epithelial cells, lymphocytes and cardiomyocytes [145, 201]. Mitochondrial function, which is essential in energy production, protein synthesis and catabolism, is commonly impaired in patients who do not survive sepsis [58]. As a consequence, metabolic intermediates accumulate and cells become unable to maintain homeostasis and function, leading to the apoptosis of organ and immune cells, and ultimately promoting immunosuppression and multiple organ failure [149]. The role of the gut and its microbiome in the development of multiple organ dysfunction is not yet fully understood. The composition of the gut microbiome is profoundly disturbed in critical illness, which may both be attributed to the disease itself and the interventions in

critical care, such as the administration of antibiotics, vasopressors and opioids as well as perenteral nutrition. Experimental work indicates that the microbiome plays an important role in maintaining gut-barrier function and modulation of the innate and adaptive immune system, and a disturbed composition might contribute to the development of organ dysfunction [131]. One hypothesis, supported by substantial experimental evidence, proposes that the injured gut mucosa releases toxic mediators that are transported through lymph nodes and cause dysfunction in distant organs [237]. Alternatively, an impairment in the epithelial barrier function may promote the translocation of bacteria from the gut into the bloodstream, thereby contributing to inflammation and organ dysfunction [268].

Autopsy studies suggest that in most organs, organ dysfunction is mainly functional and accompanied by only minimal structural tissue damage. Exceptions are immune cells, the gut and the spleen, in which higher levels of cell apoptosis could be observed [146, 68, 362]. The organ dysfunction in sepsis may thus be a mixture of adaptive and pathogenic responses, with the former being reversible and the latter leading to irreversible damage. A cellular metabolic downregulation can be observed in early sepsis, which may be a protective mechanism, seeking to re-prioritize cellular energy consumption to limit further damage to the organs and maintain energy balance [267]. Nevertheless, the downregulation of cellular metabolism may also lead to a decreased ability to respond to infections and repair tissue damage, ultimately promoting organ failure [362]. In the lungs for example, processes that clear fluid from the alveolar space are inactivated during sepsis, resulting in the progression of pulmonary edema [345].

In summary, the pathophysiology of sepsis is dynamic and heterogeneous and involves a multitude of mechanisms that contribute to the development of organ dysfunction. A dysregulated immune response, endothelial dysfunction and resulting microcirculation and coagulation dysfunction could be identified as key drivers of sepsis in the past two decades [201]. However, knowledge gaps remain: The understanding of the pathophysiological events during sepsis is incomplete, and how immunological alterations predispose to sepsis as well as the long term effects of sepsis on immunity are fairly unknown [295]. Understanding the pathophysiology of sepsis is thus an active area of research, and new insights are expected to contribute to the development of novel diagnostic and therapeutic strategies [233].

2.2.2.3 Sepsis Diagnosis and Therapy

With increasing knowledge of the pathophysiology of sepsis, the definition and diagnosis of sepsis have evolved over time. The most recent update occurred in 2016 when an expert consensus process, led by the European Society of Intensive Care Medicine and the Society of Critical Care Medicine, published the “Third International Consensus

Definitions for Sepsis and Septic Shock", known as *Sepsis-3*. As depicted in Figure 2.11, it redefined sepsis as a "life-threatening organ dysfunction caused by a dysregulated host response to infection", thereby emphasizing the significance of organ dysfunction in sepsis [320].

For clinical operationalization, organ dysfunction is assessed using the Sequential Organ Failure Assessment (SOFA) score (cf. Figure 2.12), which evaluates the function of 6 organ systems, namely the respiratory, cardiovascular, central nervous, liver, renal and coagulation system. It ranges from zero to 24, with higher scores indicating more severe organ dysfunction [349]. Sepsis is diagnosed when the SOFA score increases by at least two points within 24 h. The assessment of the SOFA score comes with considerable time and resource requirements, as laboratory tests are needed to evaluate the functionality of some organ systems (e.g., coagulation, liver, and renal system), making it less suitable for rapid bedside evaluation. Therefore, the quick Sequential Organ Failure Assessment (qSOFA) score was introduced as a simplified version of the SOFA score, focusing on 3 clinical criteria, namely increased respiratory rate, reduced systolic blood pressure, and altered mental status [320]. The qSOFA score was designed to quickly identify patients at risk of poor outcomes, with the purpose to ensure that patients suspected of having sepsis, especially in pre-hospital settings, general wards, and emergency departments, receive rapid treatment and are immediately transferred to an ICU, where the more comprehensive SOFA score could subsequently be determined [362]. However, due to the low sensitivity and specificity of the qSOFA score, it is not recommended for the diagnosis of sepsis, but rather for risk stratification [17].

Septic shock is a severe form of sepsis, characterized by persistent systemic hypotension. It is defined by the need for vasopressor therapy to maintain a mean arterial pressure (MAP) of at least 65 mmHg, together with a serum lactate level exceeding 2 mmol/L, despite adequate fluid resuscitation [320].

The rapid and targeted treatment of sepsis is crucial, as with every hour the treatment is delayed, the mortality increases due to irreversible organ damage [97]. A major challenge consists in the early diagnosis of sepsis. Clinical symptoms of infection are often atypical and unspecific (e.g., in most cases absence of fever), especially in elderly patients, causing common delays in diagnosis [136]. Furthermore, organ dysfunction is a late sign of sepsis, making the SOFA score unsuitable for early diagnosis. The qSOFA score has poor sensitivity and specificity [17]. To this end, an active body of research investigates biomarkers that can support both early sepsis diagnosis and risk stratification, such that preventive, diagnostic, therapeutic and palliative measures can be initiated in a timely manner [362]. An overview of the current state of the art in early sepsis diagnosis is given in Section 7.1.

Evidence-based guidelines on the management of sepsis and septic shock are provided by the Surviving Sepsis Campaign, an international initiative sponsored by the European Society of Intensive Care Medicine and the Society of Critical Care Medicine

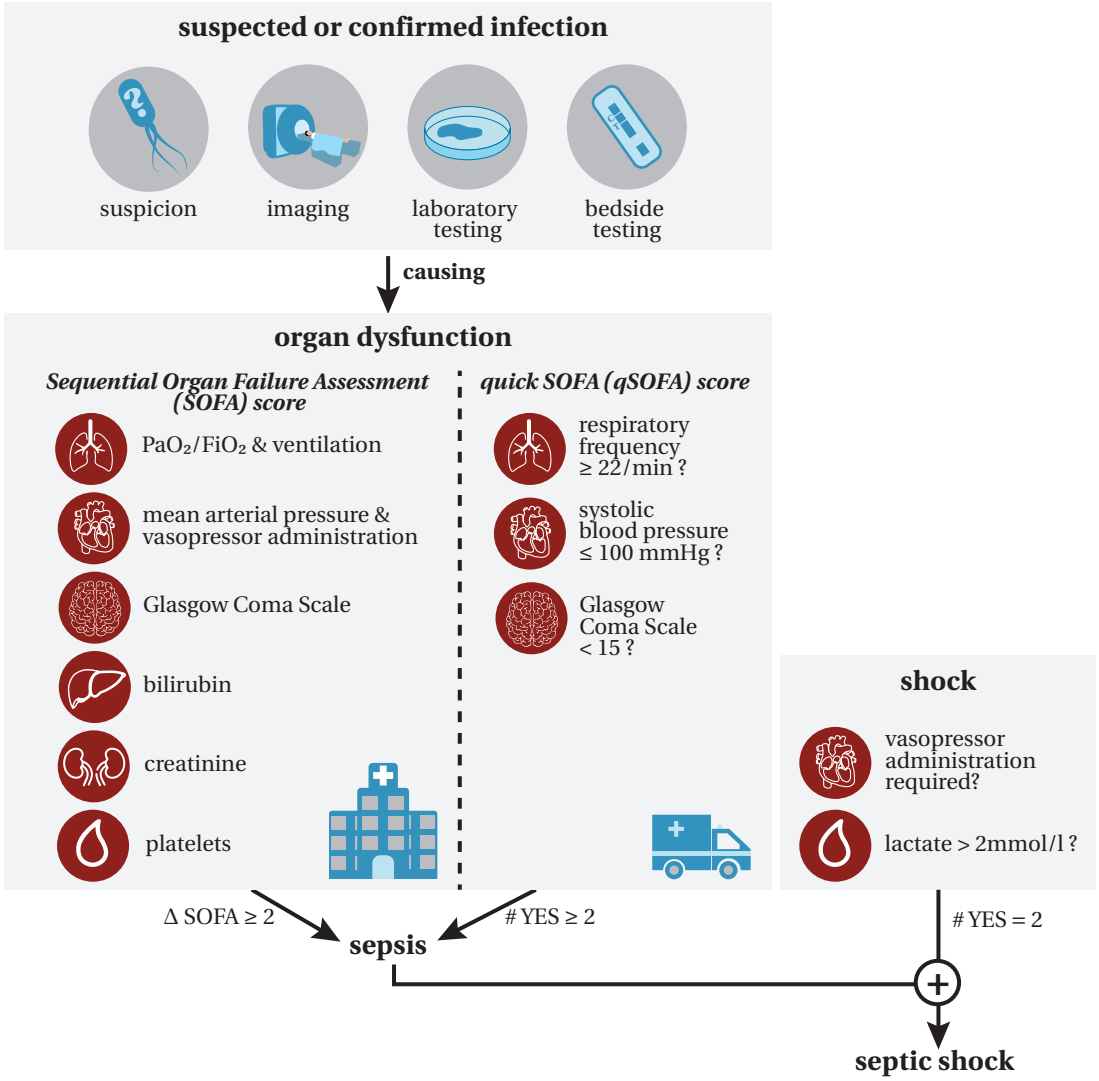


Figure 2.11: Diagnosis of sepsis and septic shock according to Sepsis-3. Sepsis is defined as "life-threatening organ dysfunction caused by a dysregulated host response to infection" [320]. The Sequential Organ Failure Assessment (SOFA) score (cf. Figure 2.12) is employed as a clinical measure to evaluate organ dysfunction, assessing the functionality of 6 organ systems using vital signs and laboratory parameters. Sepsis is diagnosed when the SOFA score increases by at least two points within 24 h. In pre-clinical and emergency settings, the quick Sequential Organ Failure Assessment (qSOFA) score is recommended due to its suitability for bedside assessment. Septic shock, a subset of sepsis, is characterized by the need for vasopressor administration and an elevation in serum lactate levels.

organ system	parameter	sub-scores 1	2	3	4
respiratory system	PaO_2/FiO_2 [mmHg]	< 400	< 300	< 200 & ventilation	< 100 & ventilation
cardio-vascular system	MAP [mmHg] & vasopressor dose [µg/kg/min]	MAP < 70	dopamine ≤ 5 OR dobutamine	dopamine > 5 OR arenaline ≤ 0.1 OR noradrenaline ≤ 0.1	dopamine > 15 OR arenaline > 0.1 OR noradrenaline > 0.1
central nervous system	Glasgow Coma Scale	≤ 14	≤ 12	≤ 9	< 6
liver	bilirubin [mg/dl]	≥ 1.2	≥ 2.0	≥ 6.0	> 12.0
renal system	creatinine [mg/dl]	≥ 1.2	≥ 2.0	≥ 3.5	> 5
coagulation system	platelets [1/µl]	< 150 000	< 100 000	< 50 000	< 20 000

SOFA score = Σ sub-scores

Figure 2.12: Definition of the Sequential Organ Failure Assessment (SOFA) score. To compute the SOFA score, the functionality of the respiratory, cardiovascular, central nervous, liver, renal and coagulation system is assessed based on vital and laboratory parameters. The resulting 6 sub-scores are summed up to compose the SOFA score.

[286, 89]. They were last updated in 2021 and recommend that within the first hour of recognition, serum lactate should be measured, blood cultures obtained and the administration of broad-spectrum antibiotics started. For patients with sepsis-induced hypoperfusion or septic shock, resuscitation therapy should be initiated. If hypotension persists during or after fluid resuscitation, vasopressors should be administered to maintain a MAP of at least 65 mmHg. Organ support measures should be applied for patients with sepsis-induced organ dysfunction, for example mechanical ventilation for patients with sepsis-induced respiratory failure, and renal replacement therapy for patients with sepsis-induced kidney dysfunction. The guidelines also recommend the use of adjunctive therapies in specific patient populations, for example, corticosteroids should be administered for patients with an ongoing requirement for vasopressor therapy [286]. Controlling the source of infection plays a major role in sepsis therapy, and surgical interventions might be needed (e.g., debridement, surgical drainage) [362].

The evidence-based guidelines on the treatment of sepsis have contributed to a decline in early sepsis mortality [144]. However, survivors often succumb to long-term complications like secondary infections [242]. Although decades of clinical research have targeted late-stage sepsis, no sepsis-specific therapies have been developed to date. Given the limited success of decades of clinical trials targeting hyperinflammation, recent research has shifted its focus to the immunosuppressive phase of sepsis and the development of novel immunomodulatory therapies [143, 233]. Following the discovery

of additional pathophysiological mechanisms in sepsis, such as the role of endothelial and coagulation dysfunction and cell apoptosis, novel therapies targeting these mechanisms are being developed [131, 233]. For example, inhibiting the apoptosis of immune cells has shown promising improvements of survival in septic animal models [145, 62].

Despite advances in the definition and guidelines for sepsis over the past decades aimed at enhancing accurate and early diagnosis, the rapid identification and initiation of treatment remain major challenges in sepsis management. The clinical methods for early identification of suspected sepsis patients lack specificity, leading to the overuse of antibiotics and other resources [89]. In fact, only 30 % to 40 % of patients receiving antimicrobial therapy for presumed sepsis are eventually confirmed to have an infection [179]. While early and empirical antimicrobial therapy can reduce mortality, its overuse increases the risk of antimicrobial resistance and adverse effects [350]. It is estimated that in 2019, nearly 5 million deaths were associated with bacterial antimicrobial resistance, with 1.27 million deaths directly attributed to it [245]. Consequently, the development of novel diagnostic and therapeutic strategies is an active research area, aiming to improve the accuracy of early sepsis diagnosis and develop more targeted treatment options.

2.3 Machine Learning

Due to the high dimensionality of SI data, the development of automated algorithms for surgical scene segmentation and sepsis diagnosis requires advanced statistical and ML methods [148]. This section provides a high-level overview of the fundamental concept of ML (Section 2.3.1), before introducing the ML models used in this thesis, namely random forests (Section 2.3.2) and convolutional neural networks (CNNs) (Section 2.3.3).

2.3.1 Concept of Machine Learning

ML is a subfield of artificial intelligence that focuses on the development of algorithms that can learn patterns from data and make predictions without being explicitly programmed [98]. The goal of ML is to develop models that make accurate predictions on data that was not used during training. The ability of an ML algorithm to perform well on previously unseen data is called *generalization*, and achieving good generalization is a key challenge in ML [123]. The ML models used in our work are trained on a dataset that consists of input-output pairs, where the input is a set of features $\boldsymbol{x}$, such as a spectrum or a 3-dimensional SI cube, and the output is a target variable y, which in our applications can be a functional parameter value (regression), a class

label (classification) or a two-dimensional organ map (segmentation). In this process, referred to as *supervised learning*, the model learns the relationship between the input features and the target variable.

To validate the model performance, performance measures that are appropriate for the given target task, algorithm and dataset structure are needed to quantify the error in the model predictions. A comprehensive overview of *validation metrics* for the analysis of biomedical imaging data, including common pitfalls recommendations in which instances to use which metrics, is given in [222]. Having chosen appropriate validation metrics, 3 data splits are needed to train and validate an ML algorithm: a training dataset, a validation dataset, and a test dataset. The training dataset is used to optimize the parameters of the model (i.e., decision rules in the case of random forests, weights in the case of neural networks). Model *hyperparameters* are parameters that control the learning process, such as the capacity of the model (i.e., the number of layers in a neural network, the number of decision trees in a random forest). To avoid overfitting when optimizing hyperparameters on the training data, a separate validation dataset is required. This dataset contains samples that are distinct from both the training and test datasets. The test dataset remains untouched until model parameter and hyperparameter optimization is finalized, and is then used to estimate the model generalization error.

ML is an active area of research since about 1950, and numerous algorithms have been developed since then to address different types of tasks and data structures. The success of ML was mainly driven by the availability of large datasets and the increase in computational power, which enabled the training of complex models on large datasets [197]. ML, particularly DL, a subfield of ML that focuses on the development of deep neural networks, has spurred substantial advancements across various disciplines. Recent breakthroughs include text-to-image generation [278], image synthesis and style transfer [290], as well as the success of large language models [348, 386, 274]. In the biomedical domain, ML has revolutionized key areas, leading to major advancements in protein structure prediction [164], drug discovery [282], personalized medicine [116], and medical image analysis [206]. For an in-depth overview of ML algorithms, their mathematical foundations, key methodological aspects, successes and open challenges, there are numerous decent textbooks, such as [249, 123, 98]. In the following, we focus on the ML algorithms used in this thesis: We utilize random forests for functional tissue parameter regression and automated sepsis diagnosis and mortality prediction from tabular data. The DL algorithms CNNs are employed for surgical scene segmentation and automated sepsis diagnosis and mortality prediction from SI data.

2.3.2 Random Forests

The random forest algorithm [51] is a popular ML method for classification and regression tasks based on tabular data. Besides offering low computational costs and being easy to implement, random forests have demonstrated high predictive performance in various real-world applications [389]. As illustrated in Figure 2.13, random forests are an ensemble learning method. Multiple decision trees are generated from the training data, and during inference, the predictions of the individual trees are aggregated.

Decision Trees Decision trees predict the value of a target variable y by learning simple if-then-else decision rules inferred from the set of features of the training data. If y is a categorical variable, this process is referred to as classification, whereas, if y is a continuous variable, it is referred to as regression. In the example illustrated in Figure 2.13, y is a categorical variable with two classes, called a binary classification. A key strength of decision trees is that they can handle features of different data types simultaneously, including categorical variables (e.g., sex, presence of a comorbidity) and continuous variables (e.g., age, weight), thus offering substantial flexibility regarding the combination of input features. In the example illustrated in Figure 2.13, x_0, x_1 and x_3 are continuous variables, while x_2 is a categorical variable.

Each decision rule splits the data into two partitions based on a single feature. If the feature is continuous, the split is based on a threshold t (e.g. $x_0 < t$) with the threshold being chosen from the set of midpoints between adjacent, unique feature values. For example, in the left tree in Figure 2.13, unique feature values of x_0 are: $x_0 \in \{-4, 5, 12, 15, 23, 25, 30\}$. The set of potential thresholds for x_0 is thus: $t \in \{0.5, 8.5, 13.5, 19, 24, 27.5\}$. If the feature is categorical, the split is based on the presence or absence of a category (e.g. $x_2 = A$) or set of categories (e.g. $x_2 \in A, B$).

The process of splitting the data into two partitions based on a decision rule is recursively repeated until a partition is only composed of samples from one target value, referred to as *pure node*, or a maximum tree depth or minimum number of samples in a node is reached. The terminal nodes, referred to as *leaf nodes*, are thus not necessarily pure nodes. Nevertheless, the same prediction is made for all samples in a leaf node, which is the mode of the target variable in the case of classification or the average target value in the case of regression [158].

An optimal decision rule is one that minimizes the inhomogeneity (also referred to as impurity) of the target variable y in the two resulting partitions L and R. For instance, a split resulting in pure nodes would be ideal, whereas a split resulting in a uniform distribution of target variables would be undesirable. The reasoning behind this heuristic is to favor shallow trees, that achieve high predictive accuracy for a minimal number of strongly discriminative rules. With an increasing number of tree nodes, the model

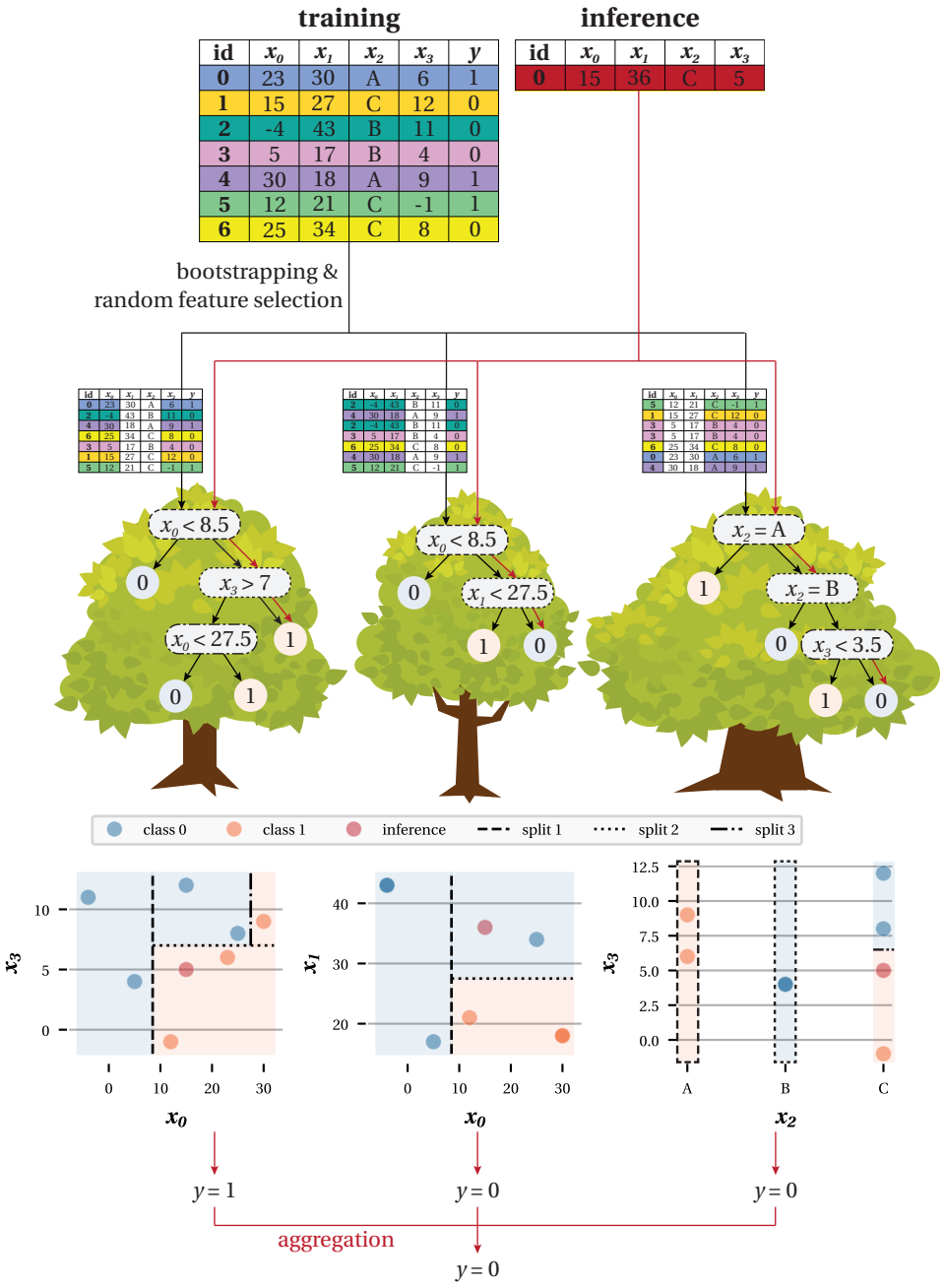


Figure 2.13: Random forest algorithm. The random forest algorithm builds multiple decision trees during training and aggregates their predictions to determine the target label y during inference. In this binary classification example with 3 decision trees, the final prediction is the mode of the tree predictions. Each decision trees consists of split nodes (white rounded rectangles) and leaf nodes (circles), with the underlying decision rules optimized by a greedy algorithm to minimize the impurity of the split. To enhance the diversity between trees and improve generalization, each tree is trained on a bootstrapped subset of the data, and at each split node, the optimal decision rule is selected on a random subset of features. For simplicity, in the example, the same random feature subset was used for all split nodes of a tree.

becomes more complex and specialized on the training data, capturing its noise and outliers. While higher model complexity can thus lead to better performance on the training data, shallow trees likely generalize better to unseen data [98, 31]. Different impurity functions have been proposed, and a review of these functions, and research targeting the optimal choice of impurity function, is provided in [276]. The most common impurity measure for classification decision trees is the gini impurity Q_g, which is defined as:

$$Q_g = 1 - \sum_{c \in C} p_c^2 \tag{2.5}$$

with p_c being the proportion of samples of class c in the partition, and C denoting the set of classes [158]. A pure node yields a gini impurity of zero. In our example of a binary classification, the gini impurity can be simplified as a function of p_1:

$$Q_g = 1 - p_1^2 - (1 - p_1)^2 = 2p_1(1 - p_1) \tag{2.6}$$

In this case, a uniform distribution of classes, which corresponds to $p_1 = 0.5$, yields the highest gini impurity of 0.5. The gini impurity for binary classification is illustrated in Figure 2.14.

The impurity measure is computed for the left partition L and right partition R of the split node, and the overall impurity is computed as the weighted sum of the impurity of the partitions, with the weights being the proportion of number of samples in the partitions ($|L|$ and $|R|$) relative to the total number of samples ($|L| + |R|$):

$$Q(\text{split}) = \frac{|L|}{|L| + |R|} Q(L) + \frac{|R|}{|L| + |R|} Q(R) \tag{2.7}$$

The intuition behind this is to penalize splits that result in small partitions, as they are more likely to be overfitting the training data. An example of a comparison between two possible decision rules with different impurity measures is illustrated in Figure 2.14.

For regression decision trees, the mean squared error can be used as impurity measure, which is defined as:

$$Q = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2.8}$$

with n being the number of samples in the partition, y_i the target value of sample i, and $\bar{y}$ the mean target value of the partition. The lowest possible value of 0 is again achieved in a pure node that only contains samples of the same target value.

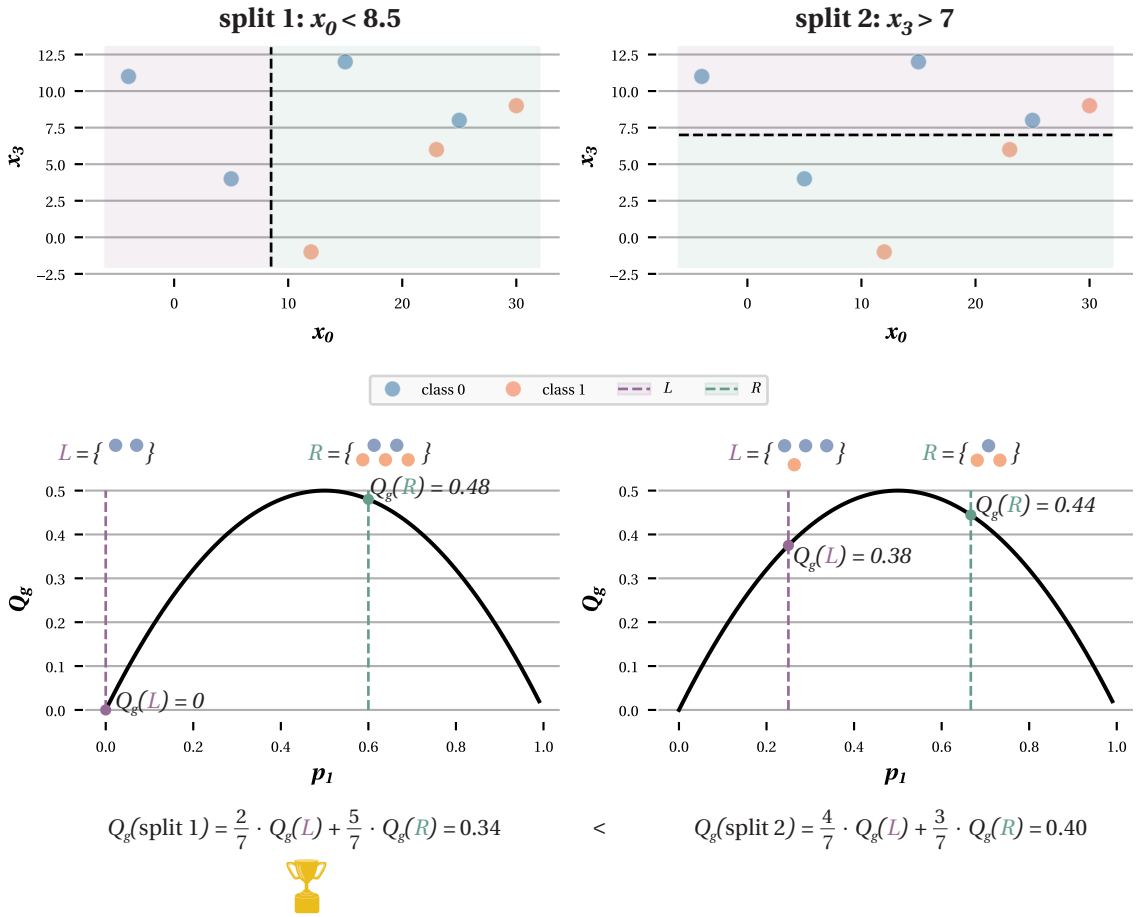


Figure 2.14: Selection of decision rules in decision trees. The optimal decision rule for a split node is the one that minimizes the impurity $Q(\text{split})$ of the resulting data partitions. Impurity can be quantified using the gini impurity Q_g, which in the case of binary classification is a function of p_1, the proportion of samples belonging to class 1 in a given partition. The impurity of a split node is calculated as the weighted sum of the impurities $Q_g(L)$ and $Q_g(R)$ of the left and right partitions L and R, with the weights determined by the proportion samples in each partition relative to the total number of samples. In this example, we evaluate two potential decision rules for the initial split in the left tree of Figure 2.13: split 1 and split 2. Split 1 yields a lower gini impurity Q_g, making it the preferred choice.

The decision tree algorithm is a greedy algorithm, meaning that it determines the optimal decision rule for a given split node without updating the decision rules of previous splits. While this local optimization does not guarantee the globally optimal set of splits, it is computationally efficient. On each split node, the decision rule is determined by iterating over all possible decision rules (including which feature to use for the split and the set of all possible splitting conditions, such as thresholds) and selecting the one that minimizes the impurity $Q(\text{split})$ of the resulting data partitions.

Pruning While a high number of tree nodes can lead to better performance on the training data, it can also result in poor generalization on unseen data [98]. To prevent overfitting, decision trees are typically pruned by limiting the maximum depth of the tree or by setting a minimum number of samples required to split a node, referred to as *pre-pruning*. Alternatively, decision trees can also be learned without these constraints, and in *post-pruning*, the complexity of the tree is reduced by removing nodes.

Bagged Decision Trees A disadvantage of decision trees is their limited robustness, as small perturbations in the training data can lead to substantial changes in the resulting sequence of decisions [98]. This sensitivity of decision trees on the training data can be mitigated by combining multiple decision trees into an ensemble model. To obtain a variety of different decision trees, each tree is trained on a randomly drawn subset of the training data, sampled with replacement – a procedure referred to as *bootstrapping*. The predictions of the individual trees are aggregated by majority voting for classification tasks and by averaging for regression tasks. The entire process is referred to as *bagging*, which stands for bootstrap aggregating [50]. Bagging can provide more accurate and stable predictions compared to individual decision trees.

Random Forests For applications in which few features dominate the decision process, the benefit of bagging could be small as the decision trees would still be highly correlated [98]. To address this issue, random forests were introduced as an extension of bagging, in which a random subset of features is selected for each split node, thus encouraging a higher degree of feature assessment variability. The random feature selection further reduces the correlation between the trees and improves the generalization of the model [389].

2.3.3 Convolutional Neural Networks

CNNs constitute a specialized architecture of neural networks developed for the analysis of grid-like data, such as images, which can be regarded as a two-dimensional grid

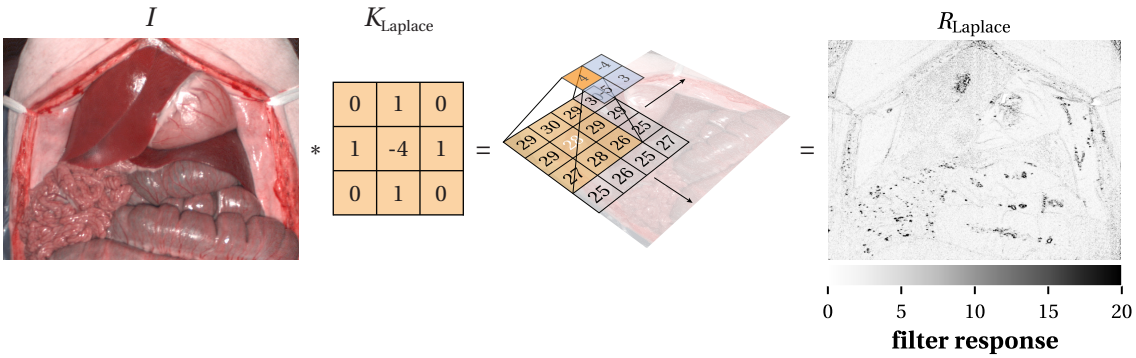


Figure 2.15: Convolution operation. The convolution operation is applied to the input data by element-wise multiplication of the kernel K with the subset of the input data I covered by the kernel, followed by the summation of the resulting products. The kernel is then shifted to the next position, and the process is repeated until the entire input data has been processed. Here, the example of a Laplace kernel K_{Laplace} is illustrated, which can be used for edge detection. Figure inspired from [311].

of pixels. In fact, state-of-the-art performance in image classification, object detection, and image segmentation tasks is achieved by CNNs [197, 123].

Convolution Operation CNNs use a mathematical operation called convolution, which involves sliding a filter, also referred to as kernel K, over the input I to extract the response R. In the example of a two-dimensional image and a two-dimensional kernel of an odd-numbered width W and height H, the convolution operation, denoted with $*$, is defined as:

$$R(i,j) = (I * K)(i,j) = \sum_{w=-\lfloor \frac{W}{2} \rfloor}^{\lfloor \frac{W}{2} \rfloor} \sum_{h=-\lfloor \frac{H}{2} \rfloor}^{\lfloor \frac{H}{2} \rfloor} I(i+w, j+h) K(w,h) \tag{2.9}$$

As illustrated in Figure 2.15, the convolution operation is applied to the input data by element-wise multiplication of the kernel with the subset of the input data covered by the kernel, followed by the summation of the resulting products. The kernel is then shifted to the next position, and the process is repeated until the entire input data has been processed. To compute the convolution for border pixels, the image borders are typically expanded, for example through padding with zeros or mirroring the image along the borders [53].

Convolutional operations have been widely utilized in the computer vision and image processing fields for many years [53]. They are performed independently at each position of the input data, making them well-suited for parallel processing on graphics

processing units (GPUs). In fact, modern GPUs have been optimized for the computation of convolutional operations, which has contributed to the success of CNNs in image processing tasks [72]. An example for a widely used convolutional operation is the Laplace kernel, which approximates a second derivative of the image I, can be used to detect edges in an image as shown in Figure 2.15 [249].

Convolutions in Convolutional Neural Networks As shown in Figure 2.15, the Laplace kernel primarily detects the edges of specular highlights, while the more relevant organ boundaries remain poorly identified. The key concept behind CNNs is that the kernel parameters, referred to as *weights*, are not fixed, but learned during training, allowing the network to discover the optimal filter for the task at hand. A major advantage of CNNs compared to fully-connected networks [389] is *parameter sharing*, where the same kernel is applied across all positions of the input data. This significantly enhances computational efficiency, particularly with high-dimensional input, as the number of parameters depends on the kernel size rather than the input size. Additionally, parameter sharing contributes to the model's translational invariance. For instance, as the same filter is applied across the entire image, the model is able to detect objects regardless of their location [123].

In CNNs, non-linear activation functions are applied to the output of the convolution operation to introduce non-linearity into the model, enabling the learning of complex patterns in the data [123]. This concept is somewhat similar to how neurons are activated in the human brain: a neuron receives input signals, which can be inhibitory or excitatory, from other neurons, and if the sum of the input signals exceeds a certain threshold, the neuron fires and transmits an output signal to the subsequent neurons. This analogy, among others, shaped the term "neural network".

Common activation functions include the rectified linear unit (ReLU) $\mathrm{ReLU}(x)$ [6], leaky rectified linear unit (LeakyReLU) $\mathrm{LeakyReLU}_\alpha(x)$ [219] and exponential linear unit (ELU) $\mathrm{ELU}_\alpha(x)$ [67] function [158], which are illustrated in Figure 2.16 and defined as follows:

$$\mathrm{ReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

$$\mathrm{LeakyReLU}_\alpha(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha x & \text{otherwise} \end{cases} \tag{2.11}$$

$$\mathrm{ELU}_\alpha(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha(\exp(x) - 1) & \text{otherwise} \end{cases} \tag{2.12}$$

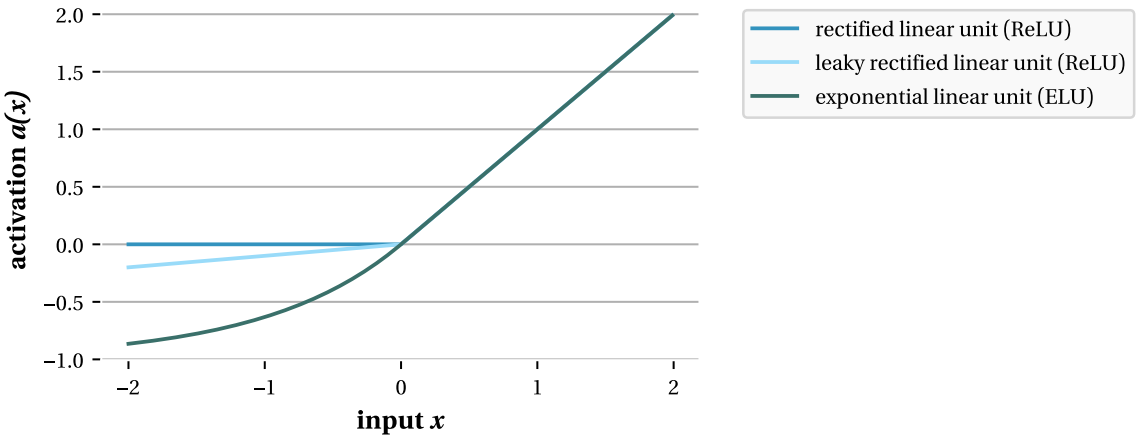


Figure 2.16: Activation functions in convolutional neural networks. Activation functions introduce non-linearity by transforming the output of the convolution operation. Common activation functions include the rectified linear unit (ReLU) $\mathrm{ReLU}(x)$, leaky rectified linear unit (LeakyReLU) $\mathrm{LeakyReLU}_\alpha(x)$, and exponential linear unit (ELU) $\mathrm{ELU}_\alpha(x)$ function.

The ReLU function is the most commonly used activation function in CNNs due to its simplicity and computational efficiency. The LeakyReLU function is a variant of the ReLU function that allows a small, non-zero gradient for negative input values, which can prevent the dying ReLU problem, where neurons that output zero are no longer updated during training [219]. The ELU function is another variant of the ReLU function that provides a smoother activation, which is beneficial in some cases [67].

Another modification to the convolution operation performed in CNNs is the addition of a bias term $b \in \mathbb{R}$, which is a learnable parameter that shifts the output of the convolution operation. It is added to the output of the convolution operation before applying the activation function, which effectively permits to perform a shift in the activation along the x-axis.

The filter response $R(i,j)$ of a two-dimensional convolution operation in a CNN for a 3-dimensional image $I(i,j,c)$, where c denotes the channels, with a total of C channels, spatial position (i,j), activation function $a(\cdot)$, filter matrix with learnable weights $\Omega \in \mathbb{R}^{W \times H \times C}$, and learnable bias b, can be expressed as [311]:

$$R(i,j) = a\left(\sum_{w=-\lfloor \frac{W}{2} \rfloor}^{\lfloor \frac{W}{2} \rfloor} \sum_{h=-\lfloor \frac{H}{2} \rfloor}^{\lfloor \frac{H}{2} \rfloor} \sum_{c=1}^{C} I(i+w, j+h, c)\Omega(w,h,c) + b\right) \tag{2.13}$$

This is the fundamental building block of a CNN. In analogy to the brain function, which involves a network of an estimated $86 \cdot 10^9$ interconnected neurons to pro-

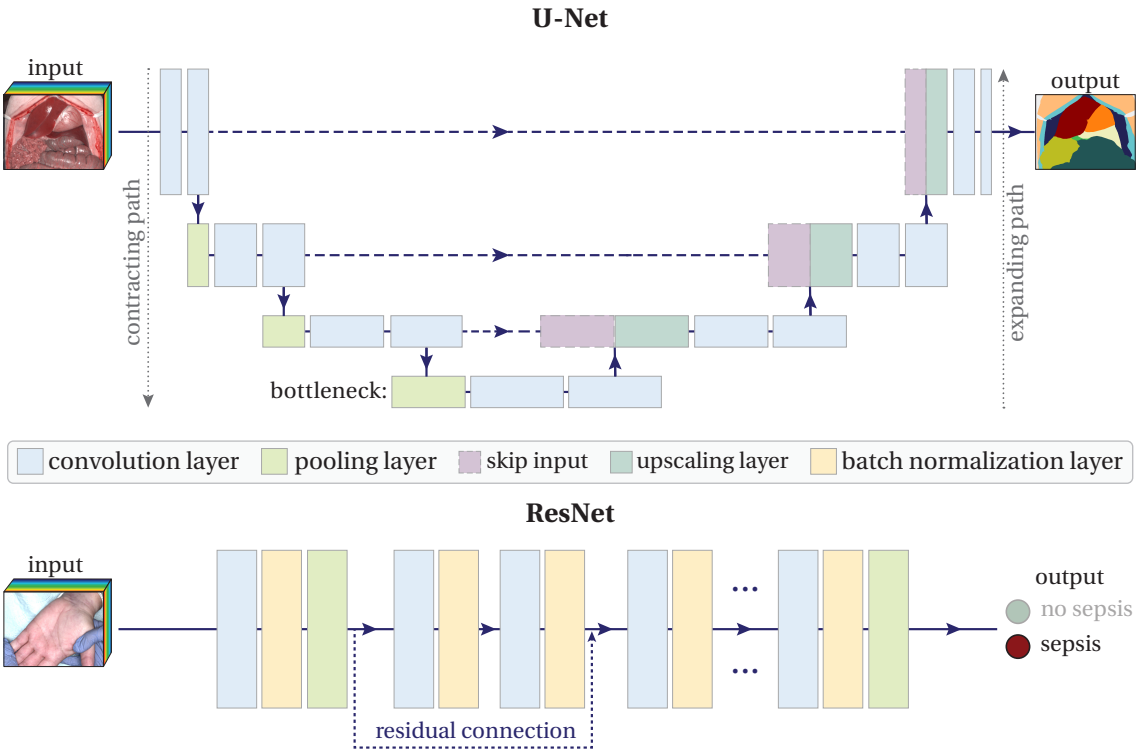


Figure 2.17: Convolutional neural network architectures. The U-Net architecture (top) combines a contracting path with a symmetric expanding path. The contracting path captures contextual information, while the expanding path enables precise localization. Skip connections between the contracting and expanding paths enable the integration of high-resolution features with contextual information. The residual network (ResNet) architecture (bottom) also uses skip connections, known as residual connections. These connections allow the network to pass the output of one layer to a deeper layer, bypassing intermediate layers. This helps to improve network convergence. U-Net sketch adapted from [308, 311].

cess information, CNNs involve a network of multiple convolutional building blocks, each with randomly initialized weights [137]. This allows each block to learn different features, enabling the network to recognize diverse patterns within the data. The convolutional building blocks are organized in layers, with the output of previous layers serving as input to subsequent layers. The layers are typically organized in a hierarchical manner, with the initial layers learning low-level features, such as edges and textures, and the subsequent layers learning high-level features, such as shapes and objects [123].

Having introduced the basic building blocks of a CNN, we now take a closer look at the CNN architectures used in this thesis.

U-Net An effective architecture for image segmentation tasks is the U-Net [291], which is characterized by a U-shaped structure as illustrated in Figure 2.17. It combines a contracting path with a symmetric expanding path. The contracting path reduces the spatial dimension of the input, thereby decreasing the number of weights in the network, which improves computational efficiency and makes the network more robust to small translational shifts in the input data [249]. Among others, this reduction is achieved through *pooling* layers, where the activation at a given location is replaced by a summary statistic of the surrounding activations, such as the arithmetic mean in *average pooling* or the maximum activation in *max pooling* [303]. The pooling and convolution process is repeated, with the number of activation maps increasing at each step, until a *bottleneck* layer is reached.

In the expanding path, upscaling layers are used to increase the spatial dimension of the activations. First, these layers perform interpolation to upsample the activation map from the previous layer, matching the shape of the corresponding activation map at the same hierarchical level in the contracting path. The upsampled activation map is then concatenated with the corresponding activation map from the contracting path, transferred to the expanding network via a *skip connection*. A convolution layer is subsequently applied to process the concatenated activation map effectively [252]. By leveraging skip connections, the U-Net architecture combines high-resolution features from the contracting path with the upsampled output, enabling the generation of high-resolution segmentation maps [291].

ResNet Residual networks (ResNets) were introduced for image recognition in 2015 [133]. A ResNet is a CNN equipped with *residual connections*, which are skip connections that transfer the output of one layer to a layer that is two or more layers deeper in the network, as illustrated in Figure 2.17. These connections mitigate a common issue in deep neural networks[3]: During backpropagation[4], gradients can become very small – a phenomenon known as the vanishing gradient problem – resulting in slow convergence or even stalled training (see [105] for a more detailed explanation). In addition to residual connections, the fundamental architecture of ResNets consists of a series of convolution, *batch normalization*, and pooling layers (cf. Figure 2.17). Batch normalization normalizes the input of each layer to have zero mean and unit variance, which mitigates the *internal covariate shift*. By enforcing that features and weights remain on a more consistent scale across layers, batch normalization promotes smoother convergence during training.

[3] Neural networks with many layers.
[4] Gradient estimation method used to update the weights of the network.

Part II

Robust Regression of Functional Tissue Parameters with Hyperspectral Imaging (RQ1)

3 Robust Functional Parameter Estimation through Automated Illuminant Estimation

As outlined in Section 1.2.1, a key challenge in safely integrating functional SI into clinical practice is the need for automated illuminant estimation. In open surgeries, illumination conditions can change dramatically throughout the procedure, and controlling the lighting or recalibrating the camera with a white reference standard is impractical due to the sterile environment and the requirement for an uninterrupted clinical workflow. This chapter analyzes the impact of varying illumination on the estimation of the functional tissue parameter StO_2 and introduces the first approach to automated illuminant estimation in the operating room (OR).

Section 3.1 provides an overview of the related work on illuminant estimation, followed by a description of our specular highlight-based approach to automated illuminant estimation and the datasets specifically acquired for this study in Section 3.2. The experimental setup and results are presented in Section 3.3, and the chapter concludes with a discussion of the strengths, limitations, and directions for future research in Section 3.4.

The research presented in this chapter was conducted in 2019, and published in the proceedings of the International Conference on Information Processing in Computer-Assisted Interventions (IPCAI) in 2020 [24], as well as in the thesis of Leonardo Ayala in 2023 [26]. It further resulted in the filing of two patents [225, 226].

3.1 Related Work

Illuminant estimation is closely linked to computational color constancy (CCC) methods, which aim to perceive constant color of an object regardless of changes in the illumination [175]. While the human visual system inherently maintains color consis-

tency despite variations in the light source's spectral composition, CCC remains an active area of research focused on replicating this capability in artificial systems [37].

CCC methods can generally be divided into ML-based approaches (cf. Section 3.1.1) and model-based (cf. Section 3.1.2) approaches. Although few methods have been specifically developed for SI data, several CCC methods originally designed for RGB image calibration can be adapted for use with SI data. The following sections provide an overview of the most promising methods, specifically those with high potential for application in open surgery SI data. For a broader survey of CCC methods, refer to [87, 208].

3.1.1 Machine Learning Approaches

CNNs are the most common ML architecture for illuminant estimation [39, 147, 5, 192, 163], consistently outperforming model-based methods on RGB data [318, 74]. However, several challenges prevented the application of ML-based approaches to surgical SI data at the time of model development and publication:

1. **Need for extensive training data:** ML approaches for illuminant estimation demand extensive training data, covering a large variety of different surgical scenes captured under a wide range of illumination conditions, to ensure model generalization to unseen data.
2. **Need for reference illuminant spectra:** In addition, corresponding reference illuminant spectra are required for the training data. However, due to sterility requirements, obtaining intraoperative surgical images with matching white reference standards is not feasible in real-world human open surgeries. This limitation makes it impractical to create an in vivo database of human open surgery images with corresponding reference illumination.
3. **Diversity of SI devices:** SI devices vary widely in spectral range, number of spectral channels, and spectral sensitivities, posing challenges for generalizing an ML model across different devices.

At the time of model development and publication, the availability of surgical SI data was limited to few, small datasets, each obtained using different SI devices [239, 73]. To this end, we decided to develop a generic, model-based illuminant estimation framework that does not depend on annotated training data.

In recent years, the availability of in vivo surgical SI data has grown, with datasets now comprising several hundred images (see Chapter 5, Chapter 6). This increase in data availability enabled us to develop the first DL-based approach to automated illuminant estimation of surgical SI data in 2024 [34]. A discussion of our DL-based approach and our model-based approach presented in this chapter is provided in Section 3.4.2.

3.1.2 Model-Based Approaches

Model-based approaches refer to automated illuminant estimation methods that depend on specific modeling assumptions. A frequent assumption is that the scene is uniformly illuminated, either due to the presence of a single illuminant or, in cases with multiple light sources, by approximating them as a single average illuminant when they are sufficiently distant or when the field of view is sufficiently small [87].

Khan et al. expanded the 4 most commonly used model-based approaches initially developed for RGB imaging – Max-RGB, Gray-world, Shades-of-gray and Gray-edge – for application to MSI data. The study reported promising performance when applied to outdoor natural scenes [175].

Max-RGB The Max-RGB approach is also known as the white patch Retinex algorithm [195]. It is based on the assumption that, for each channel c of an image, there exists at least one pixel p in the set of image pixels $\mathcal{P}$ that reflects the illuminant at its maximum level, resulting in the highest measured pixel intensity. The illuminant is estimated by collecting these maximum intensity values across all channels. With L_c representing the illuminant, and $I_c(p)$ representing the intensity of pixel p in spectral channel c, the Max-RGB approach can be expressed as:

$$L_c \propto \max_{p \in \mathcal{P}} I_c(p) = \|\boldsymbol{I}_c\|_\infty \tag{3.1}$$

Gray-world The Gray-world approach is based on the assumption that the average reflectance across a scene is achromatic, so that the average pixel intensity reflects only the characteristics of the illuminant [52]. Thus, the illuminant is estimated by computing the average intensity across all pixels:

$$L_c \propto \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} |I_c(p)| = \frac{1}{|\mathcal{P}|} \|\boldsymbol{I}_c\|_1 \tag{3.2}$$

Shades-of-gray The Shades-of-gray approach [100], which generalizes the Max-RGB and Gray-world methods, estimates the illuminant using the Minkoswski norm of the intensity image:

$$L_c \propto \left(\sum_{p \in \mathcal{P}} |I_c(p)|^m \right)^{1/m} = \|\boldsymbol{I}_c\|_m \tag{3.3}$$

While the order of the norm, m, is set to $m = 1$ for the Gray-world approach, and to $m = \infty$ for the Max-RGB approach, it is empirically set to $m = 6$ in the Shades-of-gray approach.

Gray-edge The Gray-edge approach assumes that, on average, edge differences within a scene are achromatic [346]. To minimize the impact of noise, local spatial smoothing of the intensity image $\boldsymbol{I}$ with a Gaussian filter of standard deviation σ is applied as a preprocessing step. The illuminant is then estimated by averaging the gradient of the smoothed image, I'^{σ}_{c}, across all pixels:

$$L_c \propto \left(\sum_{p \in \mathcal{P}} |I'^{\sigma}_{c}(p)|^m \right)^{1/m} = \left\| \boldsymbol{I}'^{\sigma}_{c} \right\|_m \tag{3.4}$$

The gradient I'^{σ}_{c} is computed using pixel coordinates x and y as follows:

$$I'^{\sigma}_{c}(x,y) = \sqrt{(\partial_x I^{\sigma}_{c}(x,y))^2 + (\partial_y I^{\sigma}_{c}(x,y))^2} \tag{3.5}$$

Following the recommendations of Finlayson and Trezzi, the settings $\sigma = 2$ and $m = 6$ are used for the Gray-edge approach.

Specular Highlights The spectrum of specularly reflected light closely matches that of the incident light source (cf. Section 2.1.1). This property could be leveraged to estimate the illuminant spectrum in an image by identifying specular highlights and using their spectra to infer the illuminant. However, prior work attempting to exploit this property is either limited to RGB data [99] or relies on additional modeling assumptions that do not hold in surgical scenes. For example, Imai et al. assume that the scene consists of inhomogeneous dielectric materials like plastic or paint, while Kaneko et al. incorporate assumptions about the illuminant spectrum in the form of a daylight spectrum model [151, 168].

In summary, at the time of model development and publication (2020), no prior work had explored the impact of illumination shifts on functional parameter estimation, nor had any illuminant estimation method been proposed for surgical SI specifically. Most methods developed outside the field of surgery are unlikely to generalize effectively to surgical SI data due to unrealistic model assumptions in model-based approaches, as well as the unmet need for labeled training data and limited generalization across diverse SI devices in ML-based approaches. To address these gaps in the literature, we investigate the following research questions:

RQ1.1: Can we address illumination shifts in surgeries by leveraging specularly reflected light for automated illuminant estimation?
RQ1.2: What is the impact of errors in the illuminant estimation on the performance of StO_2 estimation from SI data?
RQ1.3: How does our method perform relative to state-of-the-art model-based approaches?

3.2 Materials and Methods

This section describes our proposed approach to automated, intraoperative illuminant estimation based on specular highlights (Section 3.2.1), as well as the datasets used to address our research questions (Section 3.2.2).

3.2.1 Automated Illuminant Estimation From Specular Highlights

As illustrated in Figure 3.1, our illuminant estimation method consists of 3 main steps:

Acquisition of Low-Exposure Images Our initial approach aimed to extract the illuminant directly from the specular highlights in the MSI cubes used for functional parameter estimation. However, these MSI cubes are optimized for exposure time to capture the tissue diffuse reflectance spectra at a high signal-to-noise ratio. This typically leads to oversaturation of the camera sensor by the more intense specularly reflected light, preventing accurate measurement of the spectra of the specular highlights. To address this challenge, we propose acquiring a separate set of images specifically for illuminant estimation at a lower exposure time, which allows for the effective capture of specular highlights without oversaturation. We performed experiments to determine the optimal exposure time for this purpose, which are described in Section 3.3.

Segmentation of Specular Highlight Pixels Since both over- and undersaturated pixels provide inaccurate spectral information, we ensured that such pixels p are excluded when segmenting the specular highlight pixels.

Underexposed pixels are defined as pixels where the minimum of the intensity values $I_c(p)$ across all spectral channels c, $\min_{\forall c} I_c(p)$, is less than or equal to the sensor's dark current level $d(T_{\text{exp}})$ at the given exposure time T_{exp}. The dark current $d(T_{\text{exp}})$ is a sensor-specific parameter that needs to be determined once for a given camera.

Overexposed pixels are defined as pixels where the maximum of the intensity values $I_c(p)$ across all spectral channels c exceeds or equals a threshold intensity $I_{\text{nonlinear}}$.

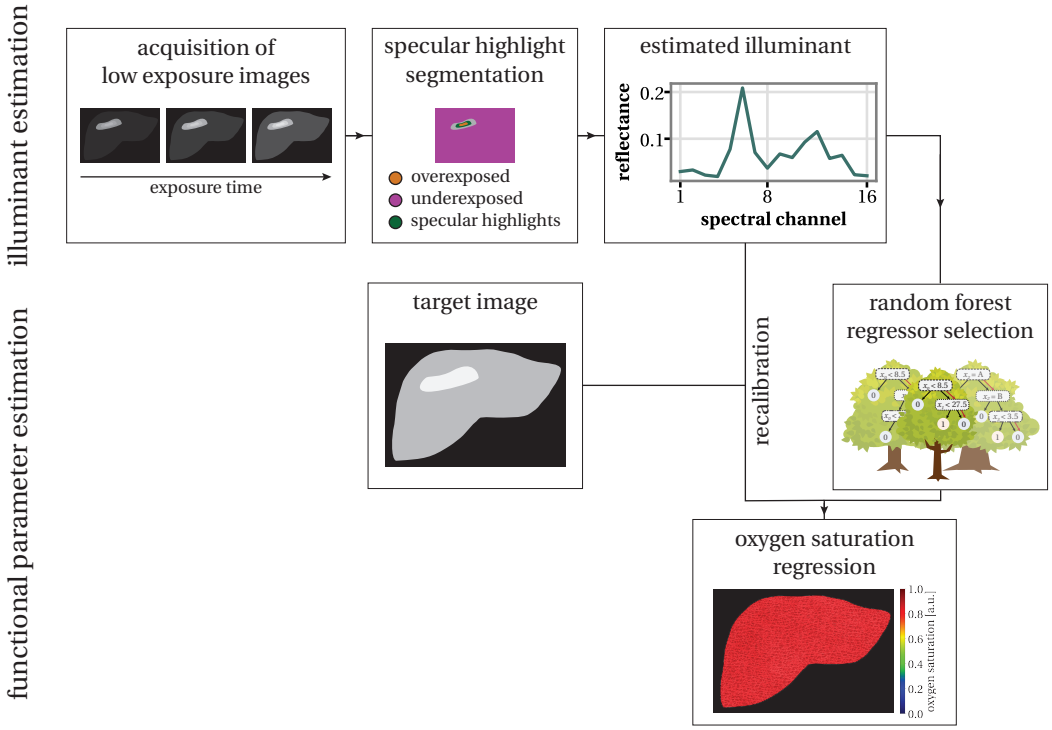


Figure 3.1: Concept of the proposed illuminant estimation method. Specular highlight pixels are segmented from a series of low-exposure images and used to estimate the illuminant spectrum. To determine tissue oxygenation in the target image, the image is recalibrated based on the estimated illuminant, and a regression model adapted to this illuminant spectrum is applied. Figure inspired by [24].

This threshold, also a sensor-specific parameter, describes the intensity level at which the sensor response becomes nonlinear, resulting in distorted spectra. In our case, $I_{\text{nonlinear}}$ is set to 950.

Thus, the set of valid pixels $\mathcal{V}$, representing the subset of image pixels $\mathcal{P}$ that are neither over- nor undersaturated, is defined as:

$$\mathcal{V} = \left\{ p \in \mathcal{P} \mid \min_{\forall c} I_c(p) > d(T_{\text{exp}}) \wedge \max_{\forall c} I_c(p) < I_{\text{nonlinear}} \right\} \tag{3.6}$$

For each pixel in $\mathcal{V}$, we compute the lightness $I_L(p)$ by averaging the intensity values $I_c(p)$ across all spectral channels c. With C representing the total number of spectral channels, this is expressed as:

$$I_L(p) = \frac{1}{C} \sum_{c=1}^{C} I_c(p) \tag{3.7}$$

Assuming that specular highlight pixels exhibit the highest lightness values, the top N pixels with the highest lightness are selected as the specular highlight pixels $\mathcal{S}$. The optimal value of N was determined experimentally, as described in Section 3.3.

Estimation of the Illuminant Spectrum Based on the assumption that specular reflection dominates over diffuse reflection in the specular highlights, the illuminant spectrum is approximated by taking the ℓ^1-normalized spectrum of the specular highlight pixels. Assuming approximately uniform illumination across the image, the illuminant spectrum $\boldsymbol{L}$ is estimated by averaging the spectra $\boldsymbol{I}$ of the specular highlight pixels $\mathcal{S}$:

$$\boldsymbol{L} = \frac{1}{|\mathcal{S}|} \sum_{p \in \mathcal{S}} \frac{\boldsymbol{I}(p)}{\|\boldsymbol{I}(p)\|_1} \tag{3.8}$$

3.2.2 Datasets

Our analysis is based on porcine ex vivo and human in vivo MSI data, supplemented by simulated tissue spectra. The simulated data enable a quantitative assessment of the StO_2 error, which is otherwise infeasible with real tissue samples due to the lack of ground truth StO_2 values (cf. Section 2.1.4).

Spectral Measurements and Light Sources The real data was acquired using the MSI snapshot camera MQ022HG-IM-SM4x4-VIS (XIMEA GmbH, Münster, Germany) described in Section 2.1.2. This camera captures 16 spectral channels in the 450–650 nm range, at spatial dimensions of 272 px × 512 px. We employed 5 different light sources, representing 4 types of illumination commonly encountered in the OR, namely xenon, halogen, fluorescent, and light-emitting diode (LED) light sources:

- L_1: Xenon (D-light P 201337 20, Karl Storz GmbH, Tuttlingen, Germany)
- L_2: Halogen (Halopar® 16, Osram GmbH, Munich, Germany)
- L_3: Fluorescent (FLS 11W 2700K, Paulmann Licht GmbH, Springe Völksen, Germany)
- L_4: Xenon (Auto LP 5131, Richard Wolf GmbH, Knittlingen, Germany)
- L_5: LED (Endolight LED 2.2, Richard Wolf GmbH, Knittlingen, Germany)

Reference Measurements As outlined in Section 2.1.3, the gold standard for correcting illumination variations involves capturing an image of a white reference standard under the same lighting conditions as the tissue. To enable a comparison with this gold standard, we acquired MSI data of a Spectralon® diffuse reflectance standard (SRT-99-050, Labsphere Inc., North Sutton, United States of America) for each of the 5 light sources L_1 to L_5. For each light source, 100 images were captured and hierarchically averaged to compute the reference illuminant spectrum.

Ex Vivo Liver Recordings We acquired MSI data of an ex vivo porcine liver alternately illuminated by the 5 light sources L_1 to L_5. To assess the robustness of our approach to changes in the imaging geometry, images were captured from 8 distinct camera poses: The liver was consistently illuminated from the east (E), while the camera was positioned at 4 angles: vertical (V) (angle of 0° to the organ surface normal), north (N), south (S), and west (W) (N, S and W at 40° to the organ surface normal). Additionally, two different camera distances were used: $D1 = 8$ cm and $D2 = 18$ cm from the liver surface.

For each camera pose, a series of 10 images per exposure time was acquired while varying the exposure time in the range 5–150 ms at increments of 5 ms. These images were used to determine the optimal exposure time for illuminant estimation and validate whether there is a benefit in aggregating the estimated illuminant across repeated low-exposure images. The chosen exposure times covered the entire range from nearly underexposed to nearly overexposed specular highlight pixels.

In Vivo Human Lips Recordings For qualitative *in vivo* validation, we acquired a MSI video stream of a human subject's lips, while switching the illumination from L_1 to

L_5 during the recording. This setup allowed us to simulate an illuminant shift in a controlled environment and assess the performance of our approach in real-time.

Simulation Framework for Oxygen Saturation Estimation We simulated tissue spectra $r^{\text{sim}}(\lambda, \boldsymbol{t}_i)$ from a vector of tissue properties $\boldsymbol{t}_i$ using the Monte Carlo simulation framework outlined in Section 2.1.4 and in [366, 367]. In this framework, the tissue is modeled as a 3-layer structure with thickness ranging from 20–2000 µm, blood volume fraction in the range 0–30 %, and StO_2 values spanning 0–100 %.

The simulated tissue reflectance spectra were subsequently converted to image intensities I_c in channel c by taking into account constant multiplicative changes of reflectance $\alpha(p)$, the noise in channel c, n_c, the illuminant spectrum $L_i(\lambda)$, and the factor $O_c(\lambda)$ which represents all linear, hardware related factors (e.g., transmittance of the optics, filter response in channel c, and quantum efficiency of the camera):

$$I_c^{\text{sim}}(p, \boldsymbol{t}_i) = \alpha(p) \cdot n_c \cdot \int_\lambda r^{\text{sim}}(\lambda, \boldsymbol{t}_i) \cdot L_i(\lambda) \cdot O_c(\lambda) d\lambda \tag{3.9}$$

In total, 15 000 tissue spectra were simulated and subsequently adjusted for each of the 5 light sources L_1 to L_5. This process resulted in a complete set of 15 000 simulated MSI spectra, each paired with corresponding reference StO_2 values, for every illumination configuration.

3.3 Experiments and Results

The details of our experimental setup are provided in Section 3.3.1, followed by the presentation of our findings in Section 3.3.2.

3.3.1 Experimental Setup

This section presents an overview of our experimental setup, detailing the validation metrics used to evaluate the performance of our illuminant estimation method, our parameter optimization strategy, and our approach to quantifying the impact of illuminant shifts on StO_2 estimation.

Validation Metrics To quantify the similarity between two spectra $\boldsymbol{L}_i$ and $\boldsymbol{L}_j$, we treated them as vectors and computed the cosine similarity, which measures the angle θ between the two vectors:

$$\text{cosine similarity}(\boldsymbol{L}_i, \boldsymbol{L}_j) = \frac{\boldsymbol{L}_i \cdot \boldsymbol{L}_j}{\|\boldsymbol{L}_i\|_2 \left\|\boldsymbol{L}_j\right\|_2} = \cos(\theta) \tag{3.10}$$

The cosine similarity ranges from -1 to 1, where 1 signifies identical spectra, 0 indicates orthogonal spectra, and -1 is obtained for opposite spectra. To assess the error in the estimation of the illuminant spectrum $\boldsymbol{L}$, we calculated the cosine similarity between the estimated illuminant spectrum $\boldsymbol{L}^{\text{est}}$ and the reference illuminant spectrum $\boldsymbol{L}^{\text{ref}}$ for each light source L_i.

To evaluate the impact of illuminant shifts on StO_2 estimation and the performance enhancements achieved through our illuminant estimation approach, we calculated the *StO_2 error*. This metric represents the absolute difference between the estimated StO_2 value and the reference StO_2 value for a given sample.

Parameter Optimization We divided our ex vivo liver dataset into two parts: a validation set for optimizing the parameters of our illuminant estimation approach, including the number of specular highlight pixels N, the exposure time T_{exp} and the required number of low-exposure images E, and a test set for evaluating the performance of our method. The validation set included data collected using light sources L_1 to L_3, while the test set comprised data obtained with the light sources L_4 and L_5.

We found that varying N between 75 and 200 had minimal impact on performance, so we set $N = 100$.

For determining the optimal exposure time T_{exp} of the low-exposure images, we introduced a goodness metric $G(T_{\text{exp}})$, with the cosine similarity between the estimated and reference illuminants increasing with G. Similar to the signal-to-noise ratio, G compares the level of informative lightness $I_L - d$ (signal) to the level of dark current d (noise) within the specular highlight pixels $\mathcal{S}$ for a given exposure time:

$$G(T_{\text{exp}}) = \underset{p \in \mathcal{S}(T_{\text{exp}})}{\text{median}} \left(\frac{I_L(p, T_{\text{exp}}) - d(T_{\text{exp}})}{d(T_{\text{exp}})} \right) \tag{3.11}$$

Averaging the illuminant estimates across multiple low-exposure images with the same T_{exp} did not result in any performance improvement. Based on these findings, we recommend capturing a total of 9 low-exposure images (at 5 ms, 10 ms, 20 ms, 40 ms, 60 ms, 80 ms, 100 ms, 125 ms and 150 ms) and then selecting the low-exposure image that achieves the highest G. With a total acquisition time of about 0.6 s, we believe that capturing the sequence of low-exposure images has a negligible impact on the surgical workflow.

Impact of Illuminant Shifts on Oxygen Saturation Estimation We divided the simulated tissue spectra with corresponding reference StO_2 values into a training set of 10 000 spectra, and a test set of the remaining 5000 spectra. This configuration was used to train and validate an ML model for StO_2 estimation for each illuminant spectrum, maintaining a consistent data split across all illuminants. Following the approach by Wirkert et al., we used a random forest regression model, with parameters as specified in [366, 367].

For each light source L_i, we compared the StO_2 estimation error across 3 scenarios:

1. **Reference illuminant spectrum:** We trained a random forest regressor using the reference illuminant spectrum L_i^{ref}.
2. **Mismatched illuminant spectrum:** We trained 4 mismatched random forest regressors, each using the illuminant spectrum from a different light source L_j ($j \neq i$).
3. **Our approach:** We trained random forest regressors using the illuminant spectrum L_i^{est}, estimated from the ex vivo liver recordings with our approach. To assess the impact of different camera poses on the accuracy of functional parameter estimates, we computed one illuminant estimate per camera pose, resulting in a total of 8 random forest regressors.

Adapting the simulated spectra to a specific (estimated) illuminant spectrum and training the random forest regressor requires less than 50 ms in total, enabling real-time recalibration of the model during surgery.

Performance Ranking Against State-Of-The-Art Illuminant Estimation Approaches We reviewed state-of-the-art methods for illuminant estimation and selected 4 methods that do not require supervised training and are potentially applicable to surgical MSI data: Max-RGB, Gray-world, Shades-of-gray, and Gray-edge (cf. Section 3.1). To ensure a fair comparison, unaffected by the optimized exposure time in our approach, we evaluated the performance of these state-of-the-art approaches across a range of exposure times using our ex vivo liver recordings.

Following the ranking and stability assessment guidelines in [364], we computed performance rankings based on the cosine similarity between estimated and reference illuminant spectra. We calculated the rank for each illuminant estimation method across 1000 bootstrap samples, each comprising 5 light source-level cosine similarity scores randomly selected without replacement. The light source-level cosine similarity scores were obtained by averaging cosine similarities across all camera poses for a given light source.

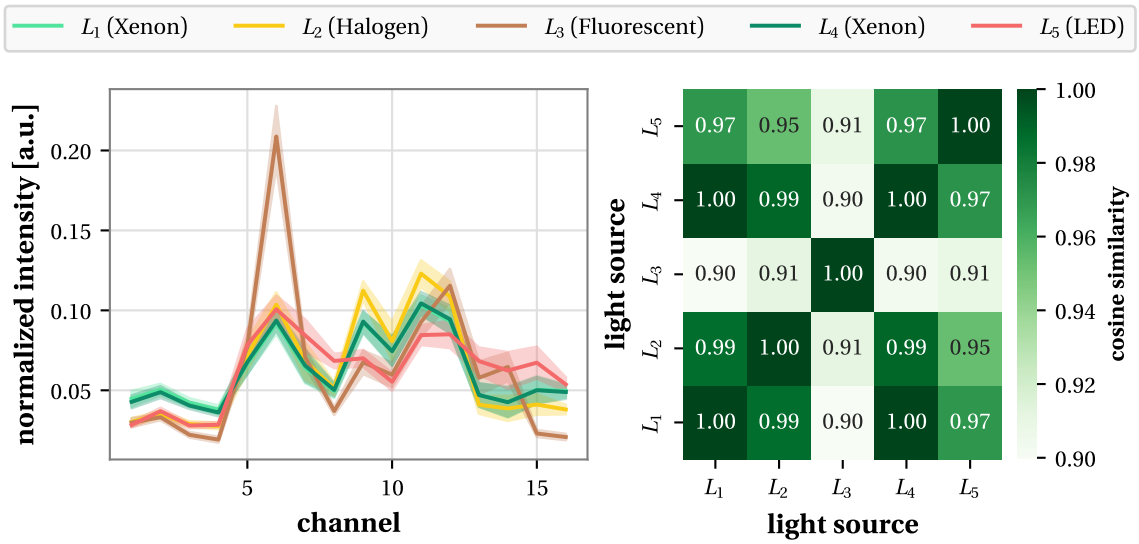


Figure 3.2: Light sources used in our study. The ℓ^1-normalized reference intensity spectra of the 5 light sources L_1 to L_5 utilized in our study are displayed on the left. These spectra were hierarchically aggregated from 100 MSI cubes of a white reference standard, with the shaded areas representing 0.1 times the standard deviation. On the right, the cosine similarity between the light source spectra is illustrated. Figure adapted from [24, 26].

3.3.2 Automated Illuminant Estimation From Specular Highlights

The primary objective of this study was to evaluate whether our specular reflection-based approach can accurately estimate illuminant spectra during surgery (RQ1.1), to analyze the impact of errors in illuminant estimation on StO_2 accuracy (RQ1.2), and to compare our method with state-of-the-art model-based illuminant estimation approaches (RQ1.3).

Accuracy of our Illuminant Estimation Approach To assess the performance of our illuminant estimation approach, we estimated the illuminant spectrum from the ex vivo liver recordings for each of the 5 light sources, L_1 to L_5, and across all camera poses, from D1_N to D2_W. An overview of the reference illuminant spectra is available in Figure 3.2. The principal component analysis projection of the reference and estimated illuminant spectra is displayed in Figure 3.3. For most light sources, distinct clusters can be observed, with the estimated illuminant spectra closely matching the corresponding reference illuminant spectra. Only for L_1 and L_4, which are both xenon light sources that only differ by the manufacturer, there is substantial overlap between the projected illuminants. In fact, Figure 3.2 demonstrates that the spectra of L_1 and L_4 are nearly identical, with a cosine similarity of $S_c \approx 1$.

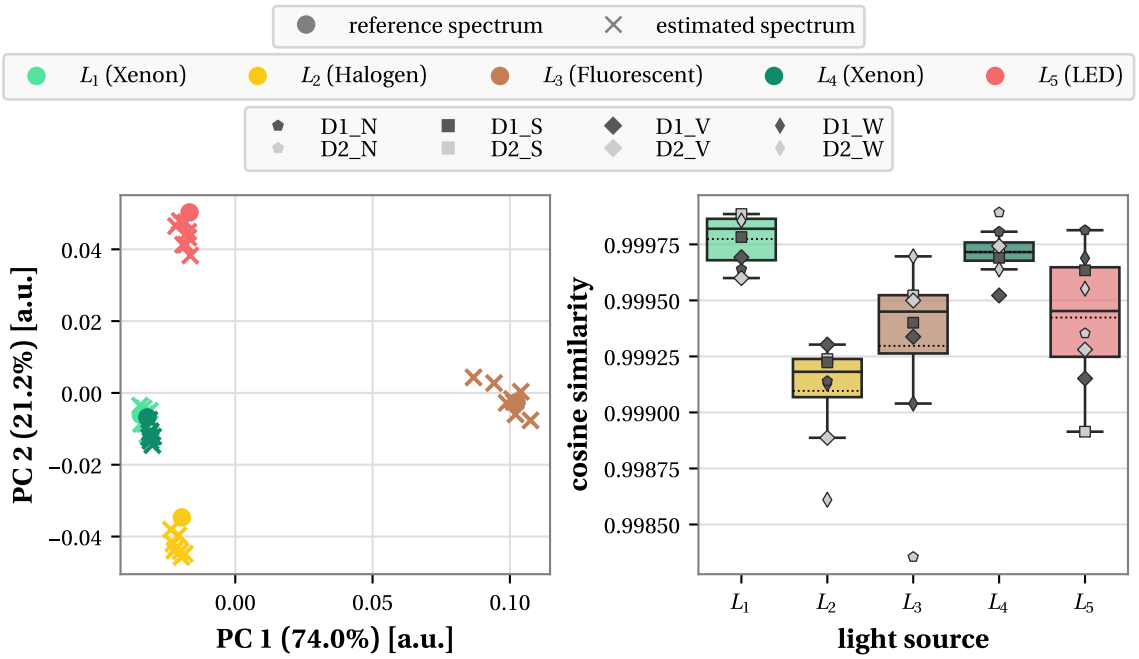


Figure 3.3: Performance of our illuminant estimation approach on ex vivo porcine liver images. The left figure shows the projection of the reference spectra (circles) and corresponding estimated spectra (crosses) for 8 different camera poses, D1_N to D2_W, and 5 different light sources, L_1 to L_5, onto the first two principal components (PC 1 and PC 2) computed from principal component analysis of the reference spectra. The explained variances for each principal component are indicated in brackets. The right figure displays the distribution of cosine similarity between the estimated and reference spectra for each light source. The boxplots represent the quartiles of the distribution across camera poses, with whiskers showing the range excluding outliers. The median is shown as a solid line, the mean as a dotted line, and the markers correspond to specific camera poses. Figure adapted from [24, 26].

We quantitatively compared our estimated illuminant spectra with the corresponding reference spectra by calculating their cosine similarity. The distribution of these cosine similarities for each light source is shown in Figure 3.3. Our method consistently achieved a cosine similarity above 0.998 across all light sources and camera poses. The performance on the test light sources L_4 and L_5 was comparable to that on the validation light sources L_1 to L_3, demonstrating that our approach generalizes effectively to previously unseen light sources.

Impact of Illuminant Estimation Errors on Oxygen Saturation Estimation As illustrated in Figure 3.4, substantial StO_2 estimation errors, up to 38.4 %, occur when a mismatched illuminant spectrum is used for estimation. In contrast, our approach effectively compensates for illuminant shifts across both validation and test light sources, as well as

camera poses. A reduction in the average StO_2 estimation error from 22.0 % (standard deviation (SD) 8.0 %) to 13.5 % (SD 2.5 %) is achieved using our approach. For comparison, the average StO_2 estimation error using the reference illuminant spectrum is 11.3 % (SD 0.8 %).

Figure 3.5 provides a qualitative validation of our approach using in vivo human lip recordings. The average StO_2 within a region of interest is shown for a continuous video stream of the lips, with an illuminant shift occurring between frames 85 and 100. When assuming constant illumination, this shift causes a noticeable drop in the average estimated StO_2 from 89.9 % before the shift to 82.4 % after the shift. In contrast, our approach effectively compensates for the illuminant change, maintaining a more stable StO_2 estimate, with an average StO_2 of 91.1 % after the shift.

Comparison to State-Of-The-Art Illuminant Estimation Approaches To evaluate the performance of our approach against state-of-the-art model-based illuminant estimation methods, we compared the cosine similarity between the estimated and reference illuminant spectra for each method, using ex vivo porcine liver data captured at 3 different exposure times. Figure 3.6 displays the distributions of the cosine similarities across the 5 light sources. Our approach consistently outperformed the competing state-of-the-art methods, with the Max-RGB and Gray-world approaches showing the poorest performance.

Figure 3.7 shows the cosine similarity-based ranking of our illuminant estimation approach compared to the state-of-the-art methods. Across all 3 exposure times, our approach consistently ranks first, followed by the Gray-edge and Shades-of-gray approaches in second and third rank, respectively. Overall, the ranking variability is higher for the state-of-the-art methods compared to our approach.

3.4 Discussion and Conclusion

This work presents the first approach to automated illuminant estimation for surgical SI data. Through a combination of in silico, ex vivo, and in vivo experiments, we identified the following key findings:

1. **Illuminant estimation based on specular highlights:** Our experiments confirm our hypothesis that specular highlight spectra from low-exposure MSI data are an accurate estimate of the illuminant spectrum in surgical scenes. Our approach demonstrates consistent performance across diverse light sources and camera poses, showcasing its robustness to variations in imaging and illumination conditions.

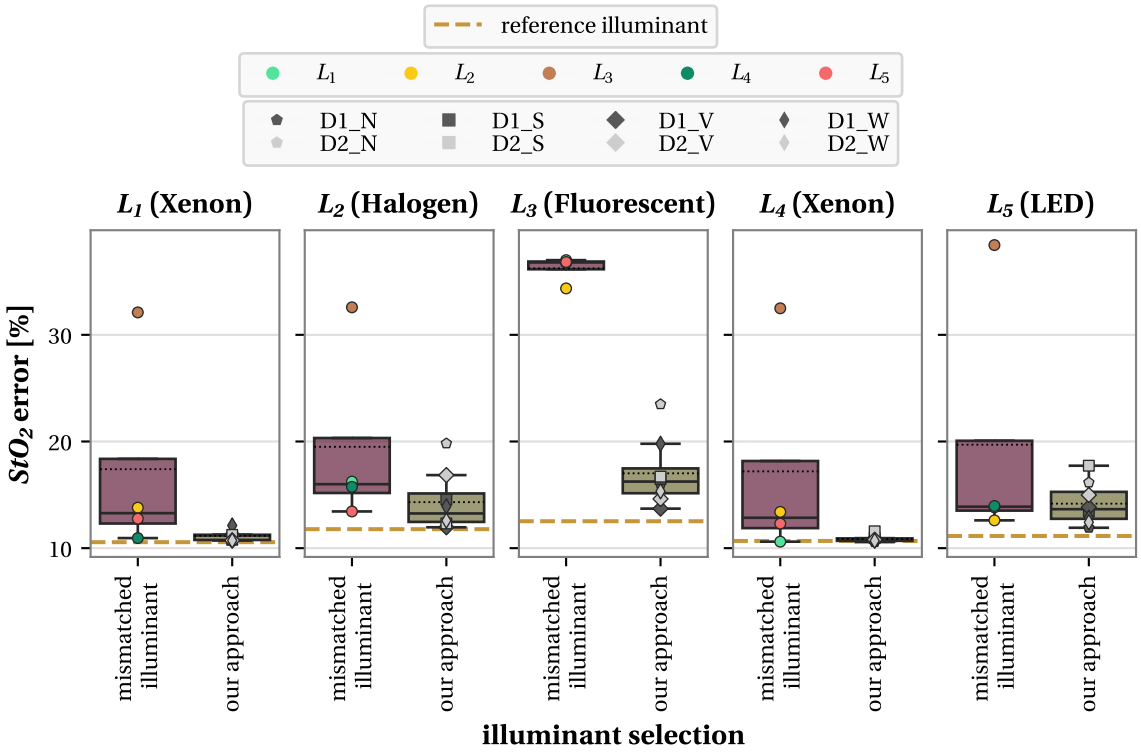


Figure 3.4: Error in the tissue oxygen saturation (StO_2) estimation using our approach compared to using a mismatched illuminant spectrum. The error in StO_2 estimation is presented for 3 scenarios: (1) using the reference illuminant (dashed lines), (2) using a mismatched illuminant, and (3) using our approach to estimate the illuminant spectrum. For a given light source L_i, the distribution of StO_2 errors is hierarchically aggregated over all other light sources $\{L_j\}_{j \neq i}$ in the mismatched illuminant scenario and over the 8 camera poses, D1_N to D2_W, for our approach. Markers denote different light sources and camera poses. The boxplots depict the quartiles of the distribution, with whiskers showing the range excluding outliers, the median indicated by a solid line, and the mean by a dotted line.

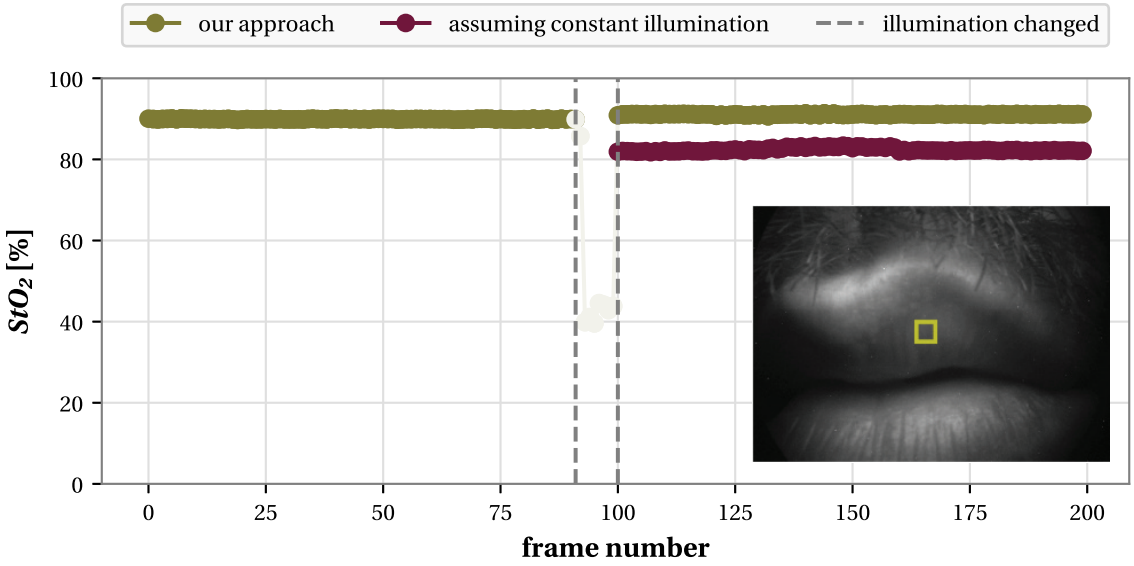


Figure 3.5: Qualitative validation of our illuminant estimation approach on in vivo human lips. A continuous video stream of a human subject's lips was recorded, with an illuminant shift occurring between frames 85 and 100. When constant illumination is assumed, this shift causes a drop in the estimated tissue oxygen saturation (StO_2) within the region of interest (highlighted by the yellow square). In contrast, our approach effectively compensates for the illuminant shift, maintaining accurate StO_2 estimates. Figure adapted from [24, 26].

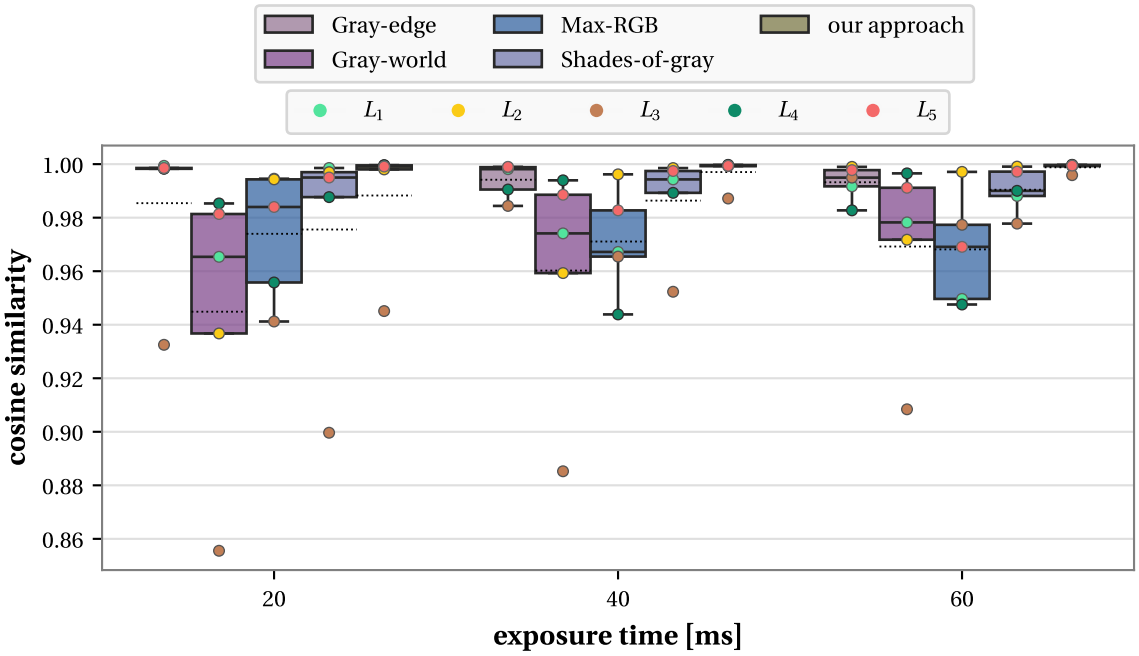


Figure 3.6: Performance of our illuminant estimation approach compared to state-of-the-art color constancy methods on ex vivo porcine liver images. State-of-the-art color constancy methods include the Gray-edge, Shades-of-gray, Gray-world, and Max-RGB approaches. Distributions of the cosine similarity between estimated and reference illuminants are shown for 3 different exposure times, and across 5 distinct light sources, L_1 to L_5, with each light source represented by different markers. The boxplots illustrate the quartiles of the distribution, with whiskers showing the range excluding outliers. The median is marked by a solid line, and the mean by a dotted line. Figure inspired from [24, 26].

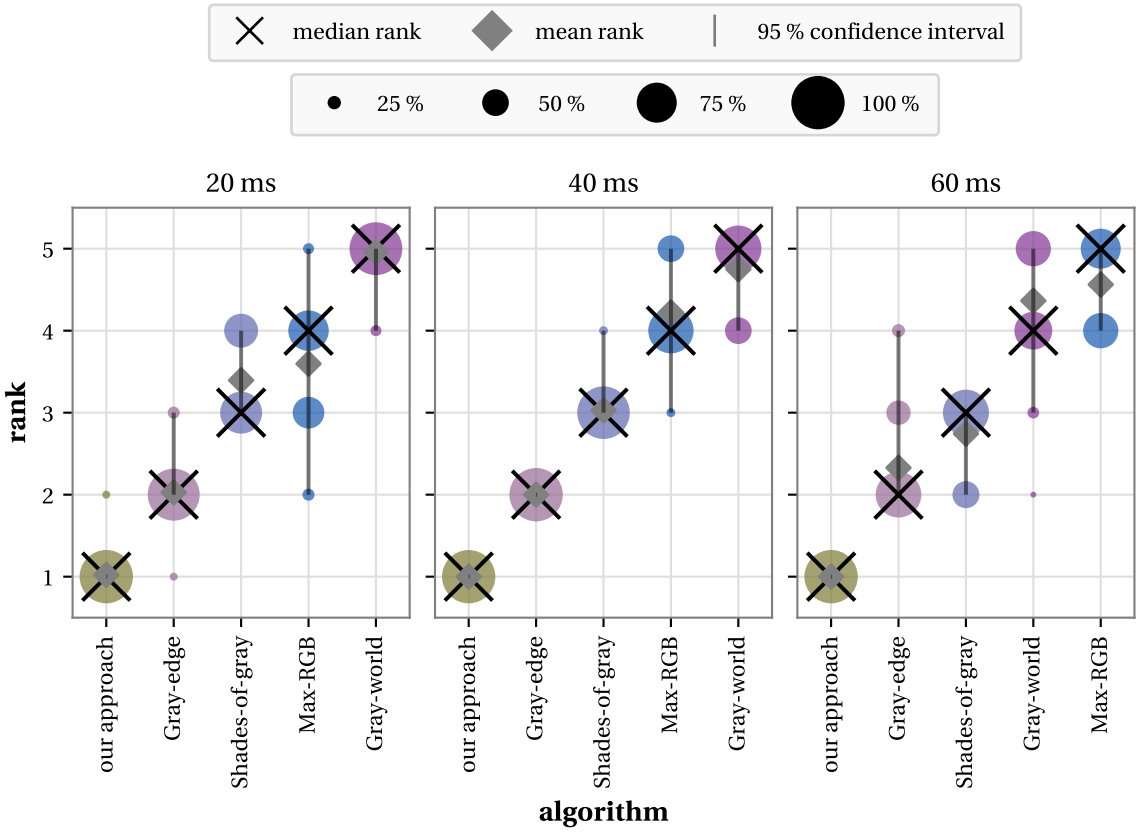


Figure 3.7: Ranking stability of our illuminant estimation approach compared to state-of-the-art color constancy methods on ex vivo porcine liver images. State-of-the-art color constancy methods include the Gray-edge, Shades-of-gray, Gray-world, and Max-RGB approaches. The ranking stability based on bootstrap sampling of the cosine similarity between reference and estimated illuminants is shown. For each blob at position (M, rank r), its area is proportional to the frequency of illuminant estimation method M achieving rank r across 1000 bootstrap samples. Each sample consists of 5 light source-level cosine similarity scores (concept adapted from [364]). For each method, black crosses indicate the median rank, gray diamonds show the mean rank, and gray lines represent the 95 % quantile of the bootstrap results. Figure adapted from [24, 26].

2. **Impact of illuminant shifts on functional parameter estimation:** Illuminant shifts can lead to substantial errors in StO_2 estimation, with deviations reaching up to 38.4 %. This presents an important challenge, particularly when precise StO_2 measurements are crucial for informed intraoperative decision-making. Our approach effectively addresses this issue, delivering accuracy close to the ideal scenario in which the illuminant spectrum is perfectly known.
3. **Comparison to state-of-the-art methods:** Our approach outperforms state-of-the-art model-based illuminant estimation methods, consistently achieving the first rank across a variety of light sources and exposure times.

The following sections highlight the key strengths of our approach (Section 3.4.1), discuss its limitations, and provide an outlook on future research directions and recent developments that build on our findings (Section 3.4.2), followed by a summary of our contributions (Section 3.4.3).

3.4.1 Key Strengths of Our Approach

A key strength of our approach lies in its ability to generalize effectively without requiring extensive training data, unlike learning-based methods. It performs robustly across unseen light sources, camera poses, and specimens, as demonstrated with liver and human lips. This adaptability is particularly valuable in surgical scenarios, where imaging setups and scene content can vary widely across devices and procedures. Additionally, the method is computationally efficient, with a total acquisition time of approximately 0.6 s for capturing low-exposure images, and below 50 ms for training a matching StO_2 regressor. This efficiency makes it well-suited for real-time applications in the OR, where rapid decision-making is essential.

3.4.2 Limitations and Future Work

While our approach represents a substantial advancement in automated light source calibration and live functional imaging during open surgery, several potential limitations should be addressed to fully realize its capabilities:

- **Nonuniform illumination:** The proposed method assumes uniform illumination across the entire image, which may not be the case in all surgical scenarios or imaging setups (e.g., multiple light sources, short distances between light sources and the surgical scene, or large field of views). Future work could investigate more advanced methods that account for spatially varying illumination, such as estimating an illuminant spectrum for each pixel.

- **Presence of specular highlights:** Our method depends on the presence of specular highlight pixels in the image. While specular highlights are generally present in surgical scenes, this reliance may limit the applicability of our approach to other types of tissue or imaging setups where specular reflections are less prominent (e.g., skin imaging or hardware setups designed to suppress specularly reflected light).
- **Limited validation:** While our approach has demonstrated its potential using two different specimens (ex vivo porcine liver and in vivo human lips), clinical validation in real open surgeries involving a broader variety of surgical tissue types is essential. Additionally, while we covered all types of light sources that commonly occur in an OR (LED, xenon, halogen, and fluorescent light), testing with a wider range of light source configurations – such as combinations of different light sources or specific setups like forehead-mounted torch lamps and surgical overhead lights of different manufacturers – would provide valuable insights. Due to the limited availability of SI devices during our study, our experiments were conducted with a single MSI setup. Further validation across other SI devices, particularly those with different spectral channels would be highly beneficial.
- **Impact on the surgical workflow:** Our method requires acquiring a separate set of low-exposure images for illuminant estimation, which, although quick (under 1 s), could disrupt the clinical workflow if frequently repeated. To minimize this impact, we propose developing a method for illumination change detection, allowing low-exposure images to be captured only when needed. Given the superior accuracy of our method compared to existing approaches, we believe the occasional acquisition of low-exposure images is a reasonable trade-off. However, future work could explore alternative approaches that enable illuminant spectrum estimation directly from the MSI data used for functional parameter estimation, eliminating the need for additional images.

In 2024, we addressed these limitations by building upon the presented work and developing a novel method for estimating the illuminant spectrum directly from the SI data, eliminating the need for additional low-exposure images. This new approach also enables pixel-wise illuminant predictions [34]. It represents the first DL-based method for automated illuminant estimation in surgical SI data, made possible by our extensive collection of in vivo porcine surgical SI datasets (cf. Chapter 5 and Chapter 6) over the past few years. These datasets were acquired using the novel medical-grade HSI system TIVITA® described in Section 2.1.2. This system offers a larger field of view compared to the MSI setup (approximately 20 cm×30 cm vs. up to approximately 3 cm× 5.6 cm). Although a direct quantitative comparison between the two methods was not possible due to the inability to capture low-exposure images with the HSI device, a comparison of pixel-wise illuminant estimates (which capture spatial inhomogeneities

in the illuminant spectrum) and image-wide average illuminant estimates (assuming uniform illumination) showed that pixel-wise estimations lead to improved functional parameter estimates in devices with larger fields of view.

3.4.3 Conclusion

In conclusion, we have shown that low-exposure MSI data is highly effective for recovering the illuminant through specular highlight analysis, particularly in scenarios where the illumination across the field of view is nearly uniform. This study marks an important first step toward enabling real-time functional imaging in open surgery.

4 Hardware-Related Sources of Variation in Hyperspectral Imaging

In the previous chapter, we examined the impact of illuminant shifts on SI measurements and derived functional tissue parameter values (cf. Chapter 3). While we demonstrated that such environmental factors can influence the accuracy of SI measurements, an equally important consideration for the clinical translation of SI is the reliability of the imaging devices themselves.

In this chapter, we focus on TIVITA® cameras (Diaspective Vision GmbH, Am Salzhaff, Germany). These devices are increasingly adopted in clinical studies due to their medical device certification, which facilitates regulatory approval, and their ease of use, enabling operation by clinical staff. In practice, different instances and generations of the same device type are deployed within and across studies, raising important questions: Do systematic shifts occur between devices? Are measurements stable over time, or are they affected by hardware-related factors such as shifts in sensor temperature or calibration drifts? This chapter provides a systematic investigation of hardware-related sources of variation in HSI measurements. Based on our findings, we propose strategies to mitigate these sources of variation in HSI study design. Following these guidelines supports unbiased data acquisition and the development of reliable, generalizable algorithms, for example for automated surgical scene segmentation (Part III), as well as automated sepsis diagnosis and mortality prediction in the ICU (Part IV).

Section 4.1 provides an overview of related work on hardware-related sources of variation in HSI devices, underscoring the need for systematic investigations in this area. Section 4.2 provides details on the devices and datasets specifically acquired for this study. The experimental setup and results are presented in Section 4.3. The chapter concludes with recommendations for HSI study design, along with a discussion of strengths, limitations, and directions for future research in Section 4.4.

Parts of the research described in this chapter were previously presented at the Institute of Electrical and Electronics Engineers (IEEE)'s 13th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS) in 2023 [310].

4.1 Related Work

An estimated 257 publications have reported the use of different generations and instances of the TIVITA® devices[1]. Despite this growing body of literature on TIVITA®-based clinical studies, only a few works have examined sources of variation in measured spectra and functional tissue parameter index values. Existing studies have primarily investigated subject-related factors such as variations in skin tone, age, and sex [258] and imaging conditions, including illuminant changes [34] (cf. our work in Chapter 3) and imaging geometries [329]. To date, however, hardware-related sources of variation – such as device shifts and temporal drifts – have not been systematically investigated.

Device Shifts In the broader field of medical imaging, the effects of device shifts have been investigated across multiple imaging modalities (e.g., magnetic resonance imaging, radiography, mammography), showing that such shifts can lead to critical clinical errors, including misdiagnosis by ML models due to shortcut learning and poor generalizability across devices [28, 374, 292]. We hypothesize that, owing to manufacturing and calibration tolerances, data shifts between different generations and instances of the same HSI device type are highly likely. However, the magnitude of these shifts and their impact on measured spectra and derived functional tissue parameter indices remains unknown.

Temporal Stability Many HSI studies compare data acquired at different time points. Examples include assessing the impact of an intervention over time or detecting tissue malperfusion [83, 289, 96, 326, 227]. In several cohort comparison studies, data are collected months or even years apart [81, 327]. Such analyses assume that measurements remain stable over time and that devices do not exhibit measurement shifts – neither in the short term (minutes to hours) nor in the long term (months to years).

However, in HSI devices, measurement shifts over time can arise from several sources, such as thermal effects: In Complementary Metal-Oxide-Semiconductor (CMOS) imaging sensors, which are used in TIVITA® devices, increasing sensor temperature leads to an exponential rise in dark current due to thermal electron generation [1] and a deterioration of linearity [354]. Push-broom HSI imagers, including the TIVITA® cameras, capture a vertical line of the scene dispersed through a grating while scanning across the other spatial dimension (see Section 2.1.3). In such systems, wavelength calibration can drift over time due to misalignment between the slit, grating, and imaging optics caused by temperature variations or mechanical stress [75]. In spectroscopy, further

[1] Data retrieved from the Digital Science Dimensions platform app.dimensions.ai using the search terms “hyperspectral imaging” and “TIVITA” and “Diaspective Vision” in the full-paper dataset on July 30th, 2025.

measurement shifts have been observed due to component aging, such as soiling of optical surfaces or changes in illuminant spectra [232]. Since push-broom HSI devices contain a spectrometer unit, these effects likely occur in TIVITA® cameras as well.

Many of these intra-device shifts could be mitigated by recalibrating the device (cf. Section 2.1.3). However, the TIVITA® manufacturer does not recommend regular calibration, stating: "No, it is not necessary to calibrate the TIVITA® Tissue. The TIVITA® Tissue system is calibrated during production and the calibration data are saved within the camera" [120]. This approach overlooks the possibility of hardware-related drifts over time. To this end, a systematic investigation of temporal stability in HSI devices, including a comparison of calibration strategies, is urgently needed.

To address the lack of research on hardware-related variation, including device shifts and the temporal stability of devices over both short and long timescales, we address the following research questions:

- RQ1.4: How do spectra and functional tissue parameter index values vary across different generations and instances of TIVITA® cameras?
- RQ1.5: How stable are TIVITA® measurements over short timescales (up to hours) in terms of spectral accuracy and functional tissue parameter index shifts?
- RQ1.6: Do long-term drifts occur in spectra and functional tissue parameter indices, and how do calibration strategies (e.g., single vs. daily calibration) affect them?

4.2 Materials and Methods

This section outlines our HSI devices (Section 4.2.1), the experimental setup (Section 4.2.2), and the resulting datasets used to address our research questions (Section 4.2.3).

4.2.1 Hyperspectral Imaging Devices

Two generations of TIVITA® cameras for extracorporeal imaging are currently in clinical use. The first generation includes the TIVITA® Tissue and TIVITA® Wound editions, while the second generation comprises TIVITA® 2.0, TIVITA® 2.0 Surgery, and TIVITA® 2.0 Wound. All devices share the same spectral specifications, with a spectral resoultion of approximately 5 nm and 100 spectral channels spanning 500 nm to 1000 nm. The resulting HSI cubes measure $640 \times 480 \times 100$ (width $\times$ height $\times$ spectral channels). Editions differ mainly in application-specific add-ons (e.g., the Surgery edition features

an elongated cardanic mount, an additional monitor, and a sterile handle adapter), whereas more substantial differences exist between the first and second generation, including:

- **Optical Components**: The field of view differs between the two TIVITA® generations as different objective lenses were installed. The second-generation devices have a fixed lens resulting in a field of view of approximately 16 cm × 11.5 cm at the recommended measurement distance of 50 cm. In the first-generation devices, the user can choose from a set of lenses with focal lengths of 8 mm, 12 mm, 16 mm, 25 mm, 50 mm and 75 mm. Since our focus is on analyzing spectral shifts across devices, we aim to minimize geometric differences from varying fields of view. To this end, we used a 12 mm lens for the first-generation devices, resulting in a field of approximately 18.5 cm × 13.5 cm, closely matching that of the second-generation devices. The optical path of the second-generation devices differs from the first generation through the addition of an RGB sensor and a semi-reflective mirror that splits the incoming light between both sensors.
- **Illumination**: The first-generation TIVITA® cameras use a halogen light source, while the second-generation devices feature a broadband LED light source. Figure 4.1 shows spectra from white reference standard images captured with two devices from each generation, reflecting the combined effects of illumination and the transmission characteristics of optical components.
- **Camera Housing**: Both generations of the TIVITA® cameras differ in the mechanical design of the camera head. As illustrated in Figure 4.2, the second-generation devices feature a more compact housing of the camera head, with the camera sensors and LED lightning unit being in proximity and covered by a glass window.

4.2.2 Experimental Setup

Our experiments were conducted using two device instances from each generation: the TIVITA® Tissue (halogen illumination) and the TIVITA® 2.0 Surgery (LED illumination). The devices are referred to as Halogen1, Halogen2, LED1, and LED2.

To investigate the effects of hardware-related sources of variation, we conducted experiments on both phantoms and in vivo human skin (cf. Figure 4.2). Phantoms have the advantage of possessing stable optical characteristics over extended periods of time such that shortcomings in accuracy and shifts in measurements can be immediately related to the measurement hardware. In vivo skin measurements, by contrast, allow for the assessment of hardware-related sources of variation on the spectra and resulting functional tissue parameter shifts in a realistic setting, closely reflecting the devices' intended clinical applications, such as automated sepsis diagnosis and mortality prediction (cf. Chapter 7).

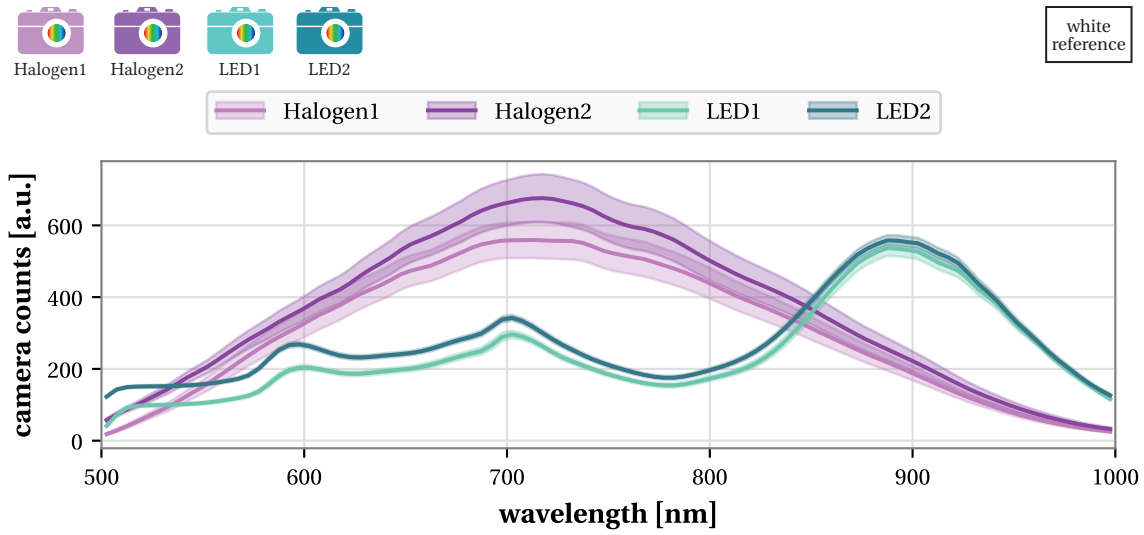


Figure 4.1: Comparison of halogen and light-emitting diode (LED) illumination in TIVITA® devices. Over one month, images of a white reference standard were captured on different days using two TIVITA® Tissue devices (halogen illumination: cameras Halogen1 and Halogen2) and two TIVITA® 2.0 Surgery devices (light-emitting diode (LED) illumination: cameras LED1 and LED2). Pixel spectra were aggregated at the image level. The average spectra across images are shown as solid lines, with shaded areas representing one standard deviation.

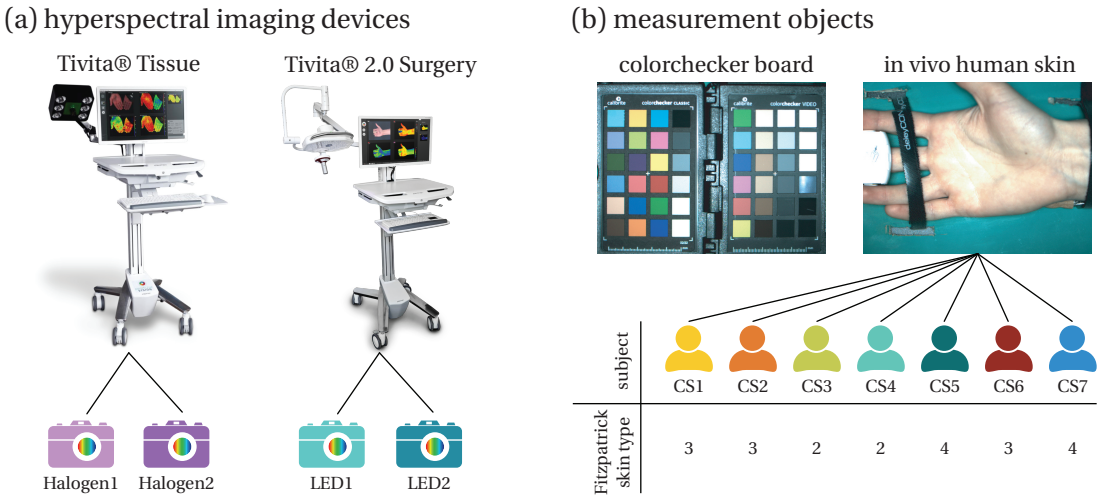


Figure 4.2: Experimental setup. (a) Experiments were conducted using two generations of TIVITA® systems: two TIVITA® Tissue devices with halogen illumination (Halogen1, Halogen2) and two TIVITA® 2.0 Surgery devices with light-emitting diode (LED) illumination (LED1, LED2). (b) Measurements were performed on a colorchecker board phantom and on the palm skin of 7 healthy volunteers (proband CS1 to CS7). The colorchecker phantom enabled reproducible assessment of device accuracy, while the human skin measurements allowed evaluation of functional tissue parameter shifts in a realistic clinical scenario, relevant to applications such as automated sepsis diagnosis and mortality prediction.

Phantom Measurements Generating tissue-mimicking phantoms for spectral imaging that closely replicate optical characteristics of biological tissues, is an active area of research. To date, reproducible and durable phantoms that accurately reflect tissue optical properties across the entire wavelength range of 500–1000 nm covered by TIVITA® devices are still lacking [337]. To this end, we utilized colorchecker board phantoms as they offer a standardized and reproducible reference that is widely used in the technical validation of MSI and HSI devices. We used a combination of the calibrite® ColorChecker Classic Mini with the calibrite® ColorChecker Passport Video (Calibrite LLC, Wilmington, USA). A sample RGB image of our combined colorchecker board is shown in Figure 4.2. It contains 48 color fields, including a white color field, 8 color fields with different skin tones and 3 color fields with different flesh tones. To rule out potential manufacturing variation between different colorchecker boards, the same colorchecker board was utilized across all measurements.

Proband Measurements Upon approval by the Ethics Committee of the Medical Faculty of Heidelberg University, Germany (study reference number: S-530/2020), we conducted measurements on skin of palm for healthy, adult volunteers. The study was performed in compliance with the Declaration of Helsinki and its subsequent revisions, and all 7 included probands, referred to as probands CS1 to CS7 in the following, provided written informed consent prior to study participation. Our study population covered Fitzpatrick skin types in the range 2 to 4 (cf. Figure 4.2 for more details). Of the participants, 4 were female, and 3 male.

To reduce inter-proband variability and avoid biases when comparing measurements of the same proband taken at different timepoints, several measures where taken: The right hand was used across all measurements of a proband, and the probands were asked to refrain from applying any creams or lotions to their hands on the day of the measurement. Probands were asked to clean their hands with lukewarm water prior to the measurements. All measurements were conducted with the proband resting in an upright seated position. After the proband was seated, a waiting period of 5 minutes was observed before starting the measurements to minimize fluctuations in skin temperature and circulation caused by prior physical activity. To ensure that the probands were in a stable physiological state, heart rate, respiratory frequency and pulse oxymetrical oxygen saturation (SpO_2) were monitored throughout the HSI measurements using a pulse oximeter mounted on the tip of the digitus medius of the right hand (Masimo MightySat Rx, Masimo Corporation, Irvine, USA). The probands were instructed to rest their arm on the measurement table with the palm facing upwards. The hands were kept in a standardized, relaxed but motionless position during the measurements, supported through mounting the hand on a custom frame placed on the table (cf. Figure 4.2 for an example image). The mounting frame further ensured a consistent background across all images.

Hyperspectral Image Acquisition Several measures were taken to control environmental factors and reduce their impact on measurements: All light sources other than the TIVITA® device were turned off, and all window blinds were closed to prevent straylight. The measurements were conducted in a temperature-controlled room with a constant temperature of 19 °C, which is within the range of operation room temperature of 0 °C to 30 °C recommended by the manufacturer [120]. The devices were operated at an imaging distance of approximately 50 cm, ensured by an integrated distance calibration system. The positioning of colorchecker boards and hands within the field of view was standardized, with both the board and the palm centered in the field of view and oriented consistently.

Hyperspectral Image Annotation In the first stage, annotations were generated automatically. For colorchecker board images, a 8×6 grid of patches with size 23 px $\times$ 23 px was fitted to annotate the 48 individual color fields. For human skin images, a DL-based segmentation algorithm was used to delineate the skin [308, 314]. All automatically generated annotations were then manually verified and corrected as needed to ensure consistency across images.

Reference Spectral Measurements To assess the accuracy of the TIVITA® measurements, we compared them to reference measurements taken with a spectrometer (HR2000+, Ocean Insight (formerly Ocean Optics) Orlando, Florida, USA) equipped with a Tungsten Halogen lightsource (HL-2000, Ocean Insight, Orlando, Florida, USA). The spectrometer captures 1131 spectral channels in the range of 200 nm to 1100 nm. Spectrometer measurements were repeated 100 times for each color field of the colorchecker board phantom, with the resulting spectra averaged to obtain a single reference spectrum per color field.

Data Preprocessing Calibration with dark and white reference spectra / images was performed for all spectrometer and HSI measurements (cf. Section 2.1.3). Spectrometer measurements were transformed to the same spectral range as the TIVITA® measurements based on the spectral sensitivities of the individual channels presented in Figure 2.6. To avoid shifts from slight fluctuations in the measurement distance, ℓ^1-normalization was applied across the spectral dimension. The tissue parameter index images StO_2, tissue perfusion index (NPI), tissue hemoglobin index (THI), and tissue water index (TWI) were derived from the HSI cubes according to the formulas described in [141].

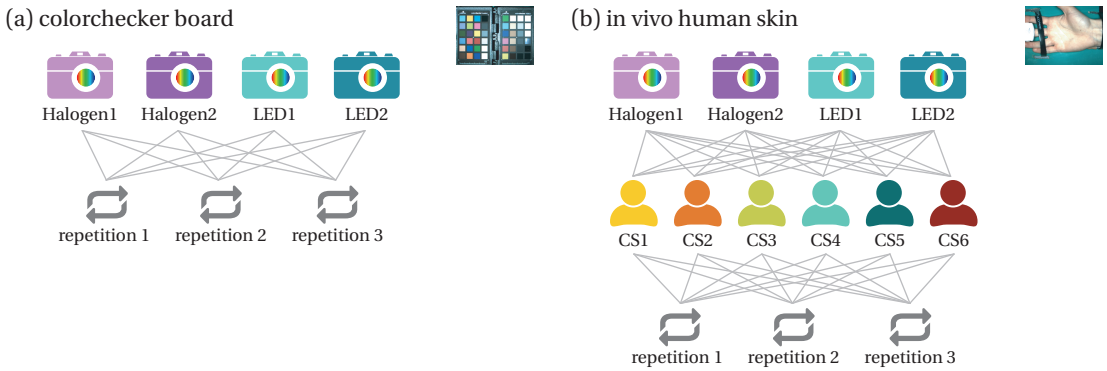


Figure 4.3: Device shift experiments. For all 4 devices Halogen1, Halogen2, LED1 and LED2, measurements were conducted on (a) a colorchecker board phantom and (b) on the palm skin of 6 healthy volunteers (proband CS1 to CS6). Each measurement object was measured 3 times per device, with the devices used in a fixed rotating order to minimize potential device-related heat-up effects and to reduce the risk of introducing bias from changes in the proband's physiological condition over the course of the measurement session.

4.2.3 Datasets

To study the impact of device shifts, as well as the sort-term and long-term temporal stability of the devices, we acquired a total of 11 028 images, constituting the following datasets:

Device Shifts As illustrated in Figure 4.3, we conducted measurements on colorchecker boards and palm skin of 6 probands (CS1 to CS6) using all 4 devices Halogen1, Halogen2, LED1 and LED2, with each measurement object measured 3 times per device. The devices were used in a fixed rotating order (Halogen1, Halogen2, LED1, LED2) which was repeated 3 times per object. This approach minimized potential device-related heat-up effects by allowing sufficient cooling time between uses. It also reduced the risk of bias from changes in the proband's physiological state by evenly distributing measurements from different devices throughout the session. As the 7 measurement sessions for the colorchecker board phantom and all 6 probands could not be performed on the same day, we ensured consistent conditions across sessions. Prior to each measurement session, devices were stored in the measurement room for at least 1 h to equilibrate to room temperature, and dark and white calibration was performed for each device.

Short-Term Temporal Stability To investigate the short-term temporal stability of the devices, we conducted series of measurements on colorchecker boards and human

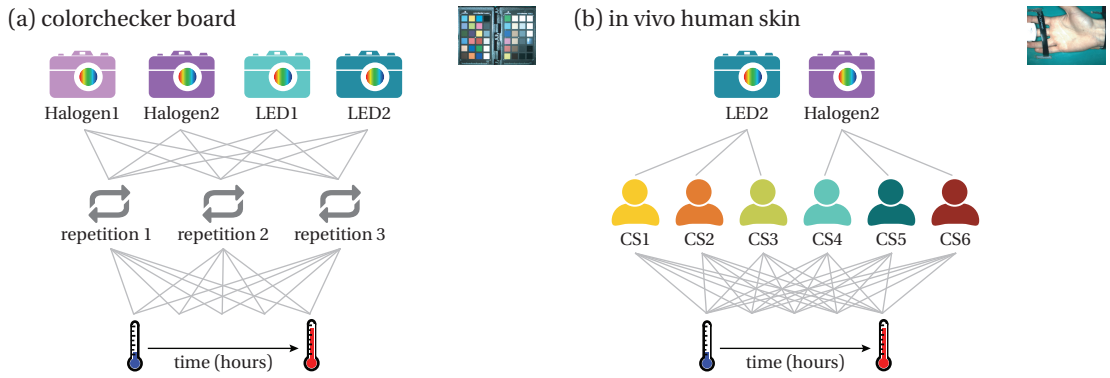


Figure 4.4: Experiments to study short-term temporal stability. Measurement series lasting approximately 2 h were conducted on both a colorchecker board and human palm skin, with measurements taken approximately every 30 s and device sensor temperature recorded throughout. (a) For the colorchecker board, each series was repeated 3 times for all 4 devices Halogen1, Halogen2, LED1 and LED2. (b) Human skin measurements were performed using the devices Halogen2 and LED2, with 3 probands per device and no repetitions.

palm skin. Each series lasted approximately 2 h, with measurements taken at an approximately 30 s interval[2]. During the measurements, the device's sensor temperature was recorded. Before each measurement series, devices were placed in the measurement room and left unused for at least 1 h to ensure they had equilibrated to ambient temperature. Also, dark and white calibration was performed prior to each measurement series. As shown in Figure 4.4, measurements on the colorchecker boards were repeated 3 times and conducted with all 4 devices Halogen1, Halogen2, LED1, and LED2. In contrast, due to the time demands on participants, human skin measurements were only conducted with the devices Halogen2 and LED2, using 3 probands per device, and without performing repetitions.

Long-Term Temporal Stability In our device shift and short-term temporal stability experiments, we recorded new dark and white reference images before each measurement series to recalibrate the TIVITA® devices. This calibration strategy is meant to account for potential long-term drifts in device performance, however, it differs from the manufacturer's recommended procedure, which suggests using the dark and white calibration files recorded during production [120].

To assess long-term device shifts in TIVITA® cameras, we conducted a multi-month measurement campaign imaging a colorchecker board phantom and the palm skin

[2]The 30 s measurement interval was generally maintained, though occasional device freezes required restarting before continuing the measurements, thereby extending the interval between measurements.

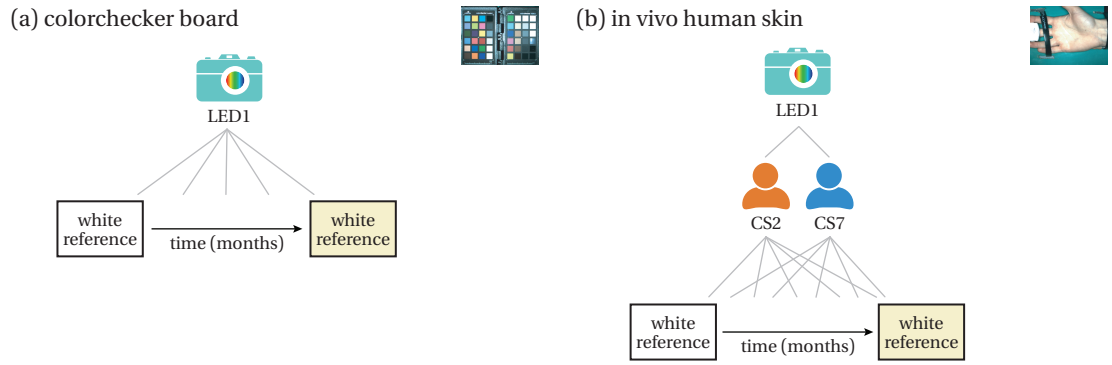


Figure 4.5: Experiments to study long-term temporal stability and calibration strategies. A multi-month measurement campaign was conducted using the LED1 device on (a) the colorchecker board and (b) human palm skin of two probands (CS2, CS7). Measurements were taken daily, excluding some weekends and public holidays. For each measurement day, calibration files were also recorded, enabling a comparison between two calibration approaches: a single calibration at the beginning of the campaign (*calibrating once*) versus calibration performed on the same day as the measurement (*daily calibration*).

of two probands (CS2 and CS7), as illustrated in Figure 4.5. Measurements were performed daily, with occasional breaks on weekends and public holidays. For the colorchecker board, the observation period spanned 418 days with 313 measurement days. For the human skin measurements, the observation period covered 204 days, with the two probands measured on separate days, totaling 183 measurement days. Due to limited device availability, only the device LED1 was used in this experiment. Before each measurement session, new white reference images were recorded, enabling a direct comparison of two calibration strategies: (1) the manufacturer-recommended approach of using calibration files recorded during production, referred to as *calibrating once*, and (2) our *daily calibration* approach, using calibration files recorded on the same day as the measurements.

4.3 Experiments and Results

This section presents the validation approach and findings for our 3 research questions on hardware-related sources of variation in TIVITA® devices. We first investigate the impact of device shifts on measured spectra and functional tissue parameter indices (Section 4.3.1), followed by an analysis of short-term temporal stability (Section 4.3.2) and long-term temporal stability (Section 4.3.3).

4.3.1 Device Shifts

To investigate the impact of device shifts on measured spectra and functional tissue parameter indices (RQ1.4), we compared measurements from two generations of TIVITA® devices, as well as measurements from different instances of the same device generation, conducted on a colorchecker board phantom and on human palm skin.

Colorchecker Phantom We hierarchically aggregated the measured spectra by first computing the median spectrum within the annotated area of each color field and then averaging the results across the 3 repetitions per device and color field. Figure 4.6 shows the resulting aggregated spectra alongside reference spectrometer measurements for 12 selected color fields from the colorchecker board phantom, including white, 8 color fields with different skin tones (desert sand, pancho, gold sand, tumbleweed, antique brass, light skin, dark skin, and tobacco brown) and 3 color fields with different flesh tones (froly, moderate red, and red).

For a quantitative assessment of the accuracy of the TIVITA® measurements, we computed the Euclidean distances between the TIVITA® spectra and spectrometer spectra for each color field in an image. Euclidean distances across the 3 repetitions were averaged. The resulting distribution of Euclidean distances across color fields and devices are shown in Figure 4.7.

Overall, the TIVITA® measurements are in good agreement with the spectrometer measurements. The largest deviations occur for the relatively dark color field "tobacco brown" measured with the first-generation devices Halogen1 and Halogen2. These devices are generally less accurate than the second-generation devices (LED1, LED2), particularly in the NIR range and below 550 nm. This may be caused by limitations of the halogen illumination: as shown in Figure 4.1, halogen light delivers high intensity around 700 nm but lower intensity in the NIR and below 550 nm.

Human Skin To assess the impact of device shifts on human palm skin spectra, we compared measurements from the 4 devices Halogen1, Halogen2, LED1 and LED2. Spectra were aggregated hierarchically by first computing the median within each annotated area and then averaging these medians across the 3 repetitions per device and proband. As shown in Figure 4.8, the standard deviation across repetitions is small. Consistent with the colorchecker board phantom results, systematic deviations across devices occur mainly in the NIR range and below 550 nm. Inter-device differences show that second-generation devices (LED1, LED2) produce more similar spectra to each other than to the first-generation devices (Halogen1, Halogen2).

The impact of device shifts on palm skin functional tissue parameter index values is shown in Figure 4.9. Index values were aggregated hierarchically by first computing the

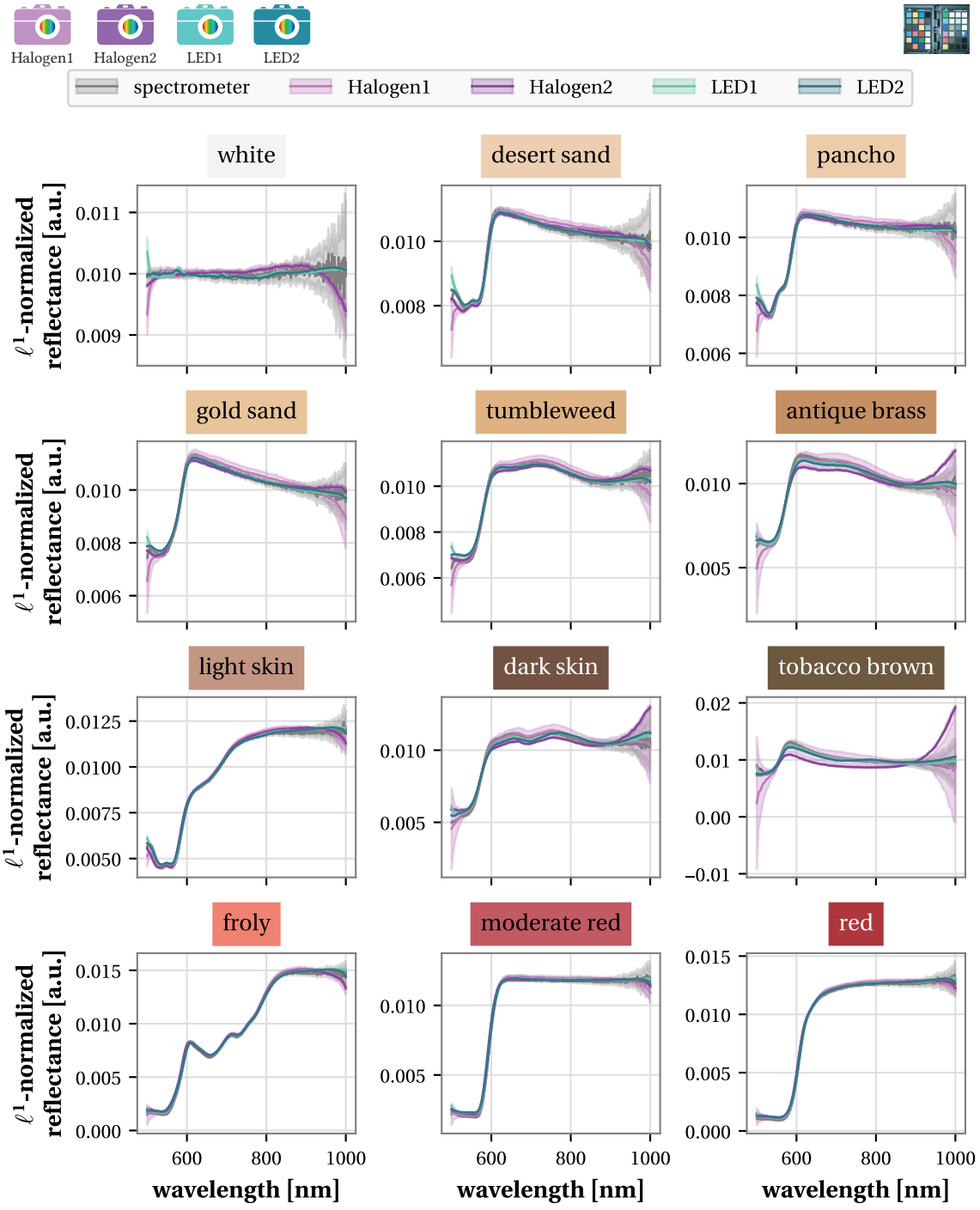


Figure 4.6: Comparison of colorchecker board spectra measured with different TIVITA® devices and a reference spectrometer. ℓ^1-normalized spectra are shown for 12 color fields of the colorchecker board, measured with the 4 devices Halogen1, Halogen2, LED1 and LED2 as well as a reference spectrometer. Spectra were aggregated hierarchically (first at the image level, then across repetitions) and shaded areas represent one standard deviation across the repetitions.

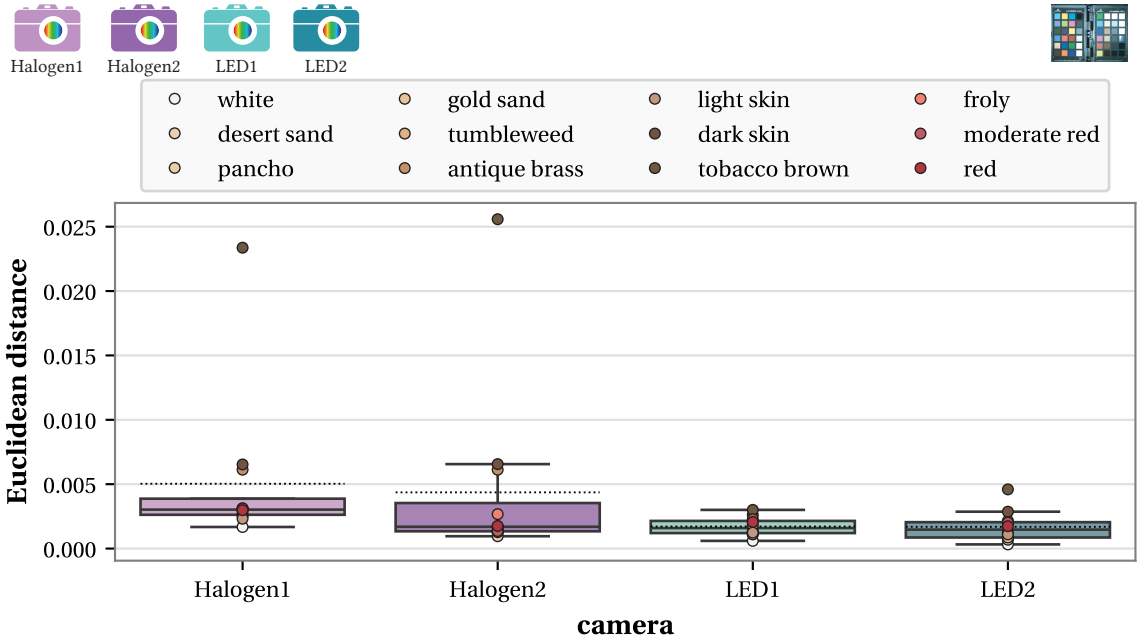


Figure 4.7: Distribution of Euclidean distance between TIVITA® measurements and reference spectrometer measurements across devices. The boxplots illustrate the Euclidean distance across the 4 devices Halogen1, Halogen2, LED1 and LED2, as well as the 12 different color fields of the colorchecker board phantom. Each box displays the interquartile range of the distribution, with whiskers showing the range excluding outliers, and median and mean indicated by a solid and dotted line, respectively. Each marker represents one color field.

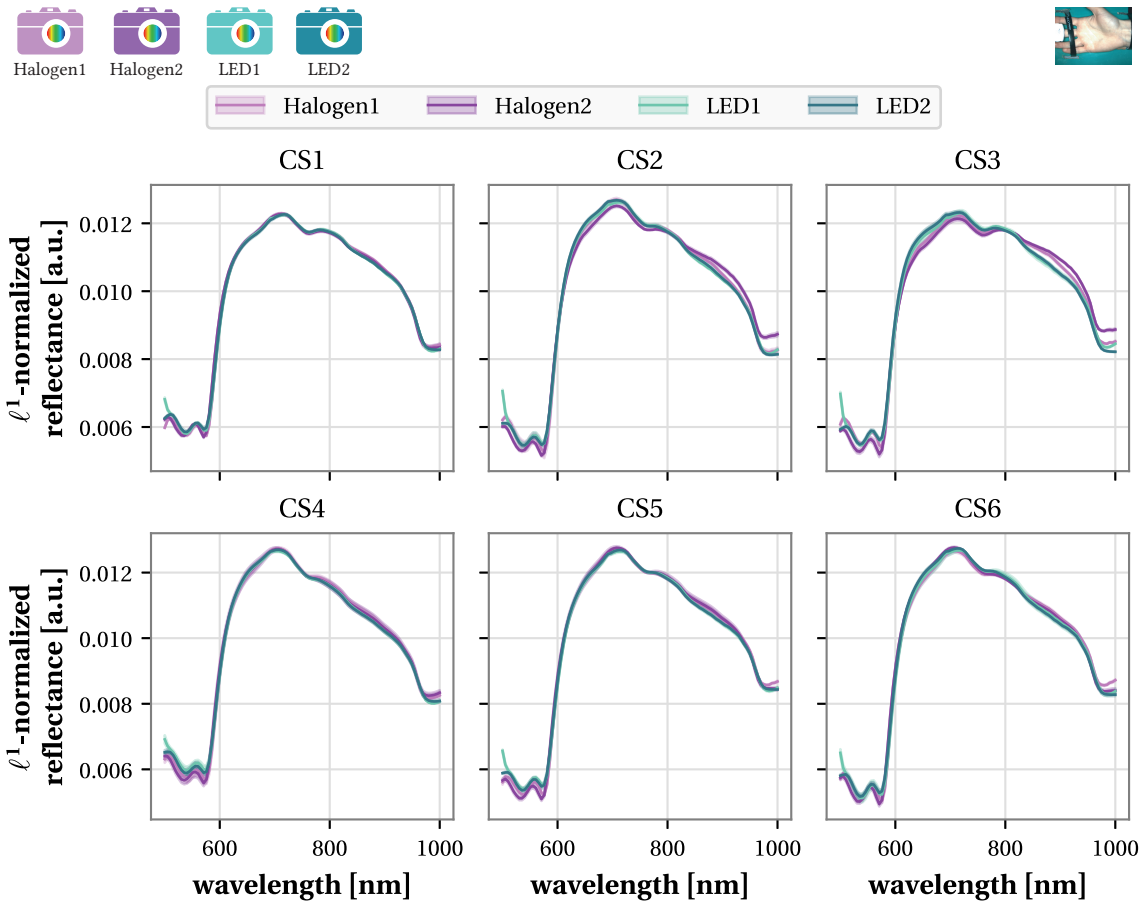


Figure 4.8: Comparison of human palm skin spectra across different TIVITA® devices. ℓ^1-normalized spectra are shown for 6 probands (CS1 to CS6), measured with the 4 devices Halogen1, Halogen2, LED1 and LED2. Spectra were aggregated hierarchically (first at the image level, then across repetitions) and shaded areas represent one standard deviation across the repetitions.

median within the annotated area of each image, and then averaging these medians across the 3 repetitions for each device and proband. Due to a high inter-subject variability in functional tissue parameter index values, we first computed a subject-specific baseline for each functional tissue parameter index p and subject s as the mean of index values $i(p, s, d)$ across all devices d in our set of 4 devices D:

$$\overline{i_s^p} = \frac{1}{|D|} \sum_{d \in D} i(p, s, d) \tag{4.1}$$

The device-specific deviation from this baseline was then calculated as:

$$\Delta_{\mathrm{m}} \,\mathrm{index}(p, s, d) = i(p, s, d) - \overline{i_s^p} \tag{4.2}$$

As shown in Figure 4.9, overall shifts in functional tissue parameter indices across devices are small. This is further illustrated in Figure 4.10, which presents sample images of proband CS1's palm skin measured with the 4 devices Halogen1, Halogen2, LED1, and LED2. The largest shift in functional tissue parameter indices occurred for tissue perfusion index in proband CS3, where values from Halogen2 and LED2 differed by 0.07.

4.3.2 Short-Term Temporal Stability

To investigate the short-term temporal stability of TIVITA® devices (RQ1.5), we analyzed how sensor temperature shifts affect spectral measurements and functional tissue parameter indices. Measurements were performed on a colorchecker board phantom and human palm skin, recording spectra every 30 s over a two-hour period. During this time, sensor temperature increased steadily, with an average rise of 3.9 °C (SD 0.8 °C) for the halogen devices and 23 °C (SD 4 °C) for the LED devices.

Colorchecker Phantom As illustrated in Figure 4.11 for the device Halogen2 and Figure 4.12 for the device LED2, the spectra of the colorchecker board phantom show a progresive shift with increasing sensor temperature. This also holds for the devices Halogen1 and LED1, as shown in the appendix figures Figure B.1 and Figure B.2, respectively.

To quantify the impact of sensor temperature shifts on spectral measurement accuracy, we calculated the Euclidean distance between spectra recorded at different sensor temperatures and the corresponding reference spectrometer spectra for each color field and device. The progression of Euclidean distances across color fields and sensor temperatures is shown in Figure 4.13 for the device Halogen2 and in Figure 4.14 for

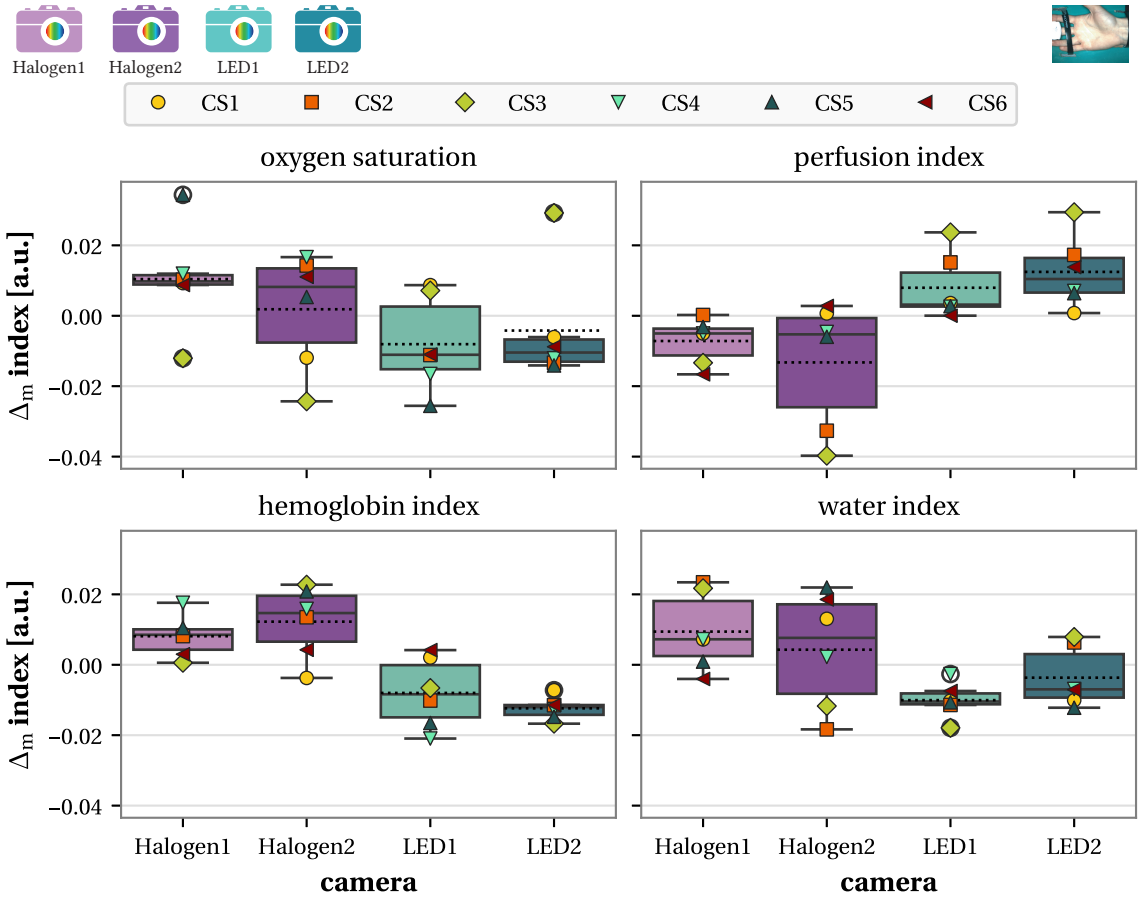


Figure 4.9: Impact of device shifts on functional tissue parameter indices of human palm skin. The distribution of Δ_m index, the relative shift in a functional tissue parameter index for a given proband and device compared to the proband-specific average value across devices (cf. Equation 4.2), is shown for the 4 parameter indices tissue oxygen saturation, tissue perfusion index, tissue hemoglobin index and tissue water index, as well as across probands CS1 to CS6 and devices Halogen1, Halogen2, LED1 and LED2. Each box displays the interquartile range of the distribution, with whiskers showing the range excluding outliers, and median and mean indicated by a solid and dotted line, respectively. Each marker represents one proband.

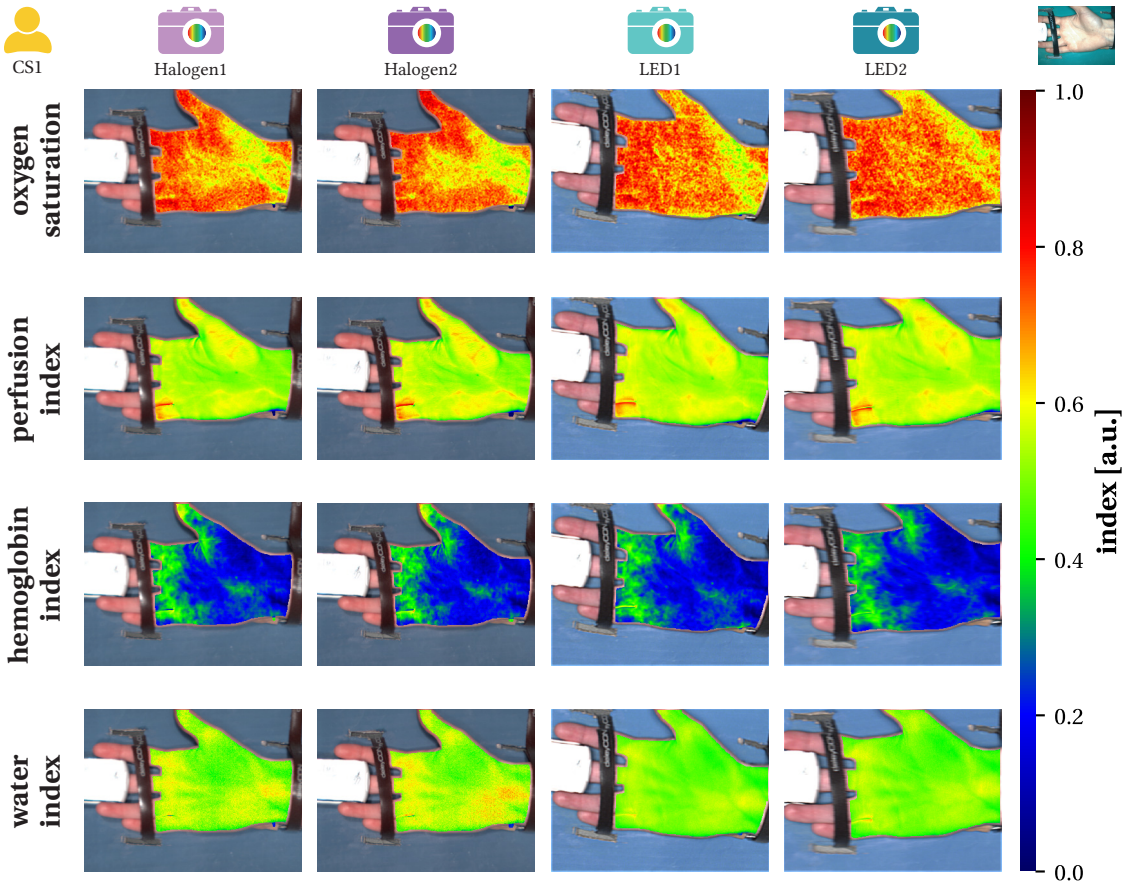


Figure 4.10: Exemplary functional tissue parameter images of human palm skin across devices. For proband CS1, reconstructed RGB images are shown overlaid with color-coded maps of the 4 functional tissue parameter indices tissue oxygen saturation, tissue perfusion index, tissue hemoglobin index, and tissue water index. Images are from a single measurement repetition for each of the devices Halogen1, Halogen2, LED1 and LED2.

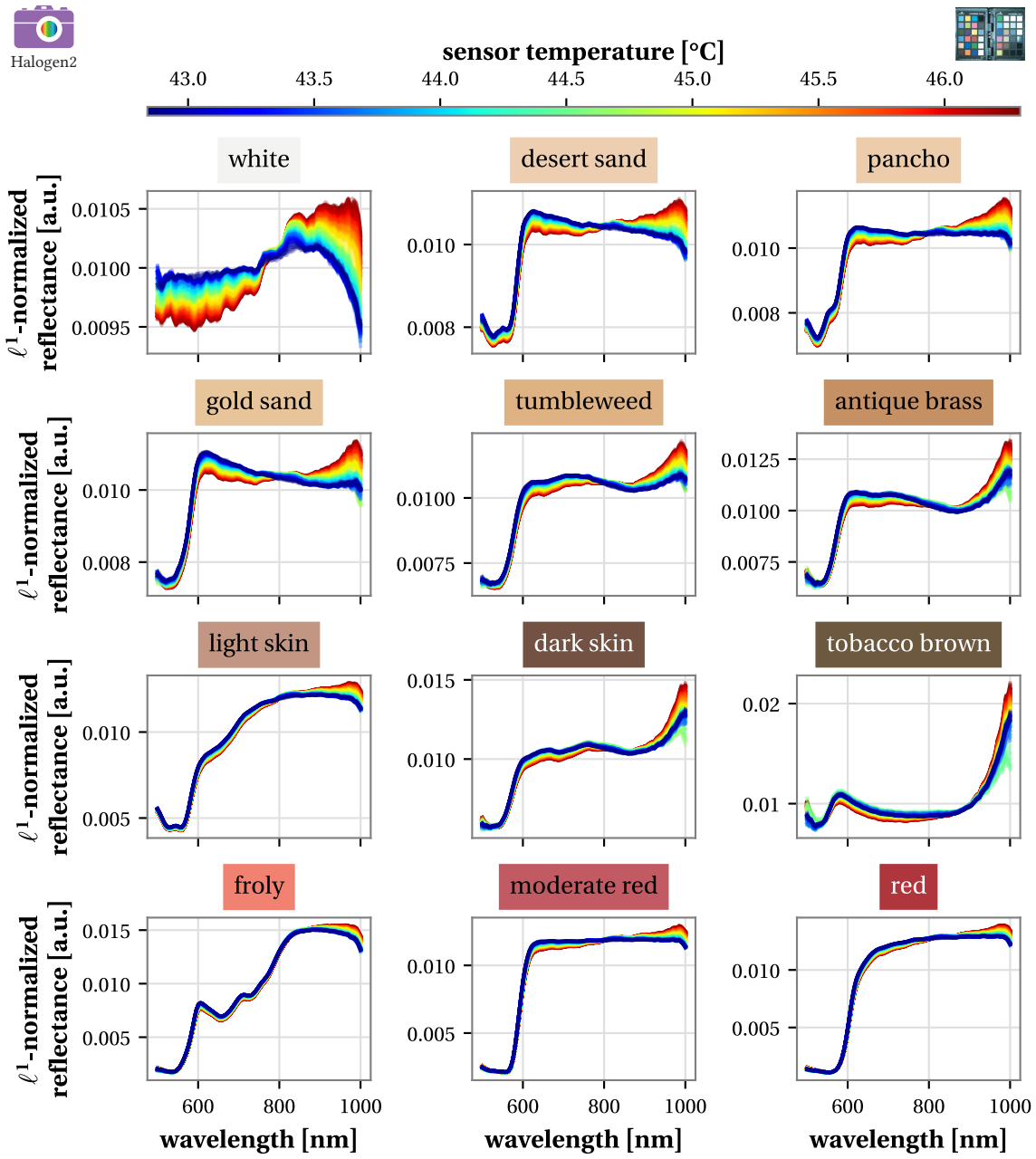


Figure 4.11: Shift in colorchecker board spectra measured with the device Halogen2 as a function of sensor temperature. ℓ^1-normalized spectra are shown for 12 color fields of the colorchecker board, with curves color-coded according to the sensor temperature at the time of measurement. For clarity, only data from one of the 3 repetitions is displayed.

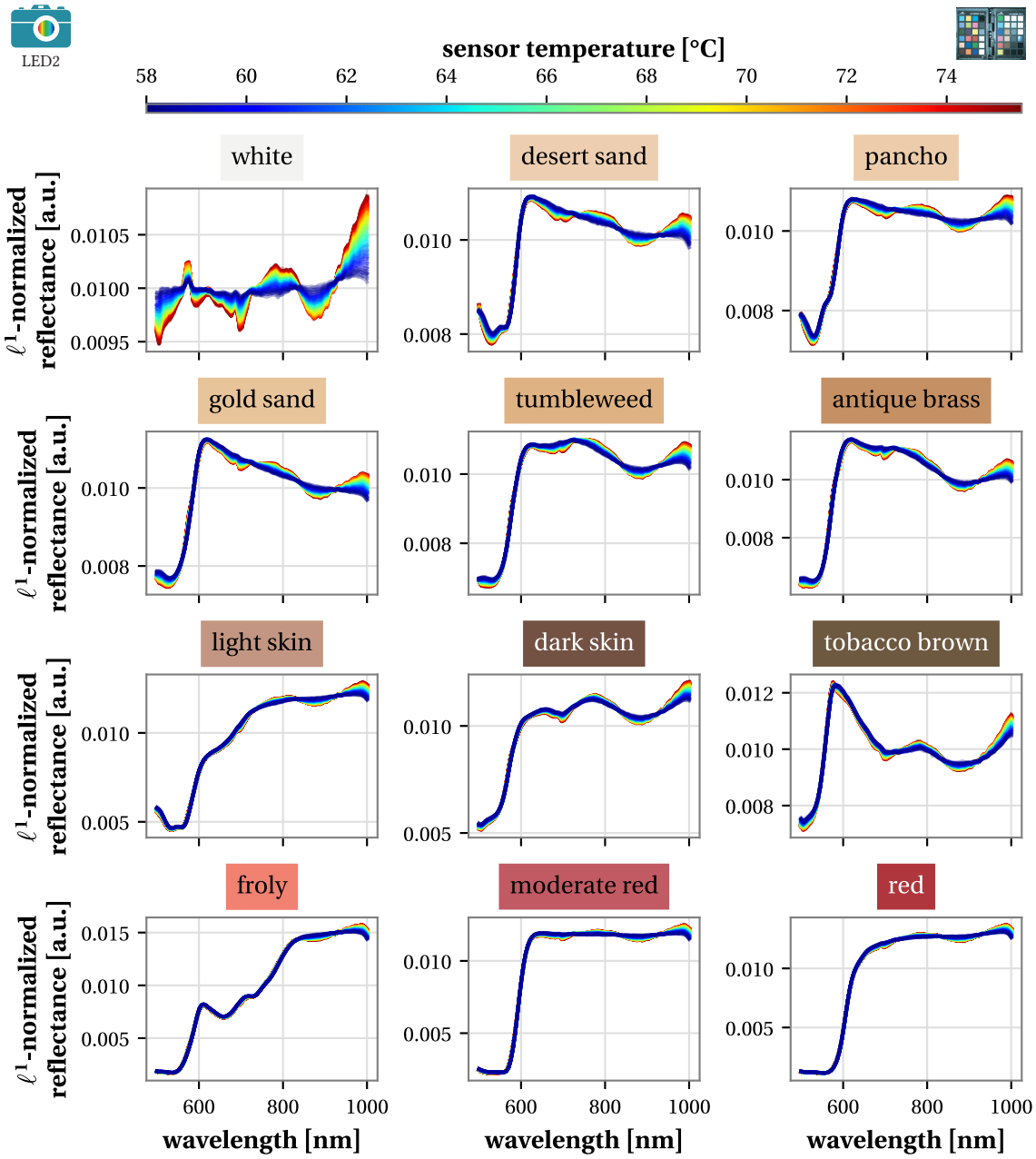


Figure 4.12: Shift in colorchecker board spectra measured with the device LED2 as a function of sensor temperature. ℓ^1-normalized spectra are shown for 12 color fields of the colorchecker board, with curves color-coded according to the sensor temperature at the time of measurement. For clarity, only data from one of the 3 repetitions is displayed.

the device LED2. Results for Halogen1 and LED1 are provided in the appendix (Figure B.3 and Figure B.4). Across all devices, the Euclidean distance between TIVITA® measurements and reference spectrometer measurements increases with rising sensor temperature, indicating a gradual loss of accuracy over the measurement series.

Human Skin To assess the impact of sensor temperature shifts on human skin spectra, we analyzed measurements from the devices Halogen2 and LED2 for 3 probands each. Consistent with the colorchecker phantom results, human palm skin spectra show a progressive shift with rising sensor temperature (cf. Figure 4.15).

Figure 4.16 shows the progression of the functional tissue parameter indices StO_2, NPI, THI, and TWI across sensor temperatures for the devices Halogen2 and LED2. Notable observations include a strong decline in StO_2 for the device LED2 with increasing sensor temperature, with a reduction in StO_2 of up to -0.27 between first and last measurement. The TWI shows a small decline for both devices, with a reduction between first and last measurement of up to -0.06. The indices NPI and THI show no consistent trends across probands.

Compared to the small inter-device shifts in functional tissue parameter indices (cf. Figure 4.9), the changes linked to rising sensor temperature are substantially larger – particularly for StO_2 estimates from the LED2 device, where shifts are roughly an order of magnitude greater. As shown in Figure 4.17, which presents sample index maps of human palm skin from Halogen2 and LED2 at both low and high sensor temperatures, these pronounced shifts are clearly visible in the functional tissue parameter images.

4.3.3 Long-Term Temporal Stability

To investigate the long-term temporal stability of TIVITA® devices and analyze the impact of calibration shifts (RQ1.6), we performed repeated measurements on a colorchecker board phantom and human palm skin over several months. Each image was processed twice, once using the calibration files recorded on the same day as the measurement (*daily calibration*) and once using the calibration files recorded at the beginning of the multi-month measurement period (*calibrating once*), thus enabling a direct comparison of the two calibration strategies.

Colorchecker Phantom To quantify the impact of calibration strategy on the accuracy of TIVITA® measurements, we compared the Euclidean distances between calibrated TIVITA® spectra and reference spectrometer measurements on the colorchecker board phantom across the different measurement days. As shown in Figure 4.18, consistently across all color fields, the *daily calibration* strategy results in lower Euclidean distances compared to *calibrating once* at the beginning of the measurement campaign.

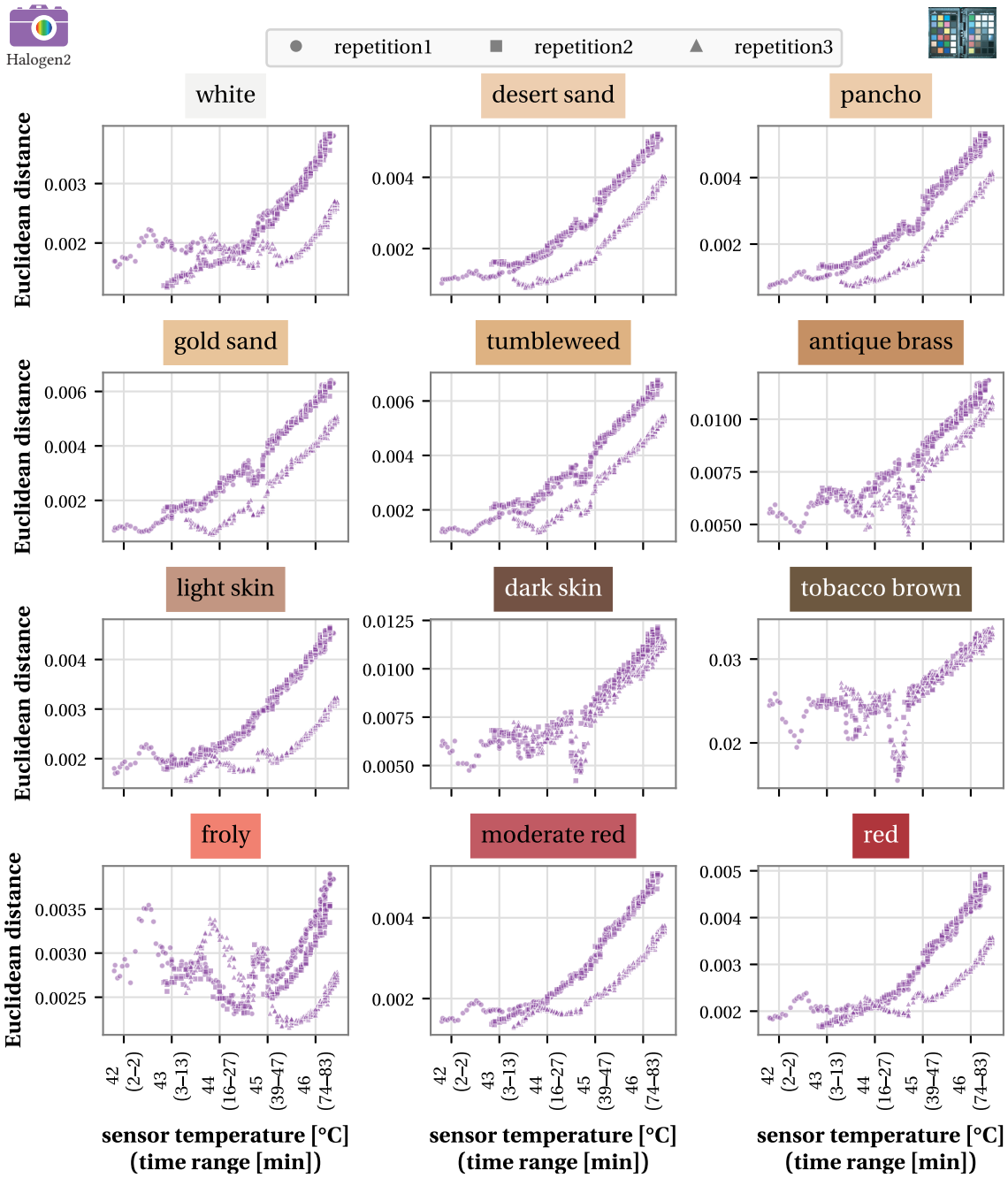


Figure 4.13: Shift in Euclidean distance between spectra measured with device Halogen2 and a reference spectrometer as a function of sensor temperature. Euclidean distance is shown as a function of sensor temperature for 12 color fields of the colorchecker board. Measurements from the 3 repetitions are distinguished by different markers.

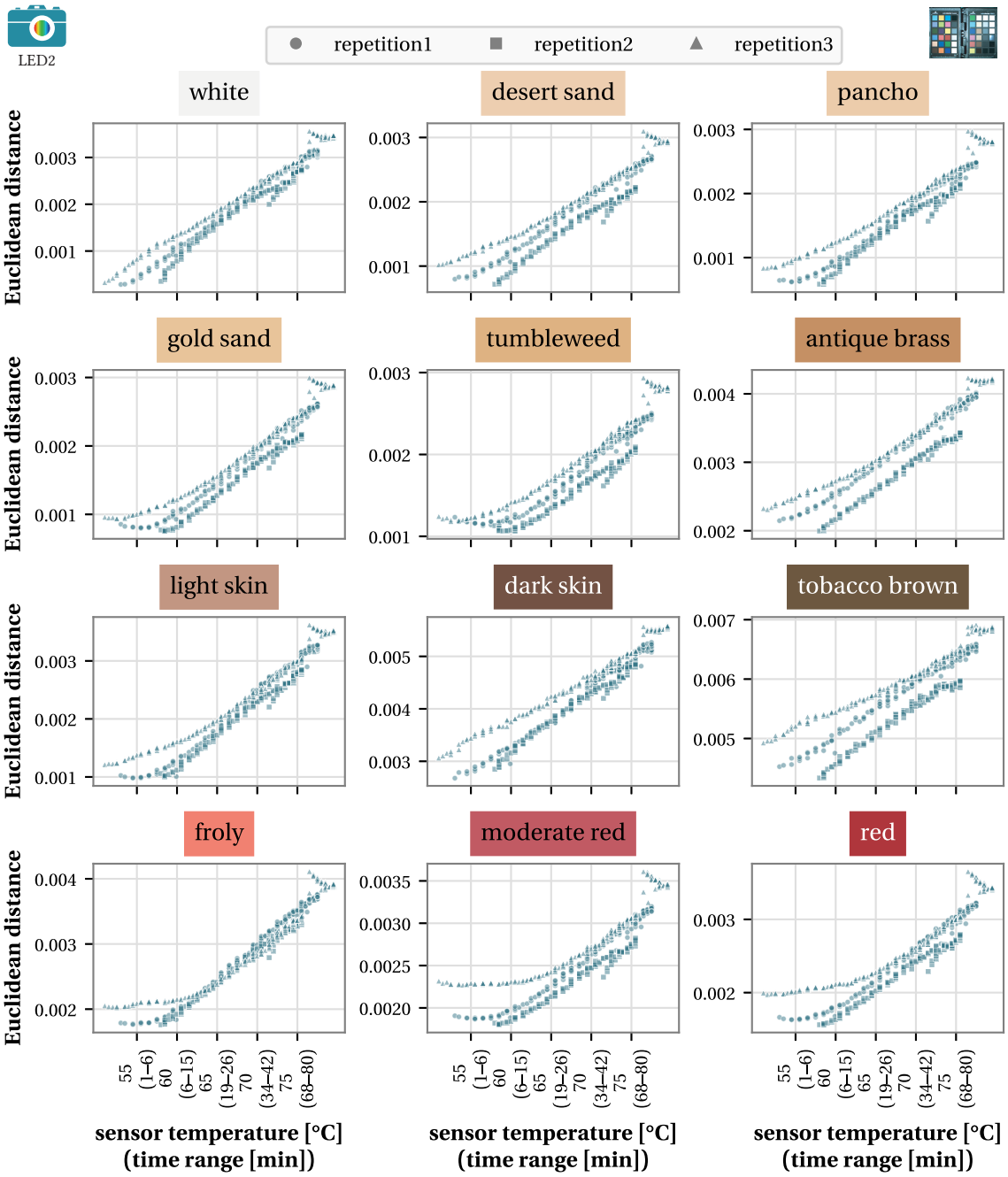


Figure 4.14: Shift in Euclidean distance between spectra measured with device LED2 and a reference spectrometer as a function of sensor temperature. Euclidean distance is shown as a function of sensor temperature for 12 color fields of the colorchecker board. Measurements from the 3 repetitions are distinguished by different markers.

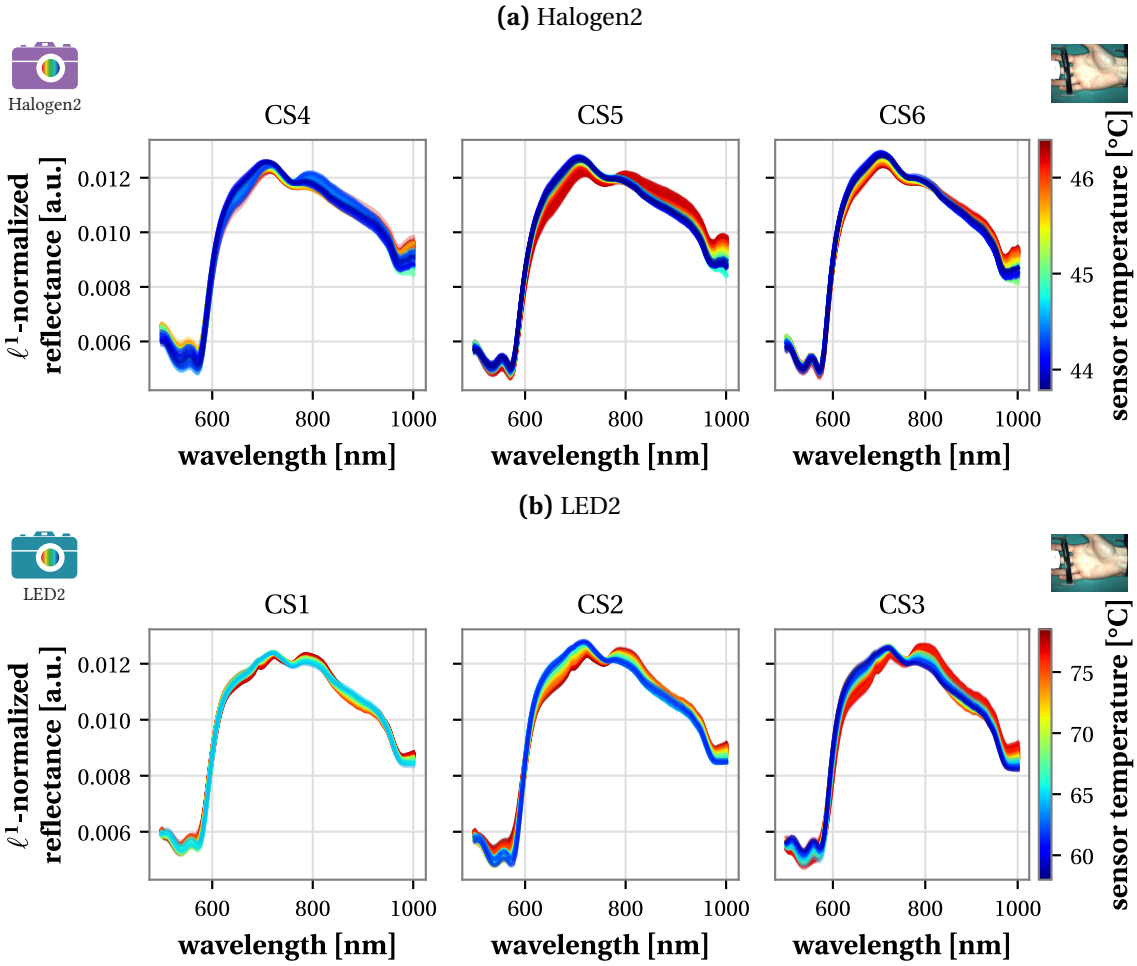


Figure 4.15: Shift in human palm skin spectra as a function of sensor temperature. ℓ^1-normalized human palm skin spectra are shown for (a) probands CS4 to CS6 measured with Halogen2, and (b) probands CS1 to CS3 measured with LED2. Curves are color-coded according to the sensor temperature at the time of measurement.

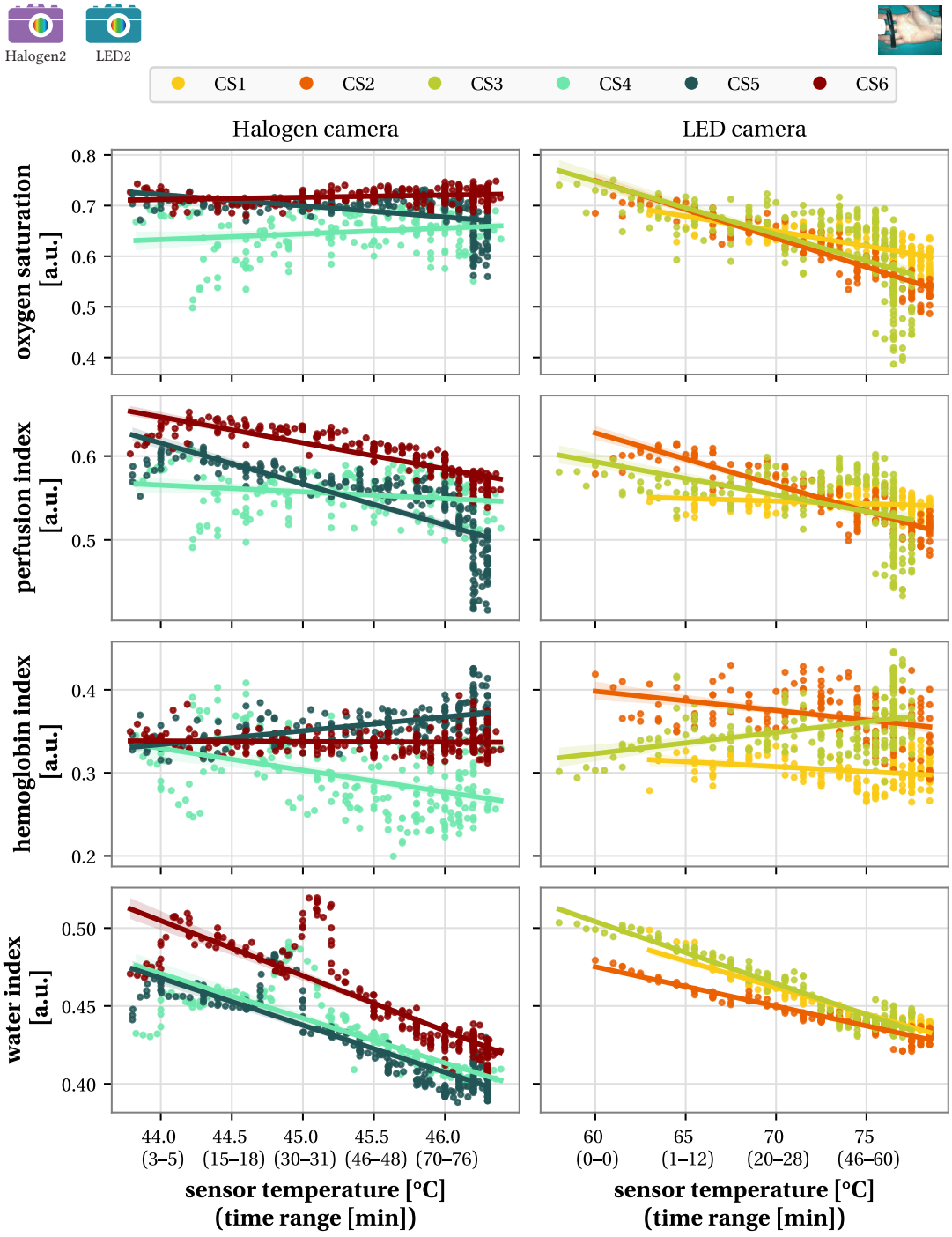


Figure 4.16: Impact of rising sensor temperature on functional tissue parameter indices of human palm skin. Scatter plots show values of the tissue parameter indices oxygen saturation, perfusion index, hemoglobin index and water index as a function of sensor temperature for the devices Halogen2 (probands CS4 to CS6) and LED2 (probands CS1 to CS3). Linear regression fits are shown as lines, with the 95 % confidence interval derived from 1000 bootstrap samples shown as shaded areas.

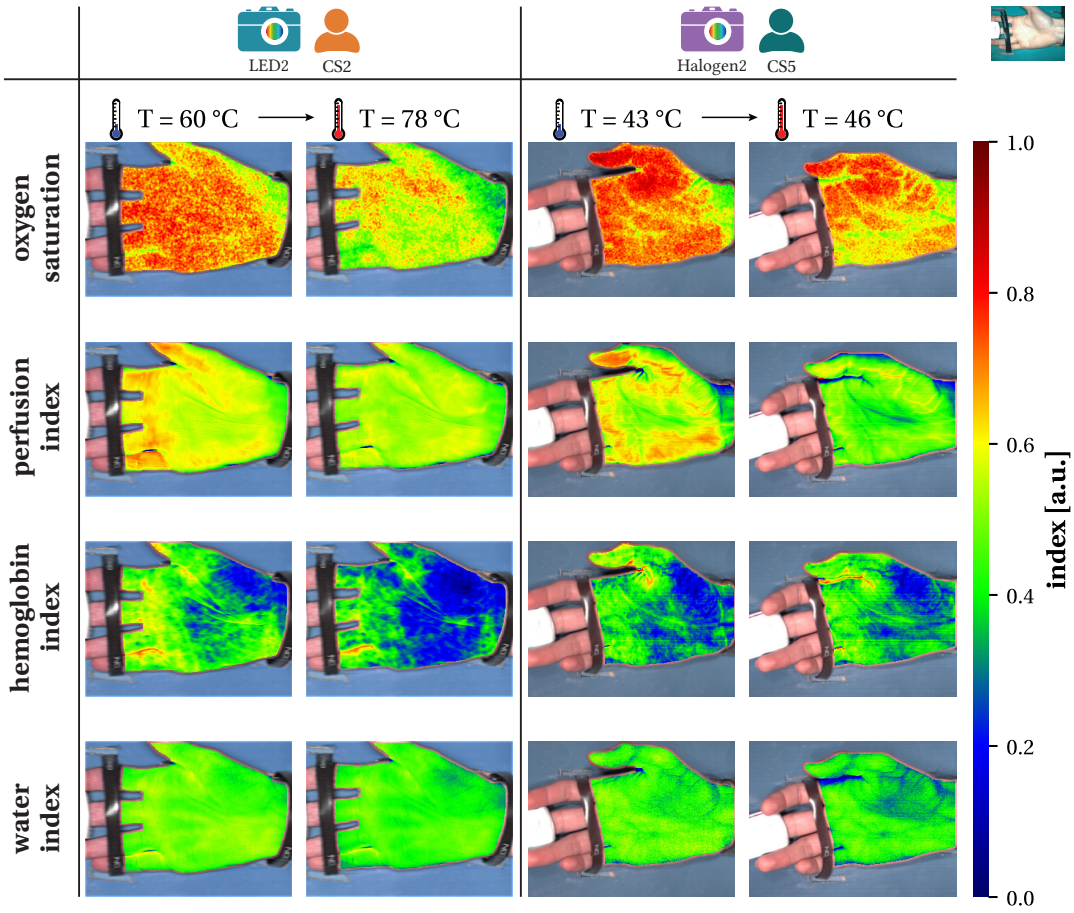


Figure 4.17: Exemplary functional tissue parameter images of human palm skin across sensor temperature. Reconstructed RGB images are overlaid with color-coded maps of the functional tissue parameter indices tissue oxygen saturation, tissue perfusion index, tissue hemoglobin index, and tissue water index. The first measurement (low sensor temperature) is compared with the last measurement (high sensor temperature) for two series: proband CS2 measured with LED2 (left) and proband CS5 measured with Halogen2 (right).

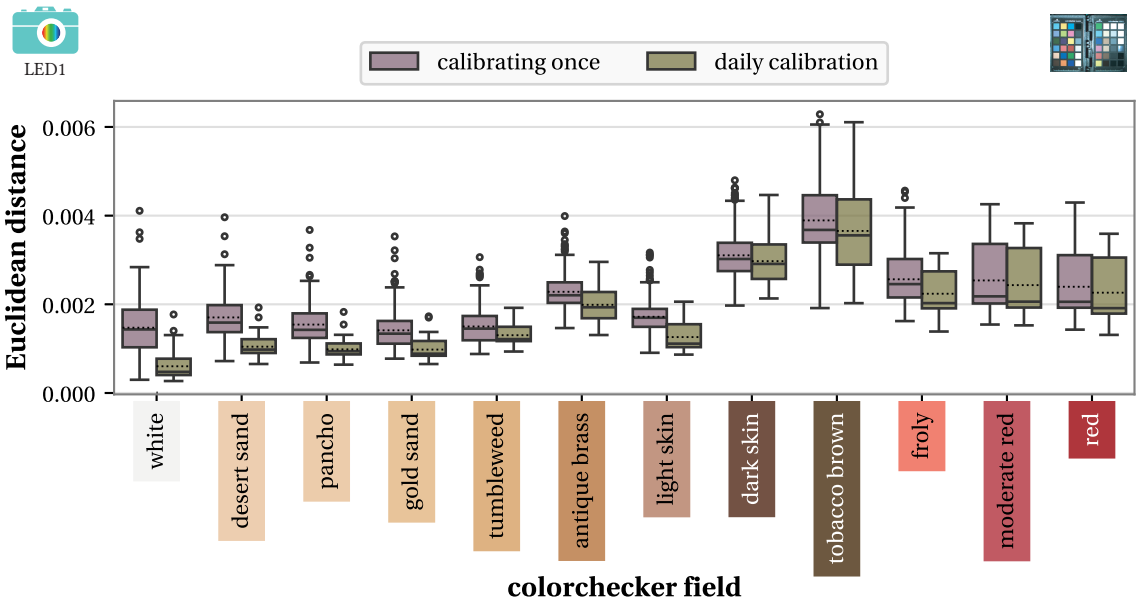


Figure 4.18: Accuracy of TIVITA® measurements as a function of calibration scheme. The boxplots show the distribution of Euclidean distances between measurements taken with the device LED1 and a reference spectrometer across 313 different measurement days. Two calibration strategies are compared: *daily calibration* and *calibrating once* at the beginning of the measurement period. Measurements were taken on 12 distinct color fields of a colorchecker board phantom. Each box displays the interquartile range of the distribution, with whiskers showing the range excluding outliers. The median and mean are indicated by solid and dotted lines, respectively.

Human Skin To evaluate the effect of different calibration schemes on human skin functional tissue parameter indices, we calculated the absolute difference between index values from pairs of processed TIVITA® images differing only in calibration method, denoted as Δ_c, index. The resulting distributions of Δ_c index are shown in Figure 4.19 for the two probands CS2 and CS7. Overall, shifts in functional tissue parameter indices due to calibration scheme are small. The largest calibration-related shifts are observed for StO_2 and TWI, with up to 0.06 and 0.08, respectively. The shifts in NPI and THI are smaller, with maximum shifts of 0.02 and 0.01, respectively.

4.4 Discussion and Conclusion

In this study, we demonstrated for the first time that TIVITA® HSI devices are sensitive to hardware-related sources of variation. Through an extensive validation on a colorchecker board phantom and the skin of 7 healthy volunteers, we investigated 3 hardware shift scenarios: (1) inter-device variability, (2) sensor temperature increases during a measurement series, and (3) calibration shifts over several months.

Our results show that TIVITA® devices exhibit systematic spectral measurement shifts across device generations and instances, sensor temperatures, and calibration schemes. Overall, second-generation devices and the use of daily calibration – rather than the manufacturer's recommended single initial calibration – yield spectral measurements that more closely match reference spectrometer data. Shifts in functional tissue parameter indices caused by changes in device or calibration scheme were generally small. In contrast, rising sensor temperature produced substantial index shifts, particularly in StO_2 estimates from the LED2 device.

4.4.1 Implications for the Design of Bias-Aware Hyperspectral Imaging Studies

Our findings highlight the need to account for hardware-related sources of variation in HSI studies, particularly when using TIVITA® devices. We recommend the following countermeasures for study design and analysis:

- **Counteracting device shifts:** To avoid device shifts, the same device should be used for all measurements in a study.
- **Counteracting sensor temperature shifts:** To mitigate sensor temperature shifts, sensor temperatures should be monitored during measurements. In our experiments, a 30 s interval was chosen to capture fast physiological dynamics. If high temporal resolution is not required, measurement intervals should be increased

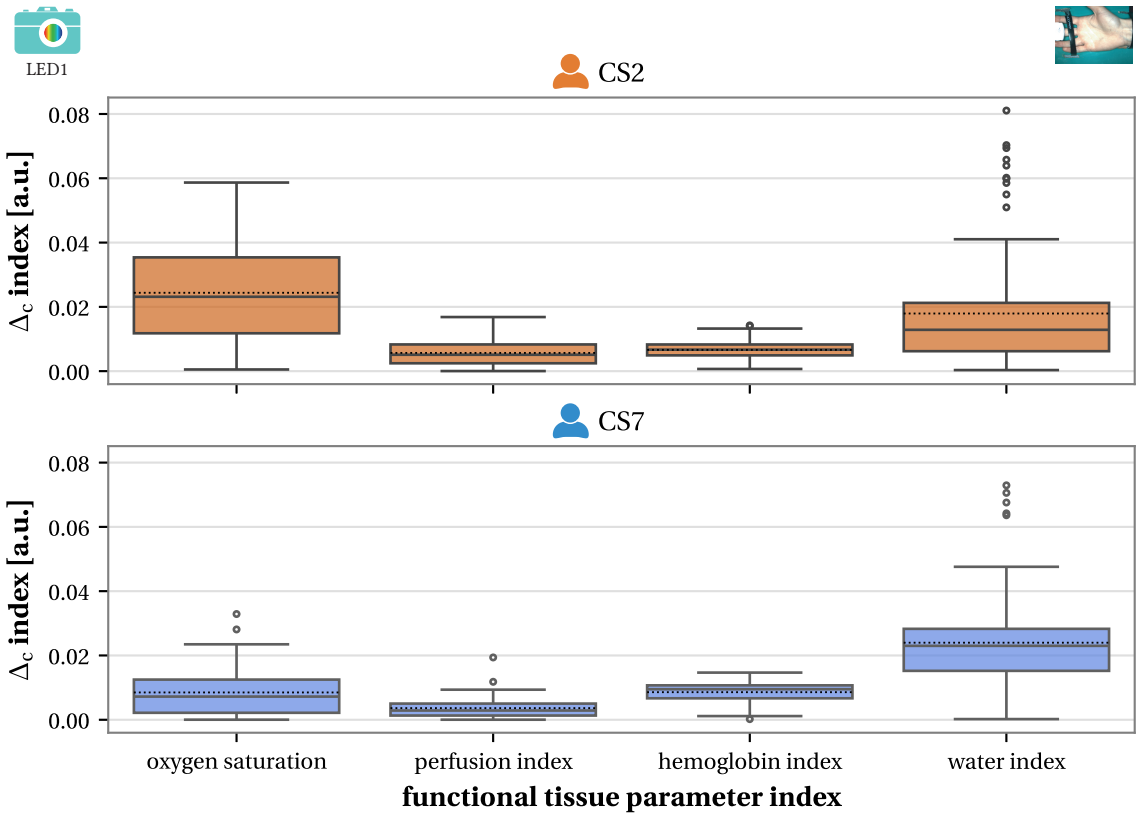


Figure 4.19: Impact of calibration scheme on human skin functional tissue parameter indices. Two calibration strategies are compared for the device LED1: (1) using calibration files recorded on the same day as the measurement, and (2) using calibration files from the start of the multi-month measurement period. The boxplots show the distribution of absolute differences in oxygen saturation, perfusion index, hemoglobin index and water index – denoted as Δ_c index – calculated by processing each image with the two calibration strategies. Each box displays the interquartile range of the distribution, with whiskers showing the range excluding outliers. The median and mean are indicated by solid and dotted lines, respectively. Measurements were performed on probands CS2 (top) and CS7 (bottom).

to reduce sensor temperature rise. For comparisons between cohorts, the order of measurements should be carefully planned – ideally randomized – to avoid systematic temperature differences between groups.

- **Counteracting calibration shifts:** To avoid calibration shifts and improve the accuracy of HSI measurements, we recommend performing frequent device calibration. Although calibration during a measurement session may not be practical in some clinical contexts (e.g., due to sterility constraints, cf. Chapter 3), calibration prior to each measurement session should be performed whenever possible. This is particularly important in long-term studies, where hardware aging could introduce gradual measurement drift.
- **Device characterization:** We recommend performing a rigorous characterization of each device in use, for example through regular measurements on a colorchecker board phantom. This allows for monitoring of device accuracy, detection of hardware-related shifts over time, and potential correction based on these observations.
- **Interpretation of HSI data:** Small differences in HSI data should be interpreted cautiously if potential confounding from hardware-related sources of variation cannot be excluded. This is particularly relevant given that several studies using TIVITA® devices have reported and interpreted changes in functional tissue parameter indices of similar magnitude to those observed from hardware-related variation, for example when evaluating intervention effects [124, 327] or comparing sepsis survivors with non-survivors [80].

4.4.2 Limitations and Future Work

In the following, we discuss the key strengths and limitations of our study and outline potential directions for future research.

Strengths and Limitations of Our in Vivo Measurements Our in vivo measurements were conducted on the palm skin of healthy volunteers, providing direct insight into hardware-related variability at a measurement site intended for our future study on automated sepsis diagnosis and mortality prediction. The calibration scheme assessment was unaffected by potential changes in the proband's physiological state, as both schemes were applied to the same set of images. In contrast, analyses of device shifts and sensor temperature shifts required comparisons between different images, relying on the assumption that each proband's physiological state remained stable throughout the measurement series.

As described in Section 4.2.2, several precautions were taken to ensure physiological stability, including prohibiting physical activity during measurements and having

probands sit for at least 5 min prior to measurements. In addition, a pulse oximeter was used to continuously monitor SpO_2, heart rate, and respiratory frequency. Figure B.5, Figure B.6, and Figure B.7 display these parameters for the 6 probands during the sensor temperature experiments, plotted against the corresponding HSI sensor temperature. The absence of parameter trends aligning with shifts in HSI spectra or functional tissue parameter estimates supports the assumption that physiological changes did not bias our results.

As an additional validation, we compared the impact of sensor temperature shifts on functional tissue parameters of human skin to that observed for the "light skin" color field of the colorchecker board phantom, which most closely resembles human skin spectra. As shown in Figure B.8, the "light skin" field exhibited similar trends to in vivo skin measurements with increasing sensor temperature, including a strong decrease in StO_2 for LED2 and a slight decrease in TWI across all devices. These findings support the robustness of our conclusions.

Generalizability of Our Findings In our colorchecker board phantom measurements, we observed that the accuracy of TIVITA® measurements depends on the measurement object. For example, Euclidean distances for the "tobacco brown" field were consistently larger than for "light skin" (cf. Figure 4.7, Figure 4.18). This suggests that the impact of hardware-related sources of variation on HSI measurements may differ across measurement objects, such as different tissue types. Furthermore, this study focused on functional tissue parameter estimation as the downstream task, since it is the most common application of TIVITA® devices. However, hardware-related shifts in HSI measurements may influence other downstream tasks in different ways. Even within our analysis, the magnitude of these shifts varied across functional tissue parameter indices – for example, StO_2 and TWI estimations were more strongly affected by sensor temperature changes than NPI and THI. In particular, ML algorithms are prone to exploiting unwanted shortcuts in the data rather than learning task-relevant features [28, 113, 385, 288]. Future work should therefore investigate the influence of hardware-related variation on a broader range of measurement objects and downstream tasks, including HSI-based classification and segmentation algorithms.

Hardware-Related Variation Beyond Spectral Shifts This study focused on spectral shifts as the primary source of hardware-related variation. However, other factors – such as geometric shifts caused by differences in field of view across devices – may also be relevant, particularly for image-based rather than pixel-based algorithms. Future work should examine these additional sources of variation and assess their effects on HSI data and different downstream tasks.

Addressing Hardware-Related Variation We have provided recommendations for mitigating hardware-related bias in study design and analysis. In parallel, there is a need for manufacturers to improve the reliability of measurements. For instance, the impact of sensor temperature shifts could be reduced through improved heat dissipation. Additionally, manufacturers should offer clear guidance on the expected accuracy of measurements under different conditions and define error margins that set meaningful boundaries for interpreting HSI data. Furthermore, algorithms could be developed to automatically detect hardware-related shifts, enabling real-time corrective actions such as initiating recalibrations or scheduling cooldown periods. Future research should focus on creating algorithms that are inherently robust to residual hardware-related variation, including more reliable functional tissue parameter estimation algorithms and ML models trained to generalize across hardware shifts.

4.4.3 Conclusion

To our knowledge, this work presents the first systematic analysis of hardware-related sources of variation in HSI measurements obtained with TIVITA® devices. We showed that these devices are sensitive to such variations, particularly increases in sensor temperature, and provided recommendations for study design and analysis to mitigate associated bias. Adhering to these guidelines will facilitate unbiased evaluation of HSI for applications such as automated surgical scene segmentation (Part III) as well as sepsis diagnosis and mortality prediction (Part IV).

Part III

Robust Surgical Scene Segmentation with Hyperspectral Imaging (RQ2)

5 Impact of Spatial Granularity And Modality on Surgical Scene Segmentation

As outlined in Section 1.2.2, fully automated surgical scene segmentation is a pivotal step towards the development of intelligent surgical systems capable of providing real-time, context-aware intraoperative decision support, as well as advancing autonomous surgical robotics. However, to date, semantic surgical scene segmentation using SI data has received little attention. Consequently, the potential advantages of SI data over other modalities, as well as how best to optimally represent input data for DL-based segmentation algorithms, remain largely unexplored. In this chapter, we address this important knowledge gap by investigating the impact of the input spatial granularity (e.g., pixels, superpixels, patches, or images) and imaging modality (e.g., RGB, HSI, or processed HSI data) on the performance of DL-based surgical scene segmentation algorithms. Our work contributes the largest semantically annotated intraoperative SI dataset to date, alongside a comprehensive framework for surgical scene segmentation across spatial granularities and modalities, which is publicly available in our GitHub repository (https://github.com/IMSY-DKFZ/htc)[312].

Section 5.1 provides an overview of the related work on surgical scene segmentation, followed by a description of our datasets and DL approach to automated semantic scene segmentation in Section 5.2. The experimental setup and results are presented in Section 5.3, and the chapter concludes with a discussion of the design choices, strengths, limitations, and directions for future research in Section 5.4.

The research presented in this chapter was published in the Medical Image Analysis journal in 2022 [308], as well as in the thesis of Jan Sellner in 2024 [311]. It was further presented at the IPCAI in 2022 [307].

5.1 Related Work

Only few works have addressed the task of surgical scene segmentation using MSI or HSI data. To this end, we provide a broad overview of related work in surgical scene segmentation using RGB data (Section 5.1.1), followed by a more detailed discussion of the existing literature on multispectral and HSI-based intraoperative tissue segmentation (Section 5.1.2).

5.1.1 Surgical Scene Segmentation Using RGB Data

Segmentation of RGB data of surgical scenes has been explored in various studies, with most efforts focusing on medical instrument segmentation [297]. This trend is driven by challenges in the field, such as the CATARACTS challenge on automatic tool segmentation in microscopic cataract surgery [10], and the Endoscopic Vision Grand Challenges in laparoscopic colorectal surgery [43, 13, 224]. The release of additional public datasets for instrument segmentation, such as those for gastrointestinal endoscopy [161] and robot-assisted prostatectomy [27], has further spurred research in this area.

In contrast, relatively few studies have focused on the segmentation of anatomical structures in surgical scenes. A snapshot of related work is presented in Table 5.1. These studies either target specific organ classes – such as the uterus [69], liver [119, 106], or recurrent laryngeal nerve [122] – or, more commonly, address full scene segmentation in various surgical contexts, such as robotic nephrectomy [12], laparoscopic hysterectomy [220], laparoscopic cholecystectomy [231], and robotic rectal resection [185]. The datasets used in these studies differ widely in terms of the spatial and temporal resolution of the video frames, as well as the number of classes considered (cf. Table 5.1). Surgical scene segmentation approaches for RGB data predominantly rely on CNNs that process entire images.

The vast majority of studies on automated surgical scene segmentation focus on microscopic or minimally invasive procedures, likely because RGB imaging systems are routinely utilized in these procedures. We are aware of only one study that specifically addresses segmentation in open surgical scenes using RGB data ([122]). Compared to minimally invasive surgeries, automated semantic segmentation in open surgeries poses additional challenges due to the greater variability and complexity of the surgical scene. For instance, Gong et al. demonstrated that shifts in imaging conditions, such as changes in lighting or variations in camera distance, substantially affect segmentation performance [122].

Previous research has highlighted several key challenges for automated surgical scene segmentation using RGB data, including substantial variability in tissue appearance

both across patients [69, 119] and within images due to factors such as occlusions or deformations [239]. Incorporating additional spectral information could be key for overcoming these challenges, since SI may depend less on spatial context while offering enhanced clinical insights (e.g., functional tissue parameters) [95].

5.1.2 Surgical Scene Segmentation Using Spectral Imaging Data

We only identified 15 studies addressing the task of intraoperative organ segmentation using MSI or HSI data. A tabular summary of these studies is presented in Table 5.2. Of these, 10 studies employ DL methods. Most studies focus on tumor segmentation in brain surgery or ex-vivo specimens, while only two studies tackle surgical scene segmentation in visceral surgeries, which is the focus of our work.

The key limitations of existing studies include:

- **Benefit of SI over RGB data**: Only few studies have explored whether SI provides advantages over RGB imaging for organ segmentation. Moccia et al. performed segmentation of 6 abdominal organs during laparoscopic hepatic surgery. To demonstrate the benefit of MSI over RGB, they compared MSI to a selection of 3 spectral bands. However, the narrow spectral bandwidth of these bands does not accurately represent realistic RGB data [239]. Garifullin et al. compared MSI and RGB data for segmenting vessels, optic disc, and macula in retinal imaging, finding only marginal performance improvements from MSI [110]. Similarly, Garcia Peraza Herrera et al. observed only mild improvements in image-wide performance metrics when using HSI compared to RGB data for segmenting 35 classes in oral and dental SI data. However, certain tissue classes (e.g., attached gingiva) were significantly better recognized with HSI [109]. Despite these efforts, the potential advantages of HSI over RGB data for fully semantic scene segmentation in visceral surgeries remain unexplored.
- **Spatial granularity**: The spatial granularity of segmentation algorithms varies substantially across studies. Some focus on pixel-level segmentation (e.g., [9, 279, 92]), while others employ superpixels (e.g., [239, 214]), patches (e.g., [340, 59, 70, 339]), entire images (e.g., [110, 109]), or a mixture of granularities (e.g., [339, 202, 214]). The choice of spatial granularity likely influences the performance of surgical scene segmentation algorithms and their ability to generalize to unseen surgical scenarios. However, a systematic investigation into this relationship has not yet been performed.
- **Amount of training data**: The number of subjects used for training and validation ranges widely, from as few as one subject [9] to 169 subjects in the most recent study [30]. Some researchers justify using smaller input spatial granularities, such as patches, with limited dataset sizes (e.g., [59, 70]). They argue that

Table 5.1: Snapshot of related work on surgical scene segmentation using RGB data. Related publications are organized chronologically by publication year and include details on the type of surgery, the number of subjects (N_s), the number of classes (N_c), and the name of the dataset in case a publicly available dataset was used.

publication	year	type of surgery	N_s	N_c	dataset
Collins et al. [69]	2015	laparoscopic (uterus)	126	1	private
Gibson et al. [119]	2017	laparoscopic (liver)	13	1	private
Fu et al. [106]	2019	laparoscopic (liver)	13	1	private
Kadkhodamohammadi et al. [165]	2019	laparoscopic (liver)	5	14	private
Laves et al. [196]	2019	endoscopic (larynx)	2	7	private
Allan et al. [12]	2020	robotic nephrectomy	19	12	EndoVisSub2018
Madad Zadeh et al. [220]	2020	laparoscopic hysterectomy	8	3	SurgAI
Maqbool et al. [231]	2020	laparoscopic cholecystectomy	?	19	m2caiSeg
Scheikl et al. [302]	2020	laparoscopic cholecystectomy	2	6	private
Gong et al. [122]	2021	open thyroidectomy	130	1	private
Grammatikopoulou et al. [125]	2021	microscopic (cataract)	25	8–25	CaDIS
Jin et al. [162]	2022	mixed	25/19	8–25/12	CaDIS/EndoVisSub2018
Bhattarai et al. [38]	2023	mixed	25/19	8–25/12	CaDIS/EndoVisSub2018
Ghamsarian et al. [115]	2023	microscopic (cataract)	30	12	Cataract-1K
Kolbinger et al. [185]	2023	robotic (rectal resection)	32	11	Dresden Surgical Anatomy Dataset
Luo et al. [216]	2023	microscopic (neurosurgery)	12	19	private
Liu et al. [209]	2024	mixed	25/19/8	8–25/12/8	CaDIS/EndoVisSub2018/private
Urrea et al. [344]	2024	laparoscopic cholecystectomy	17	13	CholecSeg8K

Table 5.2: Overview of related work on multispectral and hyperspectral imaging-based intraoperative tissue segmentation. Related publications are organized chronologically by publication year and include details on the target tissues, the number of subjects (N_s), the number of classes (N_c), the spatial granularity of the segmentation algorithms, and the level of detail of the provided annotations. In the context of the selected input spatial granularity, the term "mixed" denotes models that combine multiple levels of spatial granularity. Our own work on surgical scene segmentation ([308, 314, 309, 315]), is excluded from this list, as it is discussed in detail in the following chapters.

publication	year	target	N_s	N_c	spatial granularity	annotations
Akbari et al. [9]	2008	abdomen	1	5	pixels	semantic
Fabelo et al. [93]	2016	brain	22	4	mixed	sparse
Ravì et al. [279]	2017	brain	18	2	pixels	sparse
Fabelo et al. [94]	2018	brain	22	4	mixed	sparse
Garifullin et al. [110]	2018	retina	?	3	images	semantic
Moccia et al. [239]	2018	abdomen	7	6	superpixels	semantic
Fabelo et al. [92]	2019	brain	16	4	pixels, patches	sparse
Trajanovski et al. [340]	2019	ex-vivo specimen	14	2	patches	semantic
Cervantes-Sanchez et al. [59]	2021	abdomen	7	4	pixels, patches	sparse
Collins et al. [70]	2021	ex-vivo specimen	22	2	pixels, patches	sparse
Trajanovski et al. [339]	2021	ex-vivo specimen	14	2	pixels, patches, mixed	semantic
Garcia Peraza Herrera et al. [109]	2023	mouth	30	35	pixels, images	sparse
Leon et al. [202]	2023	brain	34	4	pixels, mixed	sparse
Lotfy et al. [214]	2023	ex-vivo specimen	30	3	superpixels, mixed	sparse
Bannone et al. [30]	2024	abdomen	169	13	patches	sparse

dividing SI cubes into smaller regions generates more training samples, thereby mitigating overfitting [70]. However, the relationship between the number of training subjects and segmentation performance – and whether smaller spatial granularities provide a tangible advantage over larger ones – has not yet been systematically studied.

- **Annotation sparsity**: The majority of studies (67 %) rely on sparse annotations that do not accurately delineate tissue boundaries. Instead, these annotations are for example confined to small circular regions within the target organs [59] or limited to tissue areas with available histopathological evidence [93]. This limitation impedes the practical applicability of such algorithms for tissue segmentation, as (1) the algorithms are not trained to accurately identify tissue boundaries, and (2) their performance in regions outside the sparse annotations cannot be assessed, making the reported segmentation performances unreliable.
- **Flaws in algorithm validation**: Many studies lack an independent test set, instead reporting performance metrics on validation sets that were also used for hyperparameter tuning. This practice risks overestimating algorithm performance [9, 279, 92, 59, 339].

In summary, existing research on surgical scene segmentation remains limited, particularly in the context of open surgeries. The optimal spatial granularity of input data to achieve high segmentation quality and reduce the number of required training subjects has yet to be established. Moreover, no prior study has conclusively demonstrated the superiority of SI data over RGB data for deep learning-based surgical scene segmentation, particularly in open visceral surgeries. To address these gaps in literature, we investigate the following research questions:

RQ2.1: What is the optimal spatial granularity of input data (pixels, superpixels, patches, or full images) in terms of segmentation performance and the number of required training subjects?

RQ2.2: Does HSI data offer advantages over RGB data and processed HSI data (e.g., tissue parameter estimations) for DL-based surgical scene segmentation?

5.2 Materials and Methods

The following sections describe our dataset (Section 5.2.1) and deep learning pipeline (Section 5.2.2) used for automated surgical scene segmentation.

5.2.1 Dataset

We collected a dataset of 506 HSI cubes from 20 pigs undergoing midline laparotomy, with a thoracotomy subsequently performed on a subset of 8 pigs. Each image was fully annotated with 18 distinct organ classes and background. The HSI data was collected at Heidelberg University Hospital with approval from the Committee on Animal Experimentation of the Regional Council of Baden-Württemberg, Karlsruhe, Germany (G-161/18 and G-262/19). The pigs were managed in compliance with German animal welfare laws and in accordance with the European Community Council Directive (2010/63/EU). Further information regarding the surgical procedures, anesthesia, and animals are provided in [329].

Hyperspectral Image Acquisition The HSI data was acquired using the medical device-graded camera system TIVITA® Tissue (Diaspective Vision GmbH, Am Salzhaff, Germany) described in Section 2.1.2. It consists of a push-broom HSI unit, halogen illumination, a computer for data acquisition and processing, and a monitor, all mounted on a mobile cart for convenient intraoperative measurements. The camera system captures 100 spectral channels in the visible and NIR range from 500–1000 nm with a spectral bandwidth of approximately 5 nm. It captures a field of view of about 30 cm×20 cm at an imaging distance of about 50 cm, which is maintained using an integrated distance calibration unit. The resulting HSI cubes measure 640×480×100 (width × height × spectral channels) and the acquisition of one HSI cube takes about 7 seconds. TPI, including StO_2, TWI, THI and NPI, were computed from the acquired HSI data using proprietary algorithms provided by the manufacturer [141]. Additionally, RGB images were generated based on the HSI data by combining reflectances across spectral channels corresponding to red, green, and blue channels of conventional RGB cameras, respectively [141].

To ensure uniform illumination from the camera's integrated halogen lighting unit, the room light was dimmed during image acquisition. To minimize motion artifacts, the entire camera system remained stationary throughout image acquisition, eliminating any potential camera motion. Additionally, images were captured from static scenes, with no movement of objects caused by the operating surgeon. Consequently, motion artifacts were limited to natural sources such as respiration and heartbeat, making them relatively mild and primarily affecting images of thoracic organs (see Figure 5.8 for an example image).

Hyperspectral Image Annotation Two medical experts performed the semantic annotation process using vector annotation tools on the SuperAnnotate platform (SuperAnnotate, Sunnyvale, USA) [7]. To ensure consistency, all annotations were reviewed and refined by the same medical expert.

Each pixel was assigned to one out of 19 classes, including:

- two thoracic organs, namely lung and heart.
- 8 abdominal organs, namely stomach, liver, gallbladder, spleen, pancreas, small bowel, colon, and kidney. Kidney images were taken prior to removing Gerota's fascia and afterwards, and labeled as "kidney with Gerota's fascia" and "kidney", respectively.
- one pelvic organ, namely bladder.
- the anatomical structures skin, subcutaneous fat, muscle, peritoneum, omentum, and major vein.
- the label "background", which was used for inorganic objects, such as metallic objects, compresses, cloth, tubes, gloves and foil. This label appears in every image, with annotated areas covering an average of 47 % (SD 24 %) of an image.
- the label "ignore", which was assigned to regions where the organ class was unclear or ambiguous, or where the area belonged to an organic structure outside the defined 18 tissue classes. The "ignore" label appears in 221 of the 506 images, covering on average 2 % (SD 3 %) of their area. These pixels were excluded from subsequent analysis.

An example image and segmentation for each of the 19 classes are shown in Figure 5.1. The figure also includes heatmaps that highlight the typical spatial distribution of each class within an image and the average class spectra aggregated across subjects.

Data Statistics An overview of the class distribution across the images is provided in Figure 5.2. For each organ, 32 to 405 images were collected from 5 to 20 subjects. As certain organs naturally appear more frequently within the field of view of others, the number of images per organ class varies. For example, because the liver encloses the gallbladder, the liver is consistently visible in gallbladder images, whereas the gallbladder does not appear in all liver images. Additionally, differences in the surgical procedures performed on the pigs led to variations in the number of subjects per organ class. For example, thoracotomy – a highly invasive and complex procedure associated with substantial mortality and extended operating times – could only be performed on 8 of the 20 subjects, resulting in missing lung and heart HSI data for the remaining 12 subjects.

5.2.2 Deep Learning Pipeline

The primary objective of our study was to systematically evaluate the performance of DL-based surgical scene segmentation algorithms across varying spatial granularities

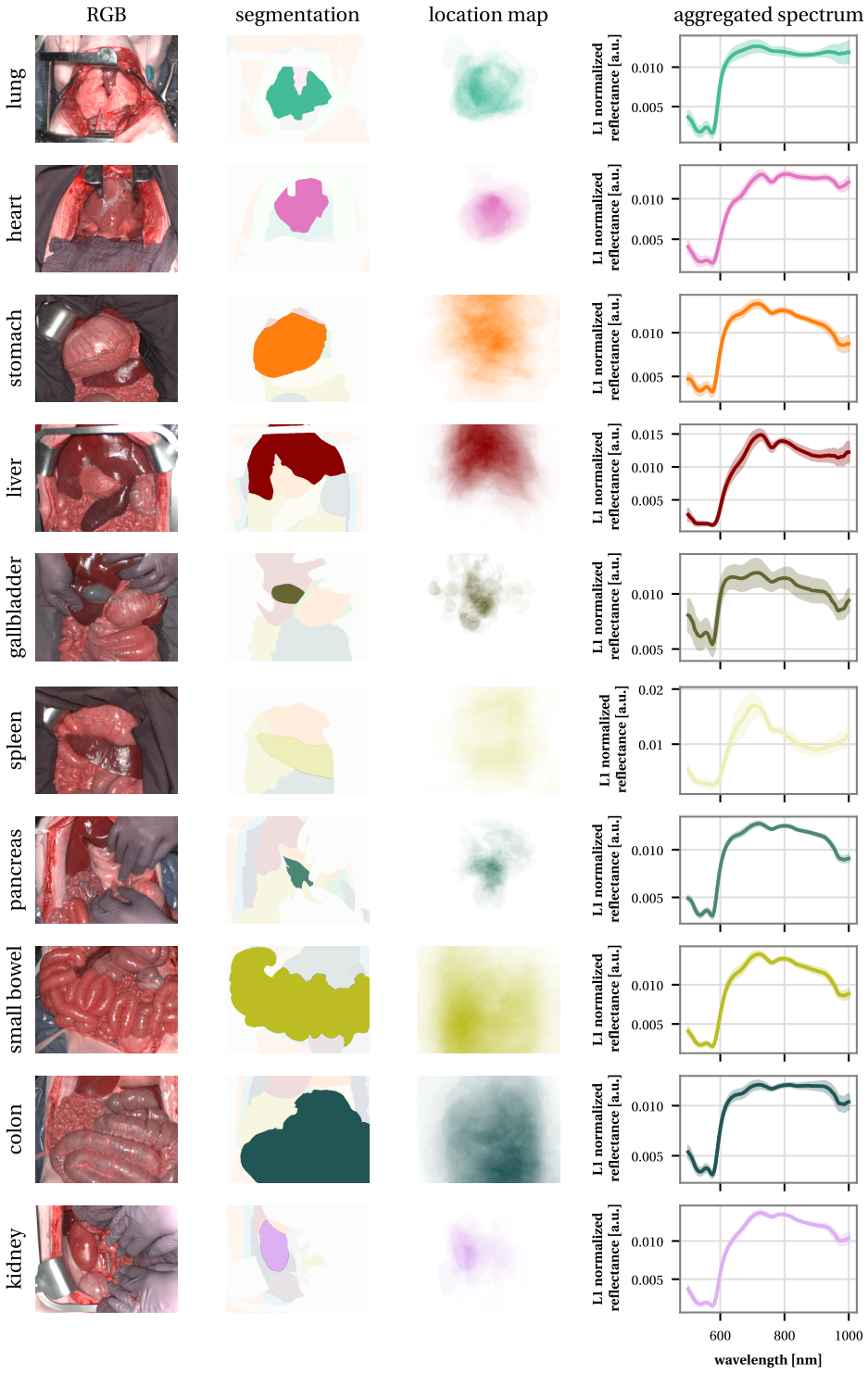
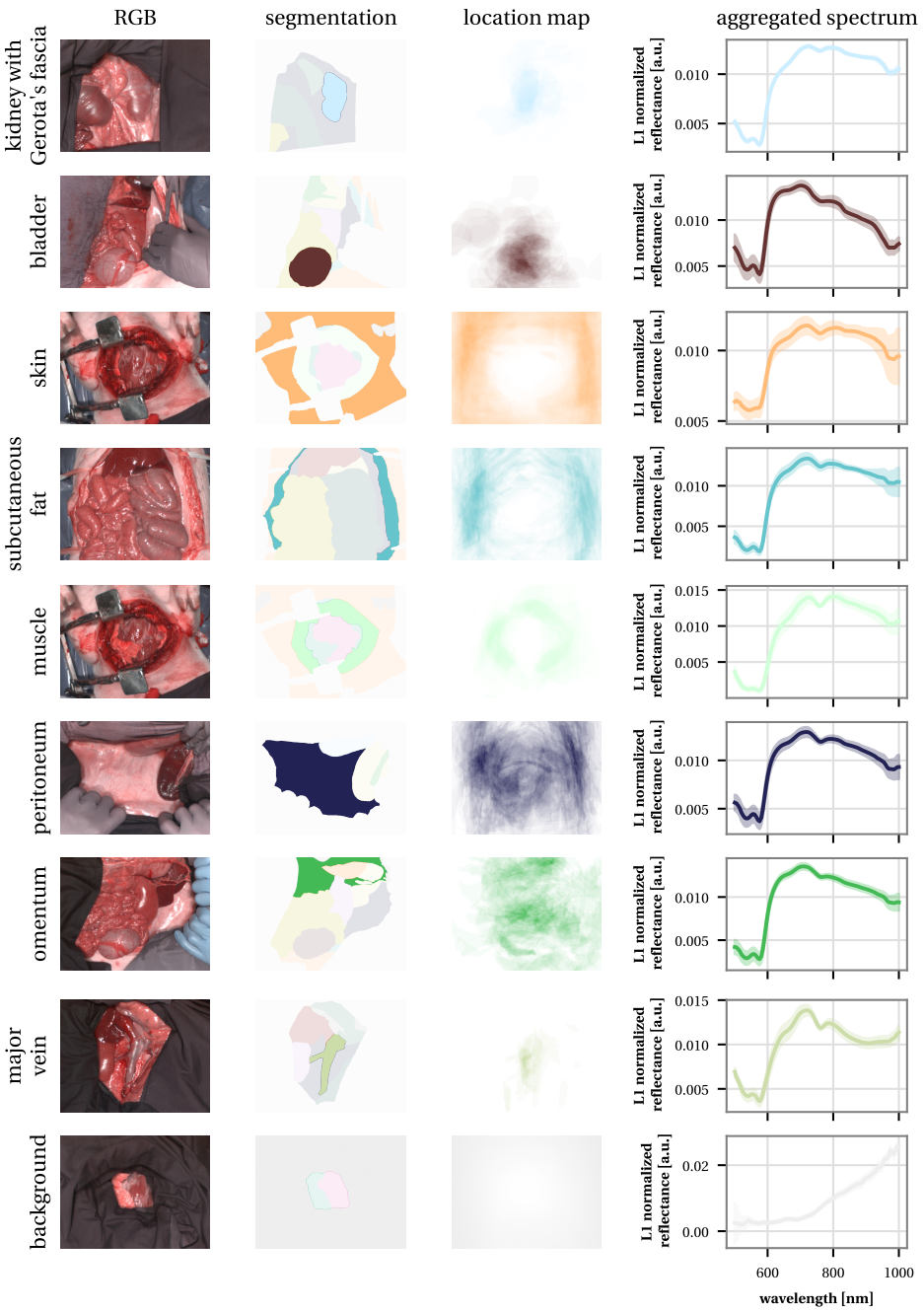


Figure 5.1: Characteristic locations and spectra for the 18 different organ classes and background. For each class, sample RGB images are displayed alongside their corresponding segmentations, location maps and characteristic spectra. Figure continued on the next page.



Continued Figure 5.1: Characteristic locations and spectra for the 18 different organ classes and background (continuation). Heatmaps of the location are generated by overlaying all available segmentations for the respective class. Exemplary RGB images were chosen to maximize the overlap between the respective class segmentation and the heatmap. Median spectra were aggregated at the subject level, with the overall mean spectrum depicted as a solid line and the shaded area representing the standard deviation across subjects. Figure adapted from [311].

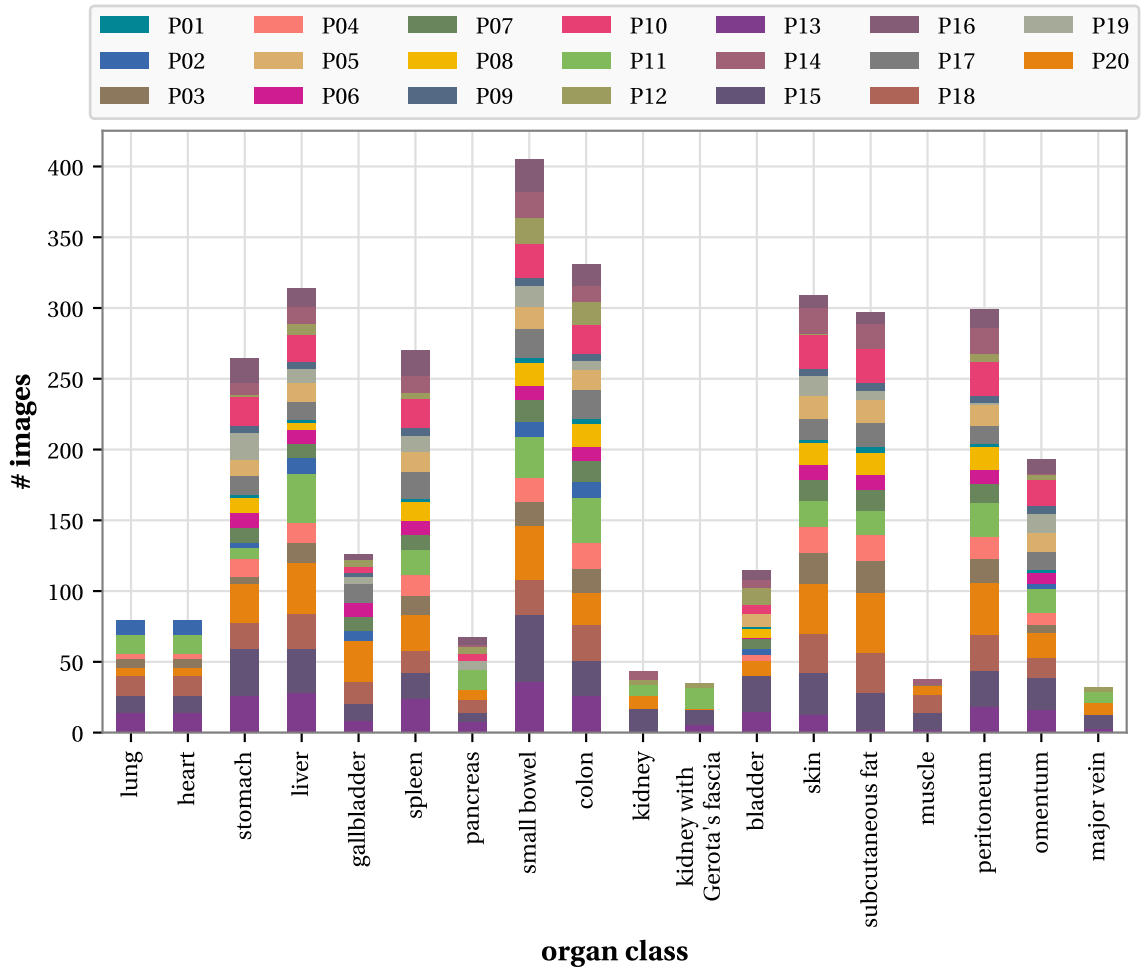


Figure 5.2: Dataset overview. The dataset comprises 506 images from 20 pigs, each with fully semantic annotations covering 18 distinct organ classes and background. The bar plots illustrate the number of images per organ class, with each pig uniquely identified by a subject-specific identifier, P*xx*, and represented by a distinct color. Figure adapted from [308].

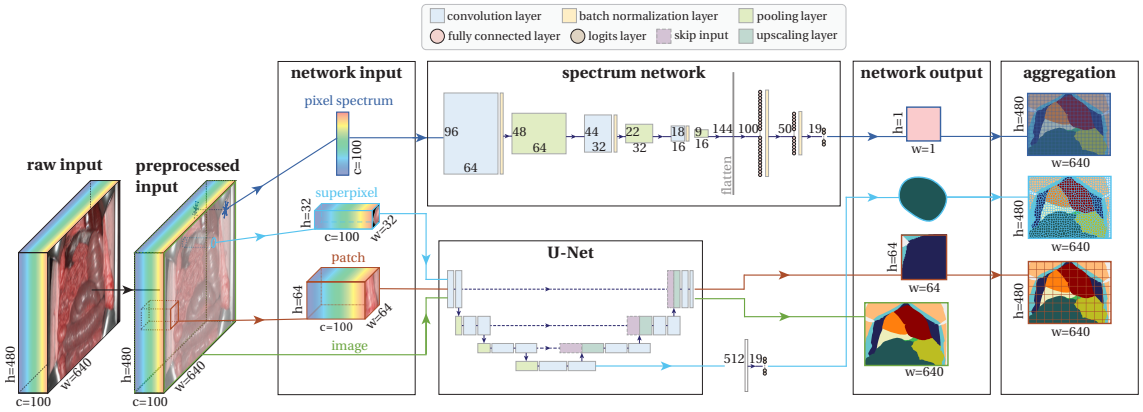


Figure 5.3: Overview of our deep learning pipeline for automated surgical scene segmentation based on hyperspectral imaging (HSI) data. After preprocessing the HSI data, including calibration and normalization, networks were trained to handle different spatial granularities: pixels, superpixels, patches, and images. Pixel spectra were input into the spectrum network for classification. For superpixel-based classification, superpixel boundaries were identified from reconstructed RGB images. For each superpixel, a minimum enclosing bounding box was computed, with pixels outside the superpixel set to zero. The resulting superpixel cube was processed through a U-Net encoder, followed by a classification head. For patch-based segmentation, fixed-shape patches were extracted from the preprocessed HSI cube and analyzed using a U-Net. For image-based segmentation, the entire preprocessed HSI cube was fed into the U-Net, yielding an image segmentation map. For pixel-wise, superpixel-wise, and patch-wise approaches, predictions from the same image were aggregated to construct an image segmentation map. Figure adapted from [308, 311].

and modalities of the input data. To achieve this, we developed a comprehensive DL pipeline capable of processing imaging data at 5 distinct spatial granularities: pixels, superpixels, patches of two different shapes, and entire images. The pipeline was designed to handle multiple imaging modalities, including RGB, HSI, and processed HSI data, the latter of which was generated by stacking TPI data for StO_2, NPI, TWI and THI. Figure 5.3 offers an overview of our DL pipeline for automated surgical scene segmentation, with further details provided in the following sections.

Data Preprocessing The HSI cubes were first calibrated using white and dark reference cubes to eliminate sensor noise and convert the spectra from radiance to reflectance [141]. Following calibration, ℓ^1-normalization was applied across the spectral channels to compensate for multiplicative changes in illumination, such as variations in the measurement distance.

Pixel-Based Segmentation Individual pixel spectra represent the smallest possible input data spatial granularity, corresponding to feature vectors of length $c = 100$ for HSI, $c = 4$ for TPI and $c = 3$ for RGB data.

Building on our previous work [329], the DL model for HSI spectra comprises 3 one-dimensional convolutional layers, using 64 filters in the first, 32 filters in the second, and 16 filters in the third layer. Each convolution uses a kernel size of 5 and after each convolutional layer, an average pooling layer is applied across the spatial dimensions with a kernel size of two. The output from the final convolutional layer is flattened and fed into two fully connected layers, with the first layer containing 100 neurons and the second layer containing 50 neurons. A final linear layer computes the class logits, with the predicted class label determined by taking the argmax of these logits. To generate a segmentation map for the entire image, class label predictions are collected for each individual pixel.

Due to the small channel size, convolutional operations across channels are not feasible for TPI and RGB input data. Therefore, the model is composed of 3 fully connected layers, comprising 200, 100 and 50 neurons in the first, second and third layer, respectively.

The model architecture was selected for its simplicity and effectiveness in analyzing spectral information. The convolutional layers capture local spectral patterns, while stacking 3 layers with a small kernel size efficiently expands the receptive field. The fully connected layers make decisions based on the global context, allowing the model to balance local and global information processing while remaining computationally efficient, with only 34 300 trainable weights for the HSI, 27 819 weights for the TPI, and 27 619 weights for the RGB modalities.

The ELU activation function [67] (cf. Section 2.3.3) was employed, with batch normalization applied to all layers except for the pooling layers. The model was optimized using the cross-entropy (CE) loss function.

Superpixel-Based Segmentation Superpixels, defined as regions of low spatial granularity that conform to local boundaries, are constructed by grouping together pixels with similar characteristics. Analogous to the pixel-based surgical scene segmentation, unsupervised superpixel clustering converts the segmentation task into a classification problem. In this approach, each superpixel is assigned a single class label based on the assumption that the entire area of a superpixel covers the same class.

To generate superpixels, the simple linear iterative clustering (SLIC) algorithm [3] was employed on RGB images that had been smoothed using a Gaussian kernel of width 3. The algorithm was configured to produce 1000 superpixels per image and perform 10 iterations, while dynamically adjusting the compactness parameter for each superpixel (SLICO mode). Subsequently, a minimum enclosing bounding box was calculated for

each superpixel, and pixel values outside the superpixel were replaced by zeros. The resulting superpixel cubes were resized via bilinear interpolation to a uniform shape of $32 \times 32 \times c$ (where c denotes the number of channels), ensuring a standardized input format across all superpixel cubes.

The superpixel cubes are processed by an EfficientNet B5 encoder [336] from the Segmentation Models PyTorch library by Yakubovskiy [372], pretrained on the ImageNet dataset [78]. This encoder was selected for its strong performance, efficiency in terms of number of parameters, minimal memory usage, and fast computation speed. To compute the class logits, the output of the encoder network is forwarded to a classification head, which consists of a fully connected layer with 19 neurons. During inference, the superpixel-wise class label is obtained by taking the argmax of these logits, and predicitons over all superpixels of an image are collected to generate an image segmentation map.

During training, fuzzy labels were employed in place of one-hot-encoded labels for the superpixels, enabling the model to account for the possibility that pixels within a superpixel belong to different classes. This approach assigns a label vector of length 19 to each superpixel, indicating the relative frequency of each class label among the pixels within the superpixel. As loss function, the Kullback-Leibler divergence [189] between the softmax output and the fuzzy labels was utilized.

Patch-Based Segmentation Patches are defined as regions of rectangular shape, with each patch containing a fixed number of pixels. In our study, we extracted patches of two different sizes from the image cubes: $32 \times 32 \times c$, referred to as patch_32, and $64 \times 64 \times c$, referred to as patch_64, where c denotes the number of channels. These sizes serve as intermediate levels of spatial granularity between the superpixel and image models (see Table 5.3). Furthermore, patch dimensions that are powers of two facilitate the integration with encoder architectures that halve the input dimensions multiple times.

To facilitate the comparison across different spatial granularities, the patches, similar to the superpixel-based segmentation, are processed using a U-Net [291] (cf. Section 2.3.3) with an EfficientNet B5 encoder [336], which was pretrained on the ImageNet dataset [78].

During training, an average of the Dice loss [236] and CE loss were computed based on valid pixels[1]. The inclusion of the Dice loss addresses class imbalance in the dataset by placing greater emphasis on misclassified pixels from underrepresented classes compared to those from overrepresented classes, such as the background. In contrast, the CE loss treats all misclassified pixels equally. By combining these two loss functions, their respective strengths are effectively leveraged [153].

[1] Pixels that do not belong to the “ignore” class.

While patches were extracted at random locations of the image during training, images were divided into a grid of non-overlapping patches during inference. As for the patch_64 model, one of the image dimension was not an integer multiple of the patch size (480/64 = 7.5), the overhanging areas of the grid were padded with zeros. The segmentation maps generated for each individual patch were aggregated to generate a complete image segmentation map, with segmentations corresponding to previously zero-padded regions excluded.

Image-Based Segmentation Entire images represent the maximum level of spatial granularity in the input data. The DL model architecture and loss functions for image-based segmentation are identical to those used in the patch-based model, except that the entire image cube is provided as input to the network.

Table 5.3: Epoch and batch sizes across the different spatial granularities. The number of pixels (# pixels), along with the epoch and batch sizes, are detailed for the 5 spatial granularities: image, patch_64 and patch_32 – referring to patches with dimensions $64 \times 64 \times c$ and $32 \times 32 \times c$, respectively, where c represents the number of channels – as well as superpixel and pixel. Table adapted from [308, 311].

spatial granularity	# pixels	epoch size	batch size
image	307 200	500	5
patch_64	4096	37 632	336
patch_32	1024	150 528	1176
superpixel	$\approx$ 300	500 760	1560
pixel	1	153 608 400	118 800

Training Setup To enable a systematic and fair comparison of the different spatial granularities and modalities, data augmentations, model optimization, and training budget were standardized, ensuring maximum comparability of the training setup across all models.

Data augmentations are widely used in computer vision to expand the diversity and size of the training data, thereby boosting convergence, generalization, and robustness to OOD data [54]. For all spatial models, training data augmentation was applied at the image level, prior to extracting smaller granularities such as pixels, superpixels, or patches, using the Kornia library [287]. Augmentations included shifting (shift factor limit: 0.0625), scaling (scaling factor limit: 0.1), rotating (rotation angle limit: ± 45°), and flipping (both horizontally and vertically). To balance computational efficiency with augmentation effectiveness, the probability p of applying an augmentation was set to $p = 0.5$.

All models used the Adam optimization algorithm [177] ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with an exponential learning rate schedule (initial learning rate $\eta = 0.001$, decay rate $\gamma = 0.99$). The models were trained for 100 epochs, and stochastic weight averaging (SWA) [156] was applied over the final 20 epochs.

For image-based segmentation, one epoch was defined as processing 500 images. For segmentation based on the other spatial granularities, the number of samples per epoch was adjusted such that the total number of extracted pixels approximately matched the total pixel count of 500 images (cf. Table 5.3). This approach ensured an approximately uniform training budget across all spatial granularities. Exact equivalence was not achievable because the epoch size needs to be an integer multiple of the batch size to ensure a balanced workload distribution among all workers in the data loader (see [308] for further details).

Recommendations in the literature regarding the optimal batch size are mixed (e.g., [322, 167]). Smaller batch sizes can accelerate the learning process by enabling more frequent updates [240], while larger batch sizes offer a better representation of the overall population, which can lead to more stable gradient estimates and improved batch statistics [152]. To strike a balance, we maximized the batch size while extending the training over a large number of epochs to mitigate the potential slowdown in the learning process. In practice, the maximum achievable batch size is determined by the available GPU memory, as well as the memory demands of the model and input samples. Consequently, the batch size was optimized individually for each model, with smaller spatial granularities allowing for larger batch sizes. The resulting batch sizes are detailed in Table 5.3.

Each model was evaluated on the validation set at the end of each training epoch by calculating the Dice similarity coefficient (DSC), with the hierarchical structure of the data taken into account. This score was subsequently used to select the best-performing model across all epochs.

To prevent overfitting, dropout regularization with a probability of $p = 0.1$ was applied to the fully connected layers in the pixel and superpixel models.

Reduction of Network Variability Training neural networks involves several sources of variation [263] that need to be minimized as much as possible to obtain reproducible outcomes and enable a fair model comparison across spatial granularities and modalities. Achieving perfectly reproducible results often requires longer training times, such as by using deterministic operations or a single, homogeneous hardware setup [263]. Due to the large number of training runs required for our study, we were unable to enforce deterministic operations and a single hardware setup. Instead, we exploited a cluster infrastructure with heterogeneous GPUs (e.g., NVIDIA® DGX™ A100, NVIDIA® GeForce RTX™ 2080 Ti (Nvidia Corporation, Santa Clara, United States of America)).

Nevertheless, we implemented several measures to reduce the network variability: We set the number of workers in each data loader to 12 and used a random seed for the training process to ensure consistent initialization of network weights and workers across all experiments. The consistent initialization of the data loading workers ensured that all models for a given spatial granularity received samples from the same spatial locations and in the same order. Additionally, the exact same sequence of data augmentations was applied across all models.

5.3 Experiments and Results

The purpose of our experiments was to identify the optimal input spatial granularity for DL-based surgical scene segmentation (RQ2.1, Section 5.3.2), considering segmentation performance and the amount of training data required. Additionally, we explored whether HSI data offers advantages over RGB data and processed HSI data (RQ2.2, Section 5.3.3). Details of the experimental setup are provided in Section 5.3.1.

5.3.1 Experimental Setup

We trained and validated our DL pipeline on the dataset described in Section 5.2.1. In the following, we provide an overview of our dataset splits, validation metrics, and the approach used for hierarchical aggregation of the results and uncertainty-aware ranking of our models. Furthermore, we describe how we evaluated the quality of our reference annotations, and present our experiment on the required amount of training data.

Dataset Splits A consistent training and validation setup was applied across all models. We divided the dataset, which comprises 506 images of 20 pigs, at the subject level, yielding a training set of 15 pigs and 340 images and a hold-out test set of 5 pigs and 166 images. The test pigs were selected randomly, ensuring that all 18 organ classes were represented in both the training and test sets. We performed 5-fold cross-validation on the training set, with folds constructed to maximize the number of organ classes across validation folds. Once model development was completed, we assessed segmentation performance on the previously untouched test set by ensembling the predictions from all 5 folds, averaging the softmax values.

Validation Metrics Following the recommendations from [284, 222, 283], we assessed the segmentation performance of our models using 3 metrics: the DSC[2] [79] (an

[2]The DSC measures the overlap between reference and predicted object segmentation.

overlap-based metric), the normalized surface Dice (NSD)[3] [250] (a boundary-based metric that accounts for annotation uncertainty), and the average surface distance (ASD)[4] [135] (a boundary-based metric). Each of these metrics has its strengths and limitations (as discussed in [283]), and by combining them, we sought to provide a comprehensive evaluation of our models.

Several design choices were required regarding the metrics: For the ASD, there is no consensus on handling missing classes[5] [135]. We opted to assign the ASD value for a missed class to the maximum ASD observed among other classes in the same image, introducing a potentially substantial image-dependent penalty when a class was not predicted. For the NSD, a clinically acceptable deviation threshold τ between the reference and predicted segmentation boundaries must be defined. Given the variation in annotation difficulty across different organs, we established a class-specific threshold τ_c for each organ class c. To determine these thresholds, 20 randomly selected images (one per pig, ensuring at least two images per organ class) were once more annotated by a second medical expert. Distances between the boundaries of the original annotation and the re-annotation were computed for each organ c and image i, and an image- and organ-specific threshold τ_c^i was obtained by averaging these distances. If an organ was missed in one of the annotations, the organ was excluded from the analysis, as distances could not be calculated. The final class-specific threshold τ_c was determined as the average of the image-specific thresholds τ_c^i. The implications of these design choices are further discussed in Section 5.4.1.

Our dataset follows a hierarchical structure, with each subject comprising one or more images, and each image containing multiple classes. To account for this hierarchy (following [140, 222]), we first aggregated metric values at the image level, yielding image-wise scores, and subsequently aggregated the scores from all images of one subject, resulting in subject-wise scores. While this approach introduces the limitation that image-level scores are influenced by the class distribution within an image, it offers the advantage of enabling an analysis of performance variability across subjects.

Uncertainty-Aware Model Ranking We evaluated model rankings and their stability concerning two sources of variability: sampling variability and metric choice. Following the method outlined in [364], model rankings were established based on the average metric value calculated across the 5 subject-level metric values in the test dataset. To assess the impact of metric choice on ranking stability, we independently performed rankings using each validation metric and compared the results.

[3]The NSD quantifies the proportion of the predicted object boundary that lies within a clinically acceptable deviation from the reference boundary.

[4]The ASD computes the average distance between reference and predicted object boundaries.

[5]Classes present in the reference annotations but not in the predictions.

To examine the stability of rankings under sampling variability, we applied bootstrapping. Specifically, we generated 1000 bootstrap samples, each comprising 5 subject-level metric values randomly selected with replacement from the 5 available subject-level metric values in the test dataset. For each bootstrap sample, the metric values were averaged, and models were ranked based on these aggregated scores, resulting in 1000 ranks for each model.

Quality of Reference Annotations To evaluate the accuracy of our reference annotations, we analyzed both inter-rater and intra-rater variability. Inter-rater variability was assessed using the set of 20 re-annotated images previously utilized for determining the NSD thresholds, while intra-rater variability was estimated by having the original medical expert re-annotate the same set of 20 images. Annotation pairs were compared using the DSC, NSD, and ASD metrics. To avoid penalizing differences in the assignment of the "ignore" class – since expressing uncertainty is a valid annotation choice – pixels labeled as "ignore" in either of the annotations were excluded from the analysis.

Training Size Experiment To evaluate model performance as a function of the available training data, a random sample of n pigs was drawn from the training set of 15 pigs without replacement, with n varying from one to 14. Our different models were then re-trained exclusively on the images from the n sampled pigs without using 5-fold cross-validation, and their performance was evaluated on the test dataset. To mitigate variations in data availability across pigs for different classes, performance was measured only for the 8 classes consistently represented across all training subjects: skin, peritoneum, spleen, liver, colon, small bowel, stomach and background. To enhance stability in the presence of inter-pig variability, the experiment was repeated 5 times, each with a different random seed, facilitating that a new set of pigs was sampled for each iteration.

5.3.2 Optimal Spatial Granularity

Our experiments aimed to determine the optimal input spatial granularity for DL-based surgical scene segmentation with respect to segmentation performance and the amount of training data needed.

Quality of Reference Annotations In addition to evaluating the segmentation performance of our models, we assessed the quality of our reference annotations by analyzing the inter-rater and intra-rater variability. This analysis serves two key purposes: (1) to identify the challenges faced by human medical experts in segmenting surgical

scenes, and (2) to provide context for interpreting model performance, as high-quality annotations are essential for training reliable models and for accurately validating their performance. We observed a DSC of 0.89 (SD 0.07), an NSD of 0.80 (SD 0.08), and an ASD of 4.88 (SD 5.33) for the inter-rater variability. The intra-rater agreement is slightly better than the inter-rater agreement, at a DSC of 0.91 (SD 0.05), NSD of 0.82 (SD 0.06), and ASD of 4.74 (SD 5.04). These results highlight the inherent difficulty of accurately segmenting surgical scenes, even for medical experts.

Figure 5.4 illustrates the inter-rater agreement for all 20 selected images. The intra-rater agreement is shown in the appendix (Figure B.9). In addition to discrepancies at organ boundaries, additional classes not present in the original reference segmentation map were annotated 8 times in the inter-rater comparison and 6 times in the intra-rater comparison. Conversely, classes present in the reference segmentation map were omitted 7 times in the inter-rater and 4 times in the intra-rater evaluations. Differences involving the "ignore" class were noted in 14 of the 20 images for both the inter-rater and intra-rater comparisons, respectively. Specifically, there were 34 063 px instances for the inter-rater case and 37 397 px instances for the intra-rater case where the ignore label was assigned to a pixel that had been assigned a different label in the reference annotation, or vice versa.

Segmentation Performance Figure 5.5 presents the test performance of our segmentation models across all 5 spatial granularities (pixel, superpixel, patch_32, patch_64, and image) and 3 modalities (RGB, TPI, and HSI), evaluated using the DSC, NSD and ASD metrics. While the performance differences across spatial granularities are less pronounced for HSI data compared to RGB and TPI data, larger input spatial granularities consistently lead to better segmentation performance. Notably, the best performing model – the image-based segmentation model using HSI data – achieved a DSC of 0.90 (SD 0.04), an NSD of 0.80 (SD 0.07) and ASD of 6.19 (SD 3.20), which is comparable to the performance obtained for a second medical expert (the inter-rater performance).

The ranking stability with respect to sampling variability is illustrated in Figure 5.6 for the DSC, with results for the NSD and ASD shown in Figure B.10 and Figure B.11, respectively. The rankings are largely in agreement. Notably, across all metrics, the first rank and last two ranks are highly stable, with more than 90 % of bootstraps yielding the same rank. Compared to the DSC, the ranking variability is smaller for the NSD and ASD.

A comparison of the rankings obtained for the different metrics is provided in Figure 5.7. Across all modalities and metrics, the spatial granularities consistently rank in the order: image, patch_64, patch_32, superpixel, and pixel (from best to worst). This finding reaffirms the observation from Figure 5.5, demonstrating that increased contextual information consistently enhances segmentation performance, regardless of the modality or metric. Overall, the rankings across different metrics are closely

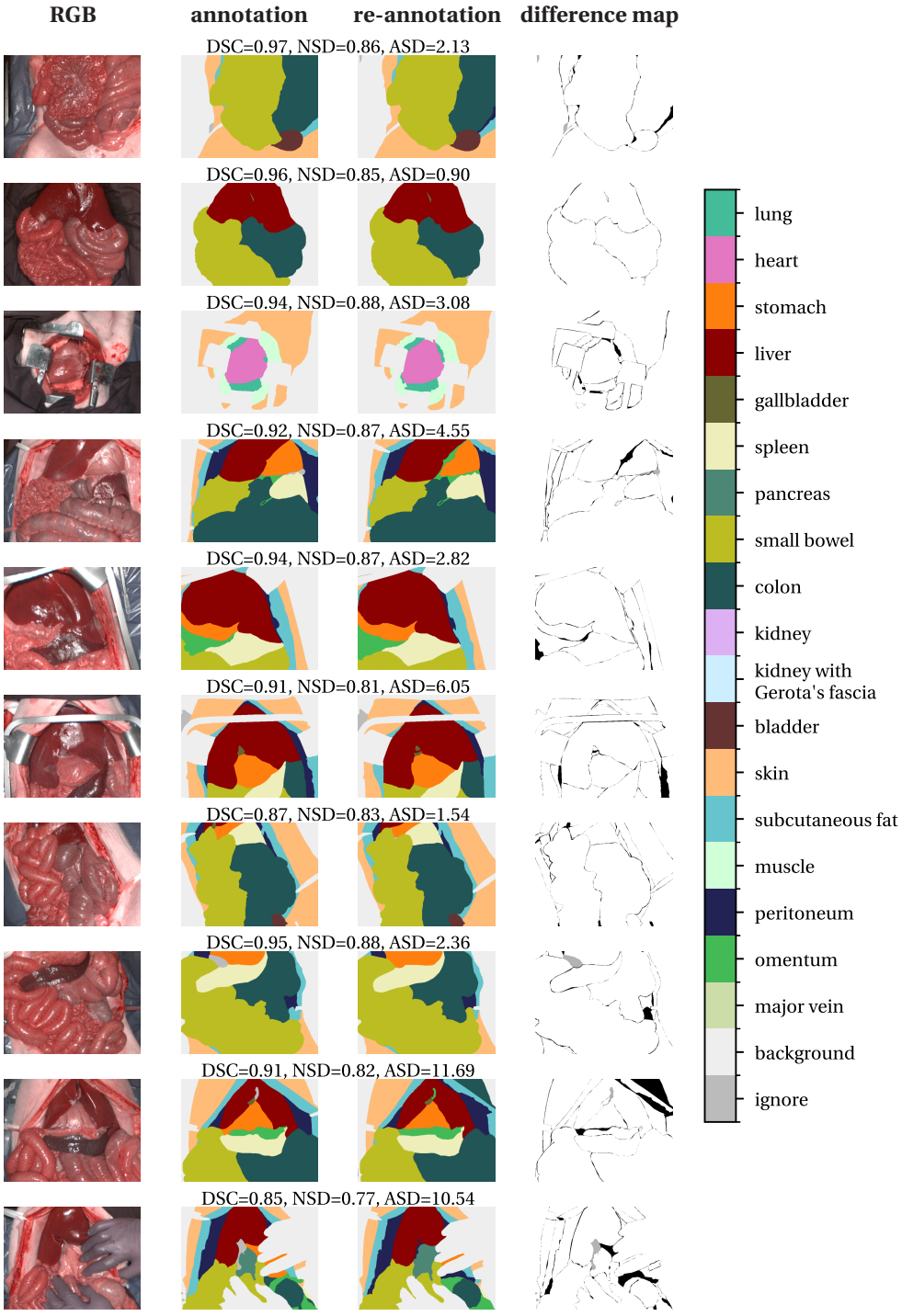
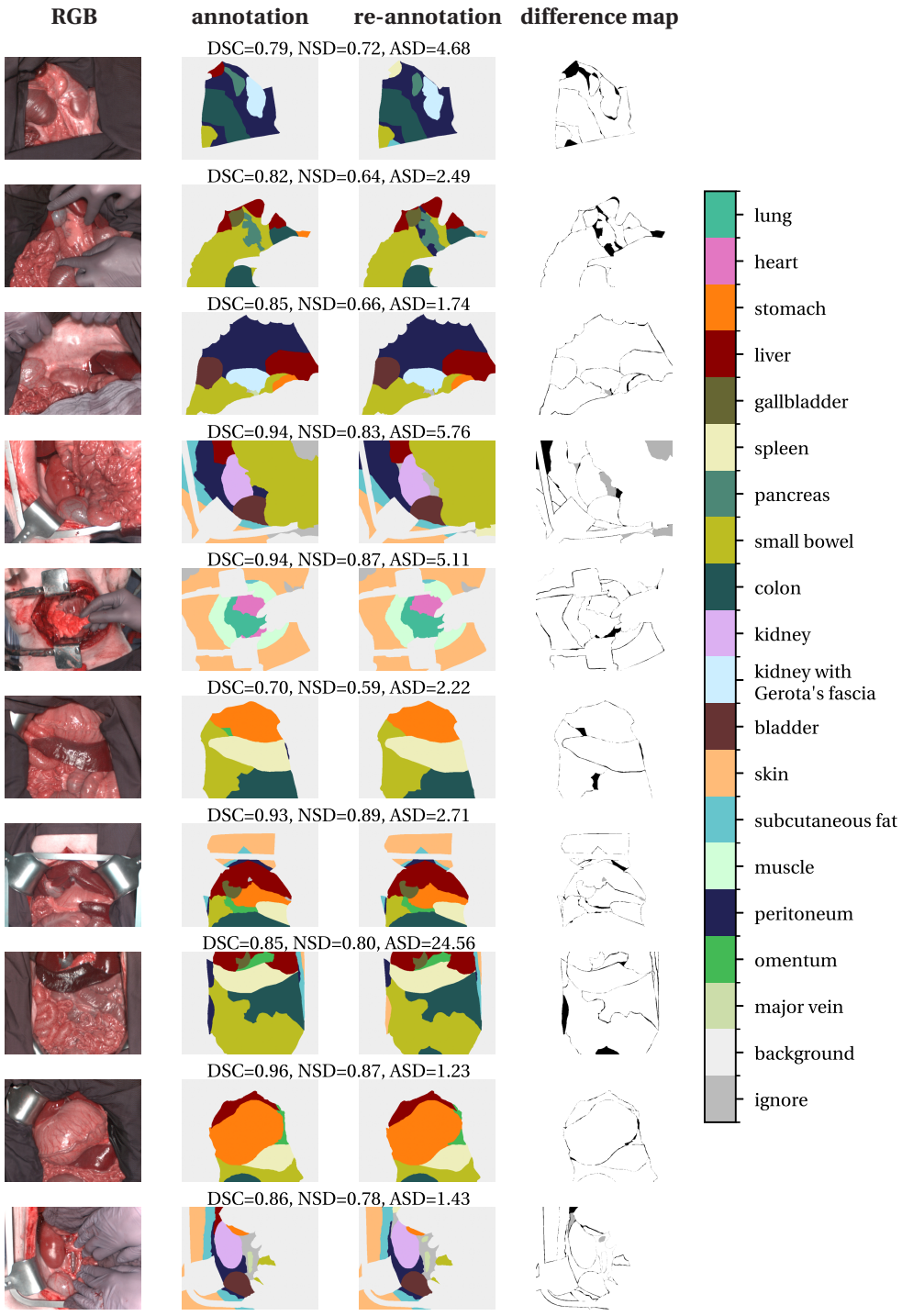


Figure 5.4: Inter-rater agreement of reference annotations. Re-annotations of the 20 selected images are shown with their RGB images, original annotations, and difference maps between the annotations. Figure continued on the next page.



Continued Figure 5.4: Inter-rater agreement of reference annotations (continuation). Mismatches between the "ignore" class and a valid class are highlighted in gray, while discrepancies between valid classes are marked in black.

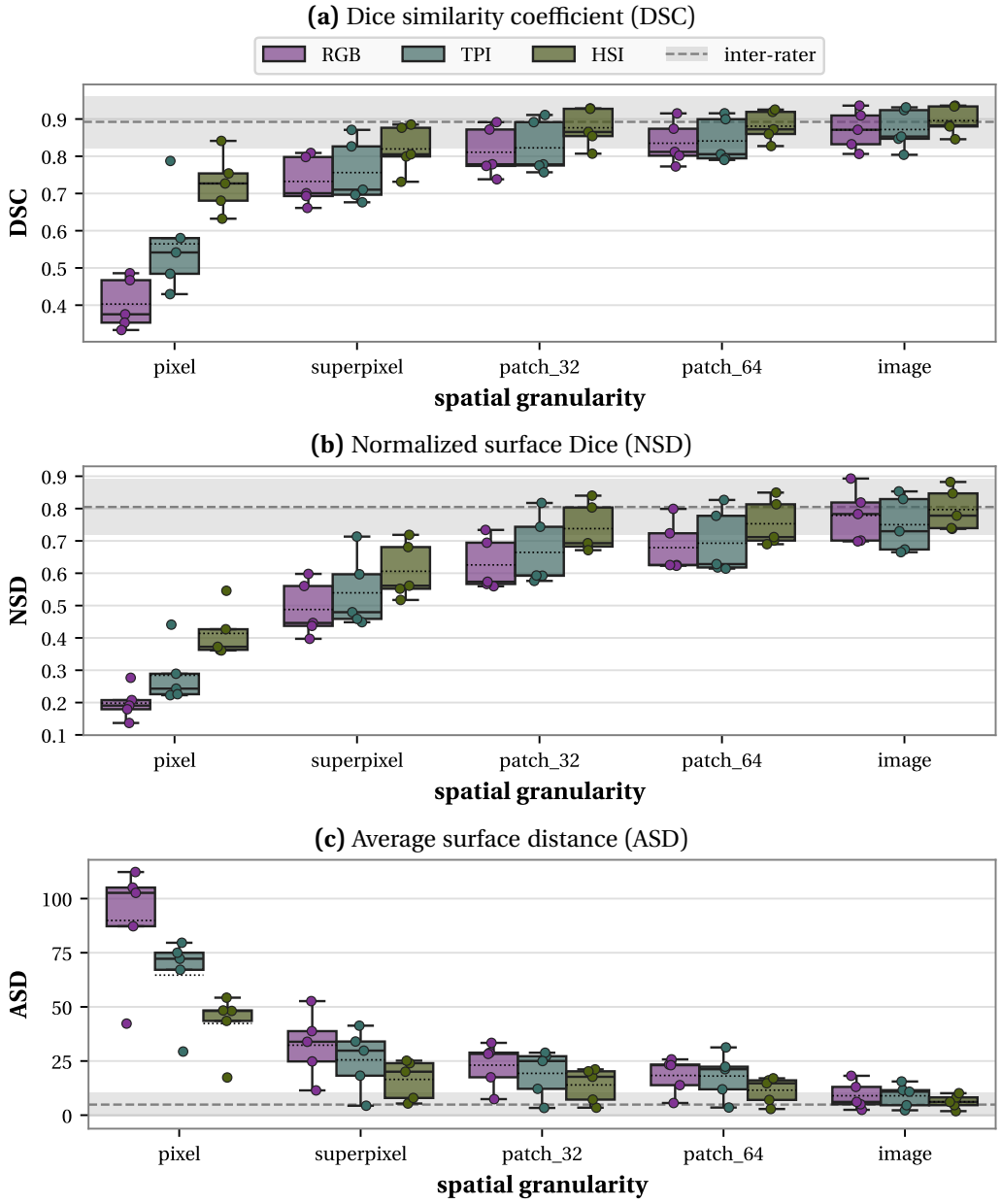


Figure 5.5: Segmentation performance across different spatial granularities and modalities. The boxplots illustrate the segmentation performance across different spatial granularities (pixel, superpixel, patch, and image) and modalities (RGB, tissue parameter images (TPI), and hyperspectral imaging (HSI)), evaluated using 3 different metrics (a-c). Each box displays the interquartile range of the distribution, with whiskers showing the range excluding outliers, and median and mean indicated by a solid and dotted line, respectively. Each marker represents one test subject. The dashed line indicates the mean of the inter-rater performance, with the standard deviation denoted as shaded area. Figure adapted from [308, 311].

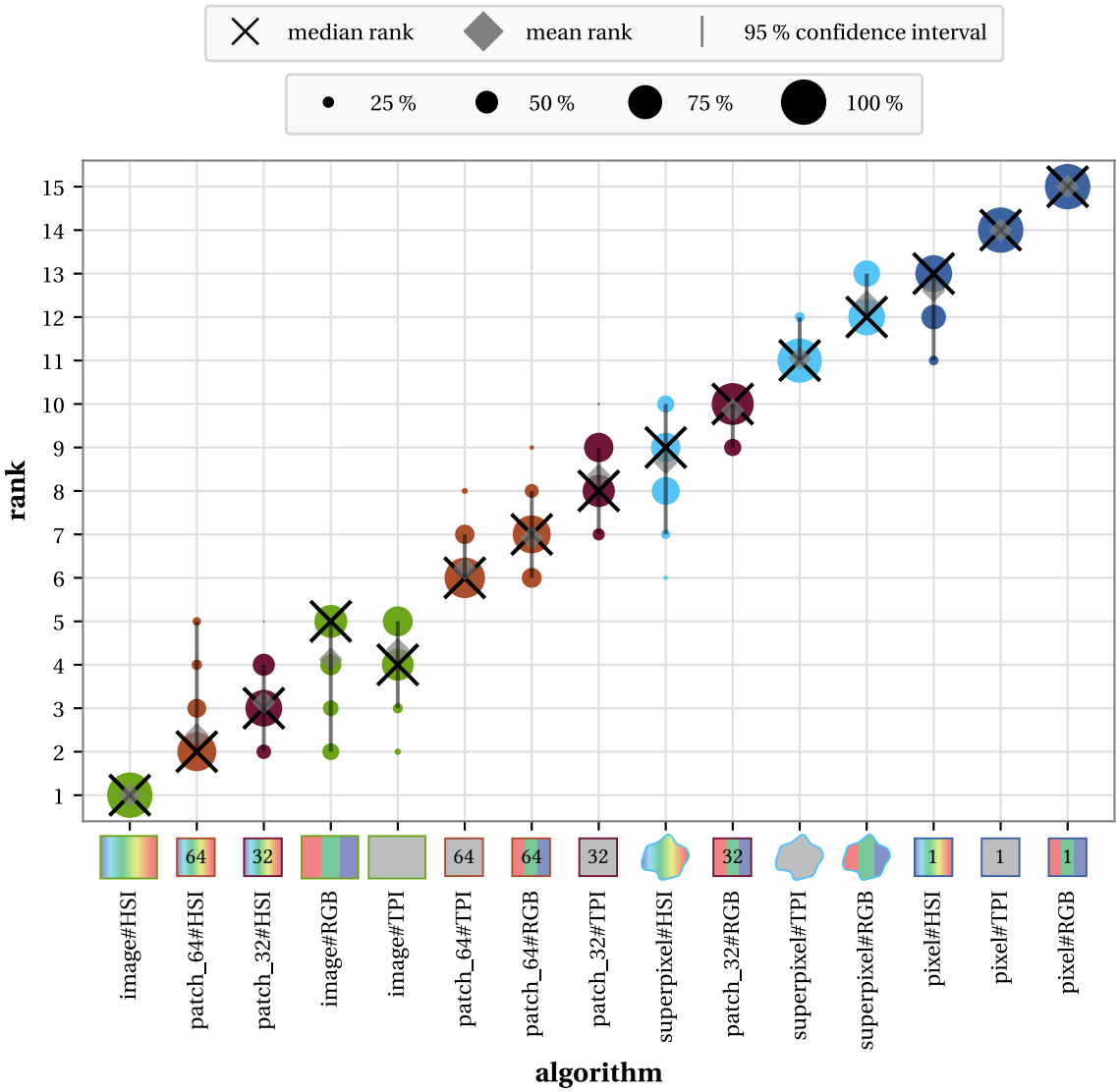


Figure 5.6: Ranking stability of our segmentation algorithms with respect to sampling variability using the Dice similarity coefficient (DSC). Following the concept from [364], bootstrap sampling was performed to assess the ranking stability of our segmentation algorithms across different spatial granularities (pixel, superpixel, patch_32, patch_64 and image) and modalities (RGB, tissue parameter images (TPI), and hyperspectral imaging (HSI)) and modalities (RGB, tissue parameter images (TPI), and hyperspectral imaging (HSI)). For each blob at position (a, rank r), its area is proportional to the frequency of algorithm a achieving rank r across 1000 bootstrap samples. Each sample comprises 5 subject-level DSC values. For each method, black crosses indicate the median rank, gray diamonds show the mean rank, and gray lines represent the 95 % quantile of the bootstrap results. Ranking stability figures for the normalized surface Dice (NSD) and average surface distance (ASD) are available in Figure B.10 and Figure B.11, respectively. Figure adapted from [308, 311].

aligned: image-based segmentation using HSI data consistently ranks first, while the bottom 5 positions are consistently occupied (from best to worst) by superpixel#TPI, superpixel#RGB, pixel#HSI, pixel#TPI and pixel#RGB. The most notable discrepancy in rankings across metrics is observed for the superpixel#HSI model, which achieves rank 6 for the ASD but falls to ranks 9 and 10 for the DSC and NSD, respectively. This discrepancy may be attributed to the ASD metric's sensitivity to boundary alignment and will be further discussed in Section 5.4.1.

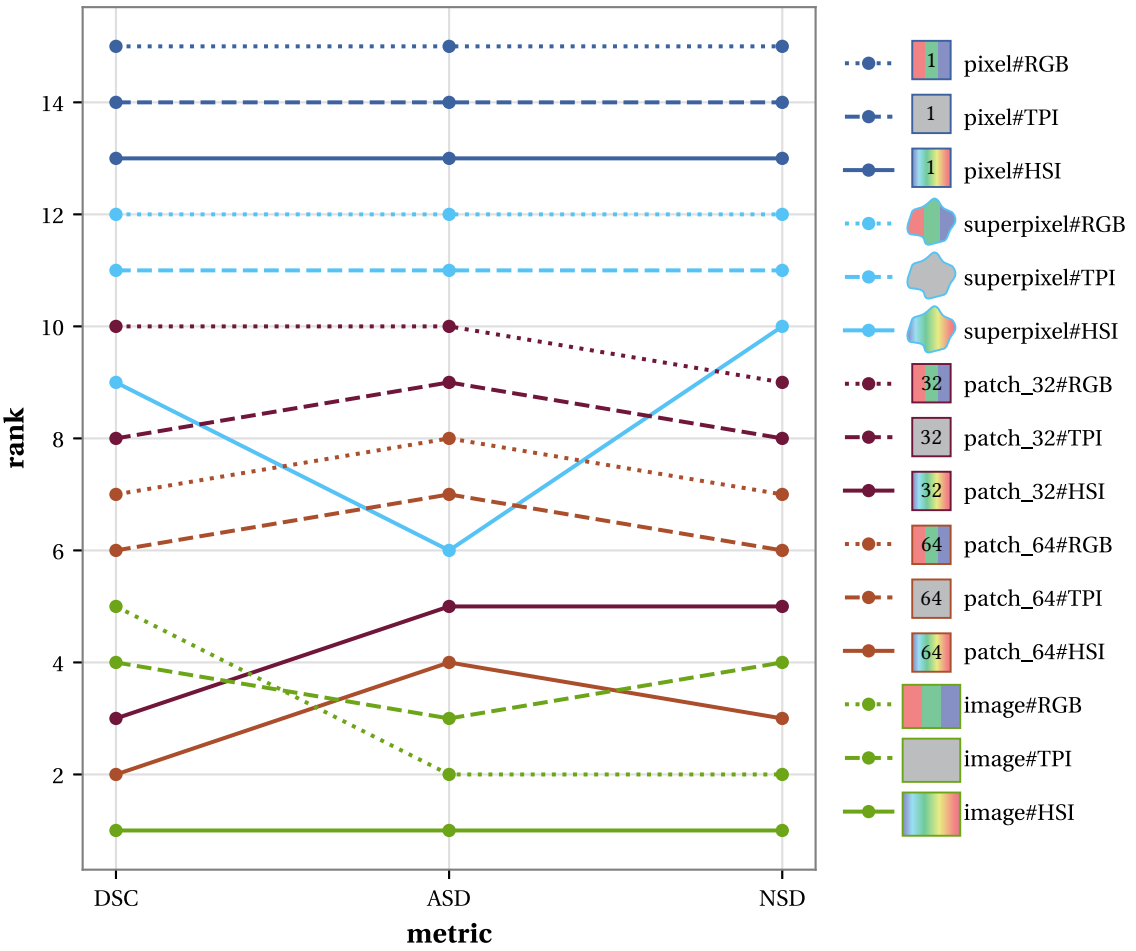


Figure 5.7: Ranking stability of our segmentation algorithms across 3 different metrics. Following the concept from [364], each line illustrates how the ranking of our segmentation algorithms varies across different spatial granularities (pixel, superpixel, patch_32, patch_64 and image) and modalities (RGB, tissue parameter images (TPI), and hyperspectral imaging (HSI)) when evaluated using 3 different metrics: Dice similarity coefficient (DSC), average surface distance (ASD) and normalized surface Dice (NSD). Figure adapted from [308, 311].

Visual Comparison of Segmentation Quality Figure 5.8 showcases example predictions across the 5 spatial granularities using HSI-based models. The images illustrate cases of poor, intermediate, and good segmentation performance, corresponding to the 5 %, 50 %, and 95 % quantiles of the average image DSC across all 5 models. Segmentation artifacts characteristic of each spatial granularity are evident: Pixel-based segmentation predictions exhibit fragmented and scattered boundaries. In superpixel-based segmentation, organ boundaries appear irregular, a result of misclassified superpixels in the boundary regions. In patch-based segmentation examples with poor performance, prominent vertical and horizontal edges are noticeable at patch boundaries, resulting from the intentionally non-overlapping patch extraction during inference (see Section 5.4.1 for a discussion of this design choice).

Amount of Training Data Required Several researchers in the related work have suggested that dividing SI cubes into smaller regions to generate more training samples could be advantageous in data-limited settings, thereby advocating for the use of smaller spatial granularities in such cases [59, 70]. Figure 5.9 illustrates the relationship between the number of training subjects and the segmentation performance of HSI-based models across different spatial granularities and performance metrics. While performance generally improves with an increasing number of training subjects, image-based segmentation consistently outperforms or matches the performance of other spatial granularities, regardless of the number of training subjects and validation metric. Conversely, pixel-based segmentation performs the worst, followed by superpixel-based segmentation. These results indicate that the additional contextual information provided by larger spatial granularities outweighs potential benefits of generating more training samples using smaller spatial granularities. It is worth noting that the observed decrease in SD with an increasing number of training subjects should be interpreted with caution. As the number of training subjects increases, the overlap between randomly selected subjects increases due to the limited pool of 15 training subjects available for sampling without replacement. For example, when selecting two sets, each comprising 14 training subjects, the overlap between the sets can include up to 13 subjects, significantly reducing the variability between them.

5.3.3 Comparison of Modalities

A primary purpose of our study was to investigate whether there is a substantial benefit in HSI-based surgical scene segmentation over RGB and TPI data.

Segmentation Performance As illustrated in Figure 5.5, the average segmentation performance using HSI data is superior to that of using TPI or RGB data across all

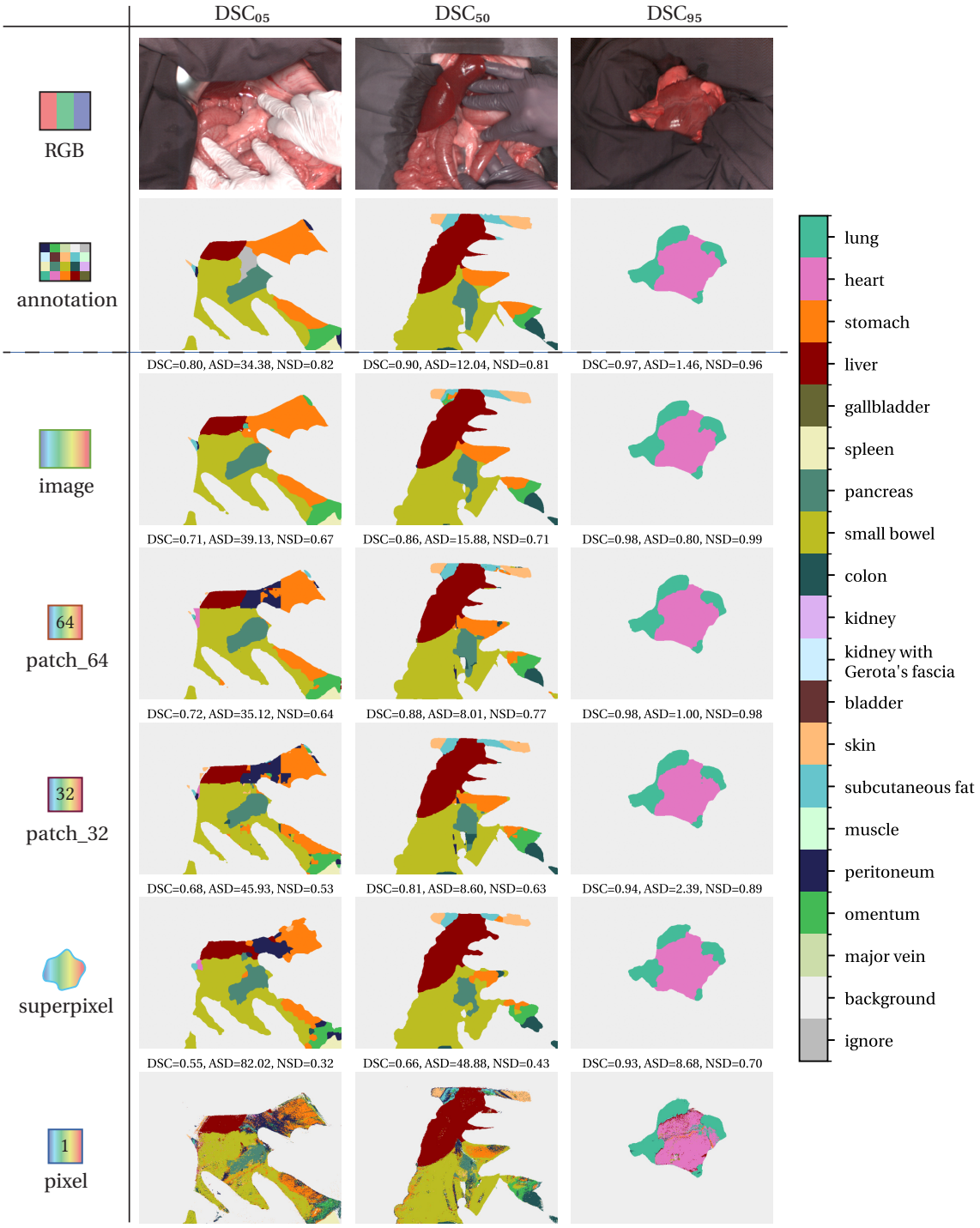


Figure 5.8: Example predictions for hyperspectral imaging-based segmentation algorithms across different spatial granularities. The images were sampled based on the average q % quantile of the Dice similarity coefficient (DSC) across all 5 granularities (pixel, superpixel, patch_32, patch_64 and image), denoted as DSC_q. Corresponding values of DSC, average surface distance (ASD), and normalized surface Dice (NSD) are displayed for each prediction. Figure adapted from [308, 311].

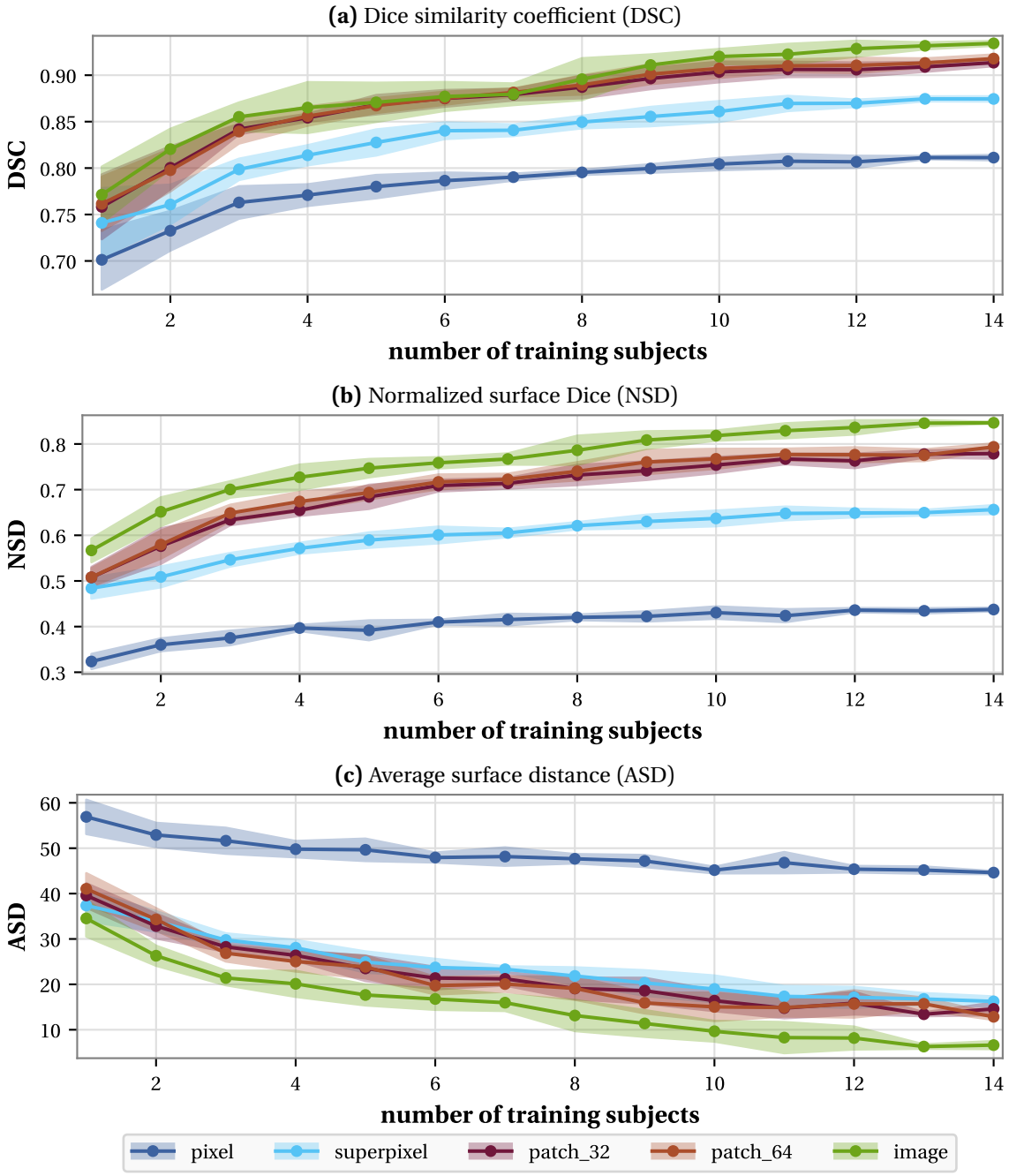


Figure 5.9: Performance of hyperspectral imaging-based segmentation algorithms across different spatial granularities as a function of the number of training subjects. To account for sampling variability, n training subjects ($n \in \{1, 2, \ldots, 14\}$) were randomly selected without replacement from the full set of training subjects, with this process repeated 5 times using different random seeds. The average performance across these runs is shown as a solid line, while the shaded area represents one standard deviation. Figure adapted from [308, 311].

spatial granularities and metrics. Consequently, models based on HSI data consistently achieve higher rankings than their TPI and RGB counterparts, as shown in Figure 5.7. In most cases, TPI-based models outperform their RGB counterparts. The performance gap is most pronounced in pixel-based segmentation, where using HSI instead of RGB results in a DSC improvement of 80.4 %. This gap diminishes as the spatial granularity increases, with improvements of 11.9 % for superpixel-based segmentation, 8.1 % for patch_32, 5.5 % for patch_64, and 2.8 % for image-based segmentation. However, it is worth noting that as the performance of image-based models is comparable to the level of inter-rater agreement, the small performance differences between image-based models using different modalities may also be influenced by the quality of our reference annotations.

Misclassification Analysis To gain a deeper understanding of the segmentation performance across different modalities, we analyzed the confusion matrices for image-based segmentation using HSI, TPI, and RGB data. The confusion matrix for HSI data is shown in Figure 5.10, while the confusion matrices for TPI and RGB data are provided in the appendix (Figure B.12 and Figure B.13, respectively). Additionally, the recall for the 3 models is displayed in Figure 5.11, highlighting notable variations in segmentation performance across different classes. For the image#HSI model, on average over 95 % of the pixels were correctly classified for 8 out of the 19 classes. The major vein exhibits the lowest recall, with only 57.1 % of its pixels correctly identified. Generally, the recall for the image#HSI model is higher or comparable to that for the image#TPI and image#RGB models for most classes, with the only exceptions being major vein and pancreas.

For the image#HSI model, the highest confusion rate of 32.2 % occured between the classes major vein and peritoneum. This is likely due to their proximity and the limited training data available for the major vein, which is present in only 32 images (see Figure 5.2). Furthermore, the visible regions of the major vein are relatively small with an average size of 4192 px (SD 3621 px). Other frequently misclassified classes include those with indistinct boundaries (e.g., omentum, peritoneum, subcutaneous fat) or those with challenging differentiation from other structures (e.g., kidney with Gerota's fascia and peritoneum). Many misclassifications in the confusion matrix involve adjacent classes within the images (e.g., stomach and omentum, heart and lung, liver and gallbladder, background and skin). These errors are likely driven by inaccuracies in the predicted segmentation boundaries, as highlighted in the examples shown in Figure 5.8.

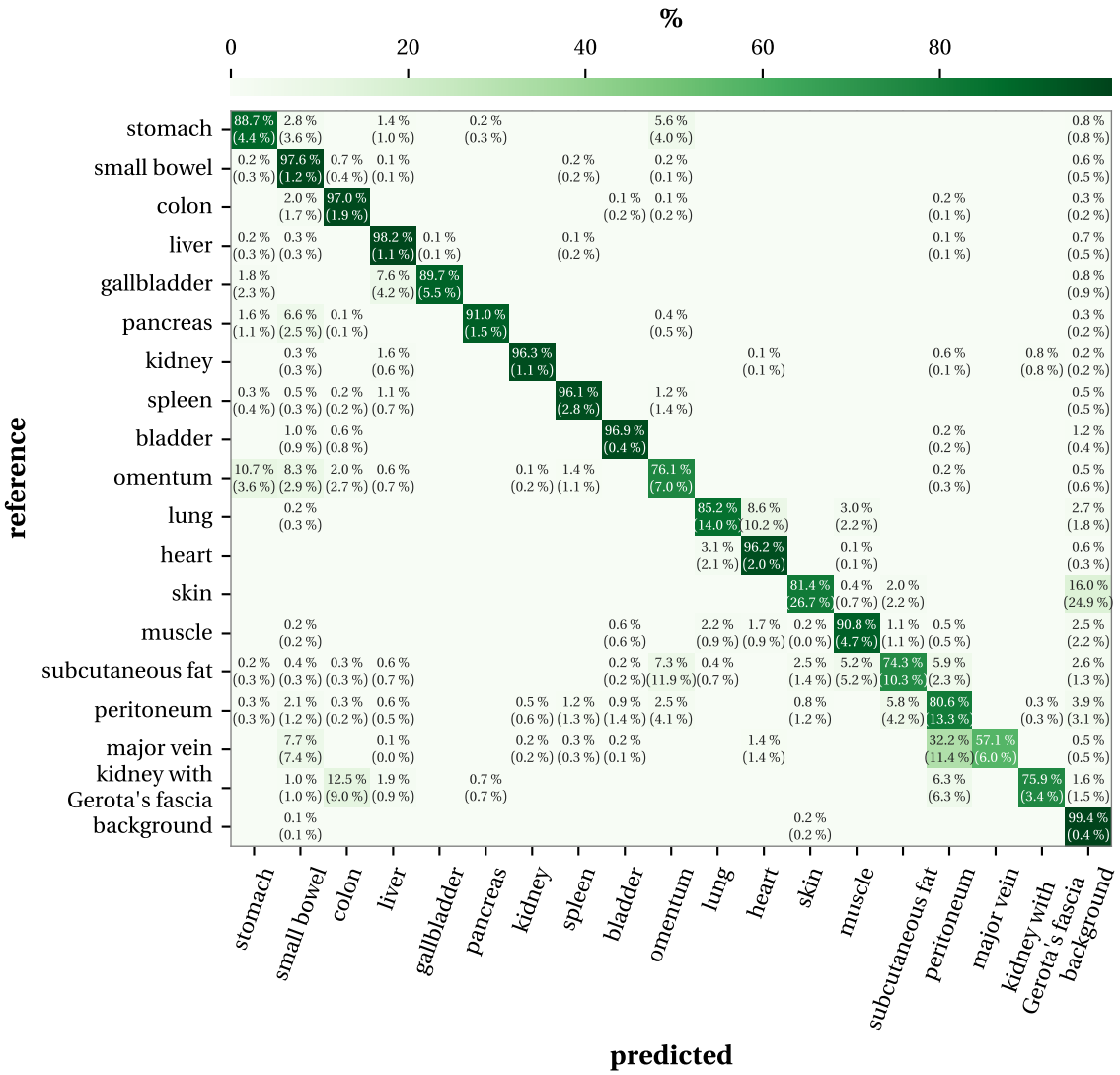


Figure 5.10: Confusion matrix for image-based segmentation using hyperspectral imaging data. Each entry (i, j) denotes the average proportion of pixels from the reference class i that are classified as class j, with values below 0.1 % omitted for clarity. Confusion matrices were row-normalized using pixel data from all images of a single subject, and the subject-specific matrices were averaged across subjects to produce the final confusion matrix. The standard deviation across subjects is indicated in brackets. Diagonal entries correspond to recall (sensitivity). Figures for the tissue parameter images and RGB modalities are provided in Figure B.12 and Figure B.13, respectively. Figure adapted from [308, 311].

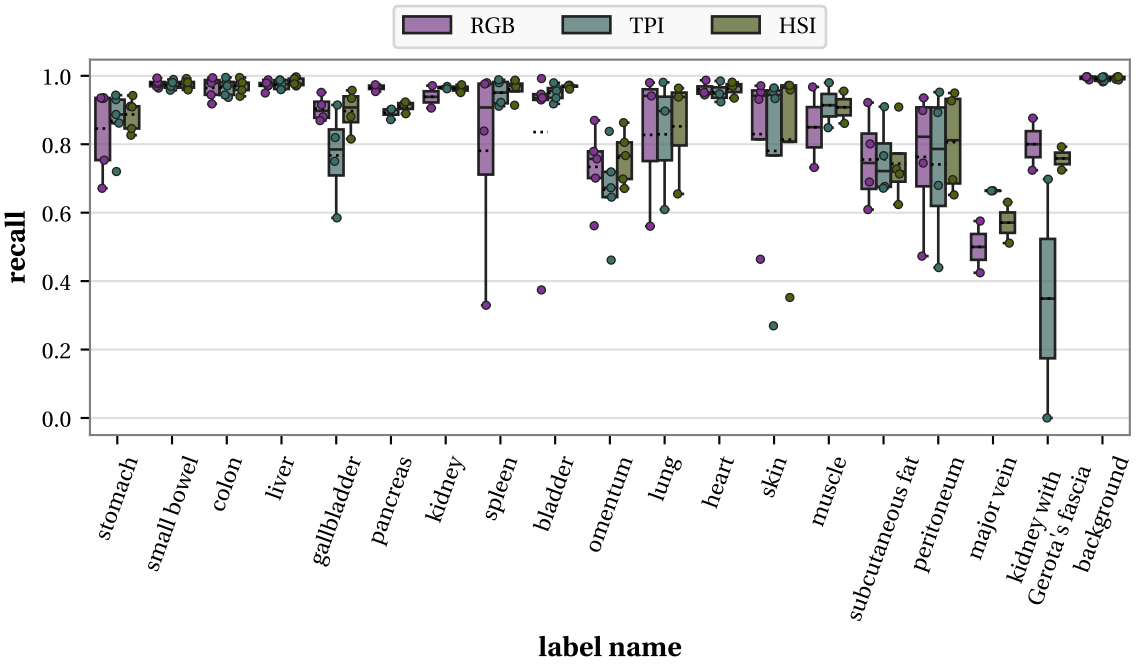


Figure 5.11: Recall of image-based segmentation across modalities and classes. The boxplots illustrate the recall across the modalities RGB, tissue parameter images (TPI), and hyperspectral imaging (HSI), as well as the 19 different classes. Each box displays the interquartile range of the distribution, with whiskers showing the range excluding outliers, and median and mean indicated by a solid and dotted line, respectively. Each marker represents one test subject. Figure adapted from [308, 311].

5.4 Discussion and Conclusion

In this study, we explored two important and previously unanswered research questions in DL-based surgical scene segmentation: (1) What is the optimal spatial granularity of input data in terms of segmentation performance and the required amount of training data? (2) Does HSI data offer substantial advantages over other modalities, such as TPI and RGB data? Our main findings are:

1. **Optimal spatial granularity:** Across all validation metrics and modalities, segmentation performance consistently improved with increasing input spatial granularity. Notably, image-based segmentation using HSI data achieved performance on par with re-annotations by a second medical expert.
2. **Required amount of training data:** Regardless of the number of training subjects, image-based segmentation with HSI data consistently outperformed or matched the performance of all other spatial granularities.

3. **Benefit of HSI data:** HSI-based models consistently outperformed their TPI and RGB-based counterparts across all spatial granularities and metrics, with the largest benefit observed at smaller spatial granularities.

The following sections provide a discussion of our design choices (Section 5.4.1), key strengths and limitations of our study as well as potential directions for future research (Section 5.4.2), and a conclusion summarizing our findings (Section 5.4.3).

5.4.1 Design Choices

The primary goal in designing our study was to ensure a fair and comprehensive evaluation of the segmentation performance across different spatial granularities and modalities. To achieve this, we made several design choices regarding validation metrics, model hyperparameter optimization, and postprocessing techniques. In the following, we discuss the rationale behind these decisions, along with their associated advantages and limitations.

Validation Metrics Following the recommendations from [284, 222, 283], we employed multiple metrics to evaluate our results and establish rankings. Specifically, we used an overlap-based metric (DSC), a distance-based metric (ASD), and a boundary-overlap-based metric that accounts for annotation uncertainty (NSD).

Each metric captures distinct aspects of the predicted segmentation map, leading to differences in model rankings depending on the metric used. For instance, a notable shift in the ranking of the superpixel#HSI model was observed when comparing rankings based on ASD with those based on DSC and NSD (cf. Figure 5.7). While the DSC- and NSD-based rankings position the patch-based counterparts ahead of the superpixel#HSI model, the ASD metric ranks the superpixel#HSI model as superior to its patch-based counterparts. Figure 5.8 reveals that sharp vertical and horizontal edges appear in patch-based predictions, while superpixel boundaries more closely match the annotated boundaries. As boundary-distance metrics like the ASD are particularly sensitive to boundary misalignment, combining multiple metrics is crucial for obtaining a comprehensive model evaluation.

A critical factor influencing the metrics is the strategy for dealing with missing classes in the prediction. Evaluation frameworks such as Medical Open Network for AI (MONAI) [56] typically yield `nan` or `inf` values in these cases, requiring the user to determine how to handle aggregation. This design choice is particularly crucial for the ASD metric, as it is unbounded. Several strategies exist for handling missing classes, such as disregarding them entirely or applying a fixed penalty, that is, for example, based on the image diagonal. In order to prevent the introduction of outliers, we opted to

set the ASD for missing classes to the maximum distance from the other classes in the image. However, this approach has the downside that the value assigned to the missing class is influenced by the predictions of the other classes in the image.

For the NSD metric, it is required to define a (class-specific) threshold for the clinical acceptable deviation of the segmentations. To establish these thresholds, a subset of images needs to be re-annotated by at least one additional human annotator [250]. As obtaining re-annotations for many images is often impractical, this subset is typically small (e.g., 20 images in our case). Consequently, errors in these re-annotations substantially impact the NSD results. Missing classes in the re-annotations also pose a challenge, as distances cannot be calculated for these missing classes. In the original publication on the NSD, this issue did not occur, as a re-annotation was performed separately for each known class [250]. However, in our study, the annotators could not be informed about which classes were present in the image, as the identification of different tissue types is an important aspect in the determination of inter-rater variability.

Another challenge lies in selecting the appropriate aggregation function to determine the threshold τ_c from the set of inter-rater distances obtained for class c. In Figure 5.12, we present several thresholds derived using different aggregation functions. The choice of aggregation function substantially impacts the resulting thresholds. Originally, Nikolov et al. utilized the 95 % quantile of the distances [250], which in our case resulted in very high thresholds, exceeding 80 px. Therefore, we utilized the mean, resulting in moderate distances consistently below 20 px. However, other aggregation methods, such as the median or alternative quantiles, could also have been suitable choices.

Furthermore, considerable differences in thresholds across classes can be observed (cf. Figure 5.12), with the largest thresholds obtained for peritoneum and the smallest for bladder. Additionally, for some classes the thresholds substantially vary across subjects – for example, the SD for the mean aggregation of skin is 2.5 times higher than the mean itself. These findings highlight that the difficulty of annotation varies between organs and reinforce our decision to establish class-specific thresholds.

Hyperparameter Optimization For our segmentation models, we used default hyperparameters wherever possible and ensured consistent hyperparameter settings across algorithms (e.g., the learning rate). When deviations were necessary, we based them on consistent criteria, such as optimizing the batch size according to GPU memory usage. However, hyperparameters can substantially impact the network performance, and given the diversity of our model architectures and input sizes, our chosen settings are unlikely to be optimal for all algorithms. Identifying the ideal hyperparameter set for each algorithm would require extensive training runs. Since training all 15 models (spanning 5 spatial granularities and 3 modalities) across 5 folds already consumed

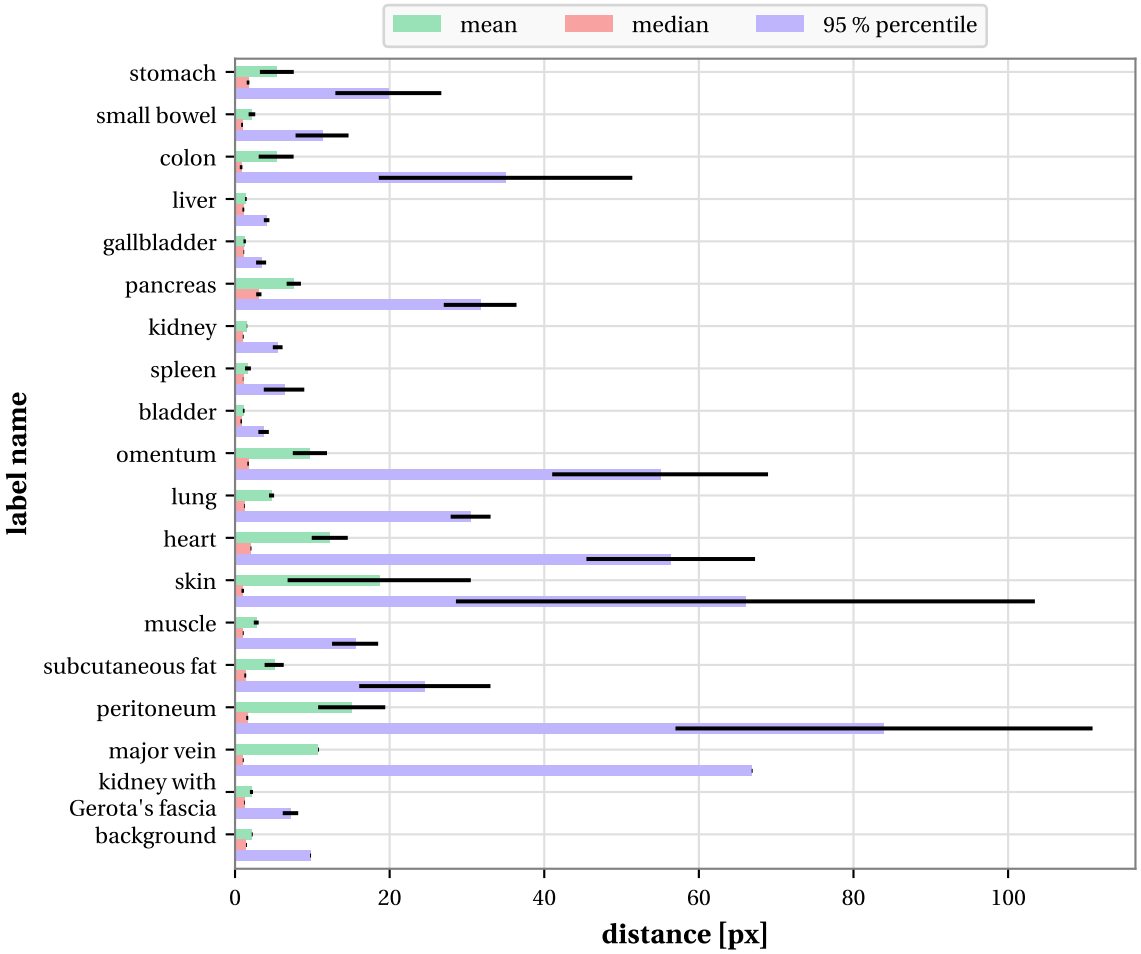


Figure 5.12: Comparison of different aggregation functions to determine class-wise distance thresholds for the normalized surface Dice (NSD) metric. Based on independent annotations from two experts, the class specific distance thresholds were determined using the mean, median and 95 % percentile of the set of distances between the paired annotations. The error bars represent 0.25 standard deviations of the aggregated values across subjects. The thresholds obtained for mean aggregation were applied in our analysis. Figure adapted from [308, 311].

approximately 292 h GPU training time[6], comprehensive hyperparameter tuning would lead to substantially higher resource costs and environmental impact.

To assess how our design choice might impact the algorithm ranking, we performed a small hyperparameter search. As the learning rate η is among the most important hyperparameters in deep learning [240], we focused our analysis on this parameter. In addition to the default learning rate of η_0=0.001, which was used in our main analysis, we trained two additional models per spatial granularity and modality: one with a reduced η_{-1}=0.0001 and another with an increased η_{+1}=0.01. We then identified the optimal learning rate for each algorithm by selecting the one among η_{-1}, η_0, and η_{+1} that yielded the highest average DSC on the validation set. Subsequently, we repeated the ranking analysis on the test data, using the optimal learning rate for each algorithm rather than a fixed default value.

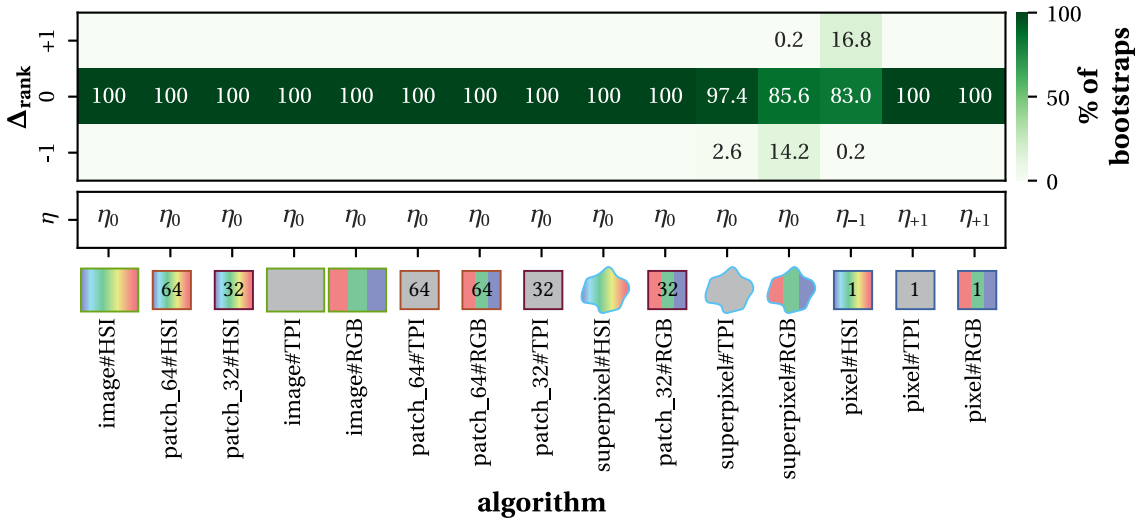


Figure 5.13: Effect of learning rate optimization on the algorithm ranking. Each algorithm was retrained using a reduced and increased learning rate (η_{-1}=0.0001, η_{+1}=0.01) in addition to the default (η_0=0.001). The optimal learning rate η for each algorithm, indicated in the lower box, was determined based on the average Dice similarity coefficient (DSC) on the validation set. Rankings were recalculated across all 1000 bootstrap samples using algorithms trained with the optimized learning rate. The heatmap displays the proportion of samples in which a given rank difference Δ_{rank} occurred compared to the default setting. Algorithms are ordered by their median rank under the default learning rate using the DSC as metric (cf. Figure 5.6). Figure adapted from [308, 311].

As illustrated in Figure 5.13, the optimal learning rate coincided with the default value for most algorithms. Deviations were observed only for the pixel-based models, yielding

[6]This corresponds to approximately 32 kg of CO_2 emissions when trained on an NVIDIA® GeForce RTX™ 2080 Ti [194].

merely marginal improvements in DSC (less than 0.007 across all pixel models). As a result, the overall ranking remained unchanged, with only minor variations across different bootstrap samples. This demonstrates that our study's findings are valid, even without performing extensive hyperparameter tuning for each algorithm.

Postprocessing Our models were designed to facilitate a fair and unbiased comparison, including key design choices such as using the same U-Net architecture and similar epoch sizes across all models. This approach ensured that the primary sources of variation were the input size and modality, rather than differences in model-specific configurations. Similarly, we avoided applying any postprocessing to the network outputs, even though certain models, auch as pixel-based segmentation, could potentially benefit from morphological postprocessing to address the fragmented and scattered boundaries (cf. Figure 5.8). During inference, each model was constrained to its predefined input spatial granularity to allow for an unbiased comparison across spatial granularities. For example, patch models were evaluated on non-overlapping patches to ensure the spatial context did not exceed the specified granularity. However, as illustrated in Figure 5.8, this design choice may lead to artifacts, especially along the patch boundaries in patch-based segmentation.

5.4.2 Strengths, Limitations and Future Work

In the following, we discuss the key strengths and limitations of our study and the methodologies explored, and outline potential directions for future research.

Strengths and Limitations of Hyperspectral Imaging Considering the modest gains in image-based segmentation performance of HSI over RGB, other benefits and limitations of HSI systems should be evaluated when selecting the optimal imaging modality for surgical scene segmentation. Beyond distinguishing tissue classes, the detailed spectral information provided by HSI systems offers additional possibilities in surgical guidance, such as assessing functional tissue characteristics like perfusion state or diagnosing pathological tissues [95, 382]. While, the HSI system used in this study has certain limitations compared to conventional RGB devices, such as longer acquisition times, higher cost, and limited availability, HSI is a rapidly evolving technology, and future iterations are expected to overcome these constraints (cf. Section 8.2 for a detailed discussion).

Upper Bound of Superpixel Performance The superpixel classification approach rests on the assumption that the entire area of a superpixel covers the same class, and therefore superpixel boundaries do not intersect organ boundaries. To evaluate this

assumption and determine an upper performance bound for our superpixel model, we assigned each superpixel the modal label across its constituent pixels. This corresponds to the scenario of all superpixels being correctly classified.

Figure 5.14, presents the results of this experiment, yielding performance limits of 0.92 (SD 0.03) for the DSC, 0.74 (SD 0.04) for the NSD, and 2.91 (SD 0.74) for the ASD. The performance of the superpixel#HSI model is consistently below these limits by a substantial margin, with a DSC of 0.82 (SD 0.06), an NSD of 0.61 (SD 0.09), and an ASD of 16.51 (SD 9.23). While a perfect superpixel-based model achieving these upper bounds would perform slightly better than the best-performing image#HSI model for the DSC and ASD metrics, the maximum achievable NSD still lags behind (0.80 (SD 0.07) for the image#HSI model), suggesting that the superpixel boundaries do not perfectly align with the annotated tissue boundaries. Given these limitations in superpixel clustering, which would require improved algorithms to better match tissue boundaries, as well as the substantial performance gap between the superpixel#HSI model and its upper performance limit, it is questionable whether superpixel-based surgical scene segmentation is a promising direction for future research.

Quality of Reference Annotations Compared to the state of the art in SI-based surgical scene segmentation, our study represents the largest available semantically annotated SI dataset for surgical scene segmentation. To enforce high-quality annotations, we implemented a rigorous two-stage process: initial annotations were provided by two medical experts, followed by a comprehensive review of all annotations by a third expert. Despite these efforts, the quality of our reference annotations remains a potential limitation of our study, as highly accurate annotations are essential for training reliable models and accurately assessing their performance. The inter-rater agreement indicates room for improvement, as reflected by the DSC of 0.89 (SD 0.07), the NSD of 0.80 (SD 0.08), and the ASD of 4.88 (SD 5.33). This aligns with feedback from our medical experts, who noted that determining which pixel belongs to which class is neither straightforward nor entirely unambiguous. Furthermore, the time-intensive nature of the annotation process, requiring approximately 30 min per image, underscores the inherent complexity and challenges of this task.

In addition to deviations in annotated tissue boundaries, we observed instances of tissue class misclassification between the two experts (e.g., labeling spleen as liver or skin as stomach). These misclassifications often occurred when only a small portion of the tissue class was visible, typically due to factors such as occlusions or image boundaries (cf. Figure 5.4). A potential strategy to mitigate such misclassifications could involve leveraging contextual information available intraoperatively, such as the unrestricted field of view, and the ability to gain alternative perspectives or haptic feedback. Although live intraoperative semantic annotation has not been feasible due to its time-intensive nature, future studies could explore the feasibility of live intraoperative

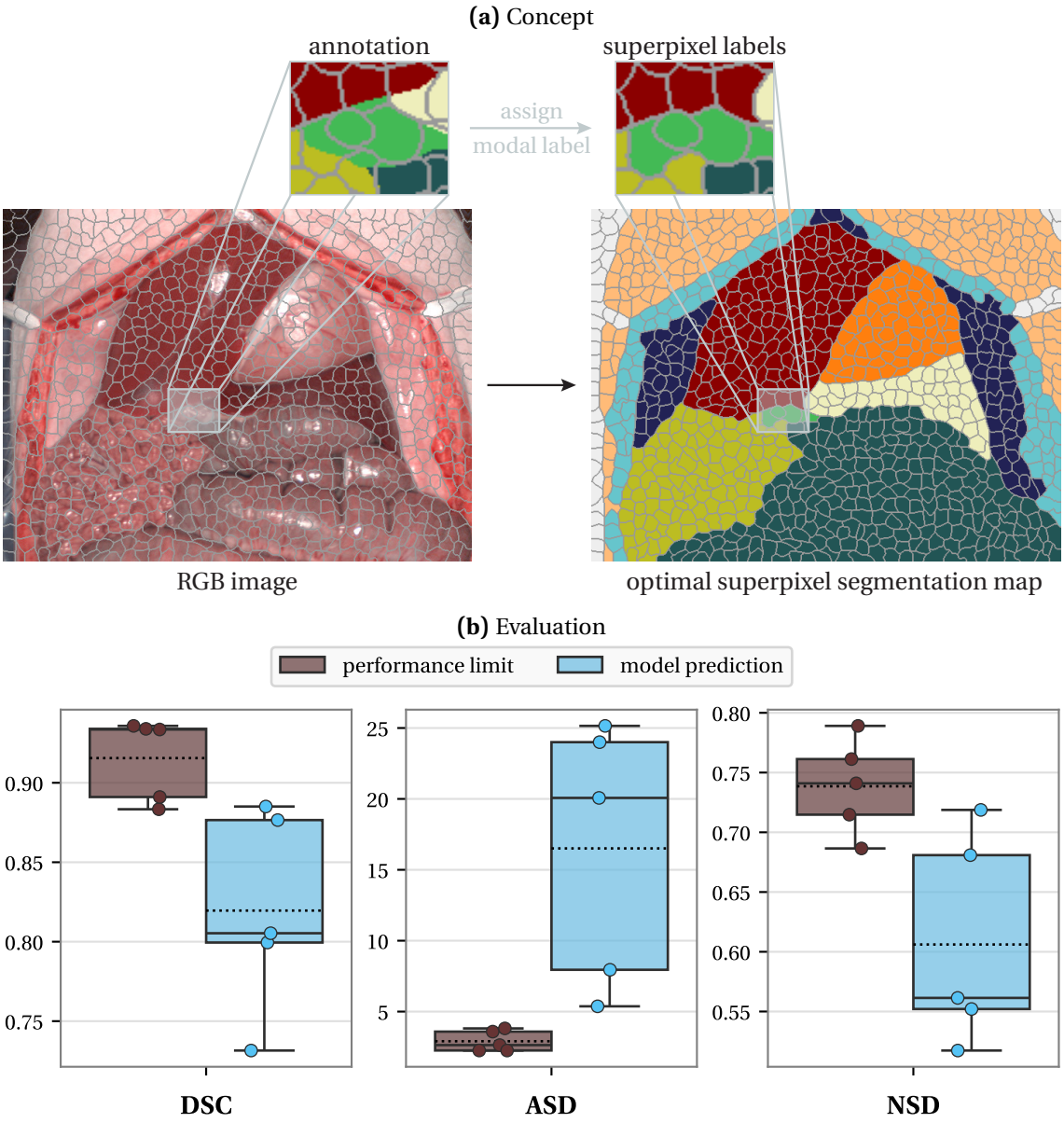


Figure 5.14: Upper bound of the superpixel-based segmentation performance. (a) The upper bound to the superpixel-based segmentation performance is computed by assigning the label for each superpixels based on the mode of the reference annotation labels of the enclosed pixels. (b) The algorithm performance of the superpixel-based segmentation algorithm using hyperspectral imaging data is compared to its upper bound for the metrics Dice similarity coefficient (DSC), average surface distance (ASD) and normalized surface Dice (NSD). Each box displays the interquartile range of the distribution, with whiskers showing the range excluding outliers, and median and mean indicated by a solid and dotted line, respectively. Each marker represents one test subject. Figure adapted from [308, 311].

sparse annotations, particularly for ambiguous tissues, to guide subsequent semantic annotations.

Moreover, the labor-intensive process of creating semantic annotations restricts the overall size of the segmentation dataset. In fact, the presented semantically annotated dataset constitutes only a fraction of our continuously growing intraoperative SI database, which now includes 46 831 images from 388 subjects across 3 species (cf. [315]). Given the impracticality of semantically annotating such a large dataset in its entirety, future efforts could explore active learning strategies to prioritize and select the most informative images for annotation, thereby maximizing the information gained from each annotation.

Spatial Granularities and Geometric Domain Shifts Our findings revealed that segmentation performance consistently improved across all modalities and metrics with larger input spatial granularity. This prompts the question of whether there are practical scenarios in which using input data with smaller spatial context might be beneficial, despite the observed reduction in performance. One potential advantage of using smaller input spatial granularity could be improved generalization to out-of-distribution data in terms of scene geometry. Such geometric variations could arise from scenarios like partially or fully resected organs. As the spatial context of the input data increases, segmentation models may become more sensitive to out-of-distribution scene geometries. Since our original dataset did not encompass radical changes in scene geometry, we investigated this research question in a follow-up study presented in the following Chapter 6.

5.4.3 Conclusion

Leveraging the largest semantically annotated SI dataset to date for surgical scene segmentation, we demonstrated that HSI data offers substantial performance improvements over both RGB data and processed HSI data, with the benefit becoming more pronounced as the spatial granularity decreases. Our findings highlight the critical importance of selecting the optimal spatial granularity for surgical scene segmentation, with larger spatial granularities consistently outperforming smaller ones. Notably, the image-based HSI model performed on par with annotations made by a second medical expert. We conclude that HSI has the potential to emerge as a powerful imaging modality for automated surgical scene understanding, offering numerous benefits over conventional RGB imaging, such as the capability to additionally extract functional tissue information. To support further research, we have publicly released our surgical scene segmentation framework, together with our pretrained models[7] [312].

[7] https://github.com/IMSY-DKFZ/htc

6 Robust Surgical Scene Segmentation Under Geometric Domain Shifts

In the previous chapter (Chapter 5), we demonstrated that DL-based surgical scene segmentation can achieve high performance comparable to human expert annotations when using HSI instead of RGB data, and entire images instead of smaller input spatial granularities. However, the generalization capabilities of surgical scene segmentation algorithms under domain shifts remains largely underexplored, despite the well-known vulnerability of DL models to substantial performance degradation when training and test data distributions differ. In this chapter, we address the critical challenge of geometric domain shifts in surgical scene segmentation: Despite geometric domain shifts frequently occur in real-world surgical scenes due to variations in procedures or situs occlusions, model development and validation are typically conducted on idealized scenes, overlooking these practical challenges. We close this important gap by presenting the first investigation of the generalizability of surgical scene segmentation models under geometric domain shifts, and introducing a novel data augmentation technique specifically designed to mitigate these shifts.

Section 6.1 provides an overview of related work on the generalization of surgical scene segmentation models, and summarizes the state-of-the-art use of data augmentation techniques in this field. Our approach to addressing geometric domain shifts, together with the datasets used, is presented in Section 6.2. This is followed by a description of our experimental setup and results in Section 6.3. The chapter concludes with a discussion of the strengths, limitations, and directions for future research in Section 6.4.

The research presented in this chapter was conducted in 2021 – 2023 and published in the proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2023 [314], as well as in the thesis of Jan Sellner in 2024 [311]. Moreover, an extended version of the MICCAI paper is available on arXiv [309].

6.1 Related Work

To our knowledge, there is no prior work on the generalizability of surgical scene segmentation models under geometric domain shifts. We are only aware of a single related work in the context of surgical instrument segmentation: Kitaguchi et al. demonstrated that surgical instrument segmentation algorithms fail to generalize to unseen surgical procedures, even when the instruments themselves are familiar but appear in unfamiliar contexts [178].

In the broader deep learning community, domain shifts are an intensively studied challenge, and data augmentation has emerged as a simple yet effective strategy for enhancing model generalizability [317, 14]. Traditional augmentation techniques can be grouped into several categories, including geometric (e.g., cropping, resizing, shifting, scaling, rotating, flipping, perspective transforms), photometric (e.g., color jittering), noise (e.g., Random Erasing [388]), kernel (e.g., blurring, sharpening), and image-mixing transformations (e.g., CutMix [379]). Among these, geometric transformations are most widely used by the general semantic scene segmentation community [169]. To determine the state of the art on data augmentation usage in DL-based surgical scene segmentation, we analyzed the related work presented in the previous chapter (cf. Section 5.1), comprising 15 works on SI data and 18 works utilizing RGB data. Our findings are summarized in Table 6.1.

As in the broader field of semantic scene segmentation, geometric transformations are the most commonly used augmentation technique in DL-based surgical scene segmentation, employed in 19 out of 20 studies reporting data augmentation usage. Photometric transformations were applied in 40 % of these studies, and kernel transformations were used in 15 %. Notably, none of the studies utilized topology-altering transformations such as Random Erasing[1] [388], Hide-and-Seek[2] [321], Jigsaw[3] [61], CutMix[4] [379], or CutPas[5] [86] (cf. Figure 6.1 for example images), likely because they were originally developed for classification and object detection rather than segmentation tasks. However, because topology-altering augmentations distort the contextual information available to the network, they hold the potential to enhance model generalizability under geometric domain shifts.

To address the gaps in the literature regarding the impact of geometric domain shifts on surgical scene segmentation performance and potential strategies to mitigate them, we investigate the following research questions:

[1] Blacking out a randomly selected rectangular region within an image.
[2] Dividing an image into a grid of patches, with randomly selected patches blacked out.
[3] Dividing images into grids of patches and swapping randomly selected patches across images.
[4] Copying a random patch from one image to another.
[5] Copying objects onto random background images.

Table 6.1: State of the art regarding usage of data augmentations in deep learning-based surgical scene segmentation. We reviewed publications on surgical scene segmentation utilizing multispectral imaging (MSI), hyperspectral imaging (HSI), or RGB data to evaluate the use of data augmentation techniques during model development. Studies marked with ✓, ✗ and ? indicate that the manuscript reported the use of the specified data augmentation, reported the use of other augmentations, or provided no details on data augmentations, respectively. Table inspired from [311].

publication	year	modality	geometric	photometric	kernel
Akbari et al. [9]	2008	HSI	?	?	?
Ravì et al. [279]	2017	HSI	✓	✗	✗
Fabelo et al. [94]	2018	HSI	?	?	?
Garifullin et al. [110]	2018	MSI	✓	✗	✗
Moccia et al. [239]	2018	MSI	?	?	?
Fabelo et al. [92]	2019	HSI	✓	✗	✗
Trajanovski et al. [340]	2019	HSI	✓	✗	✗
Cervantes-Sanchez et al. [59]	2021	HSI	?	?	?
Collins et al. [70]	2021	HSI	?	?	?
Trajanovski et al. [339]	2021	HSI	✓	✗	✗
Seidlitz et al. [308]	2022	HSI	✓	✗	✗
Garcia Peraza Herrera et al. [109]	2023	HSI	?	?	?
Leon et al. [202]	2023	HSI	?	?	?
Lotfy et al. [214]	2023	HSI	✓	✓	✓
Bannone et al. [30]	2024	HSI	?	?	?
Collins et al. [69]	2015	RGB	?	?	?
Gibson et al. [119]	2017	RGB	?	?	?
Fu et al. [106]	2019	RGB	✓	✓	✗
Kadkhodamohammadi et al. [165]	2019	RGB	?	?	?
Laves et al. [196]	2019	RGB	✓	✗	✗
Allan et al. [12]	2020	RGB	✓	✓	✗
Madad Zadeh et al. [220]	2020	RGB	?	?	?
Maqbool et al. [231]	2020	RGB	✓	✗	✗
Scheikl et al. [302]	2020	RGB	✓	✗	✓
Gong et al. [122]	2021	RGB	✓	✓	✗
Grammatikopoulou et al. [125]	2021	RGB	✓	✓	✗
Jin et al. [162]	2022	RGB	✓	✗	✗
Bhattarai et al. [38]	2023	RGB	?	?	?
Ghamsarian et al. [115]	2023	RGB	✓	✓	✓
Kolbinger et al. [185]	2023	RGB	✓	✓	✗
Luo et al. [216]	2023	RGB	✓	✗	✗
Liu et al. [209]	2024	RGB	✓	✓	✗
Urrea et al. [344]	2024	RGB	✗	✗	✗

RQ2.3: How do geometric domain shifts affect the performance of state-of-the-art RGB and HSI models for surgical scene segmentation?
RQ2.4: How does the spatial granularity of the input data influence the extent of performance degradation?
RQ2.5: Can topology-altering augmentation techniques address geometric domain shifts?

6.2 Materials and Methods

This section describes our approach to addressing geometric domain shifts in surgical scene segmentation (Section 6.2.1), as well as the datasets used to analyze the impact of geometric domain shifts and validate our approach (Section 6.2.2).

6.2.1 Proposed Approach to Address Geometric Domain Shifts

Our approach is driven by the hypothesis that topology-altering data augmentation techniques can help mitigate geometric domain shifts. Therefore, rather than altering the architecture of the segmentation networks, we propose to combine the segmentation models presented in the previous Chapter 5 with a novel data augmentation technique inspired by geometric domain shifts commonly encountered during surgeries.

Surgery-Inspired Data Augmentation Our data augmentation technique, referred to as Organ Transplantation, is illustrated in Figure 6.1. Much like how a donor organ is transferred during transplantation, our Organ Transplantation augmentation involves copying all pixels of a specific object class (e.g., an organ or background) and pasting them into a different surgical scene: From a batch of n images, we randomly select m images (based on the probability parameter p for applying the augmentation) that act as donor images from which the selected classes will be transplanted. For each of these m donor images, a class is chosen at random, and all pixels belonging to that class are transferred to another randomly selected image in the batch (the acceptor). As illustrated in Figure 6.1, the corresponding object segmentation is transferred in tandem.

Our Organ Transplantation augmentation places a class object in an unconventional geometric context while maintaining its original shape and texture, thus (1) generating an occlusion in the acceptor image and (2) forcing the model to detect donor classes independent of their surroundings. The technique can be regarded as an advancement of the image-mixing augmentation CutPas that was initially introduced for object

detection [86]. Since then, it has been adapted for instance segmentation [117] and for creating cost-effective datasets in surgical instrument segmentation by synthesizing images from a limited set of real-world images [353].

Alternative Augmentations Other topology-altering augmentations suggested for classification and object detection could also be promising candidates for addressing geometric domain shifts in surgical scene segmentation. For instance, Random Erasing [388] and Hide-and-Seek [321], which black out pixels within rectangular image regions, could be used to simulate situs occlusions. Jigsaw [61] and CutMix [379] transfer rectangular image regions onto a different scene, thus both occluding parts of the original scene and placing image segments in unusual contexts. To enable a comparison of these augmentation techniques to our Organ Transplantation augmentation, we adapted them for a segmentation task by transferring/invalidating the corresponding segmentation regions together with the image regions.

6.2.2 Datasets

To assess the performance of surgical scene segmentation models under geometric domain shifts and evaluate the improvements introduced by targeted data augmentations, we examined the following geometric OOD scenarios:

(I) *Isolation*: Abdominal linens are frequently used during surgeries to protect organs and soft tissues, absorb secretions and blood, and manage bleeding. In certain procedures, such as enteroenterostomy, it is even required to cover all but one organ [361]. In these scenarios, accurately identifying an isolated organ without contextual cues from neighboring organs is essential.

(II) *Resection*: Resection procedures involve removing parts or even entire organs, making it necessary to identify surrounding organs even when typical neighboring structures are absent.

(III) *Occlusion*: Parts of the surgical scene may be obscured due to the ongoing intervention, introducing OOD elements such as gloved hands. Despite these occlusions, the unobstructed parts of the surgical field need to be accurately identified.

Manipulated Datasets The data previously used to determine the optimal input spatial granularity and modality for surgical scene segmentation (cf. Section 5.2.1), referred to as the dataset original, consists exclusively of idealized scenes and lacks isolated or resected classes. To address this limitation, we created 4 new datasets by modifying images from original. To simulate an isolation scenario, we processed each image I and its

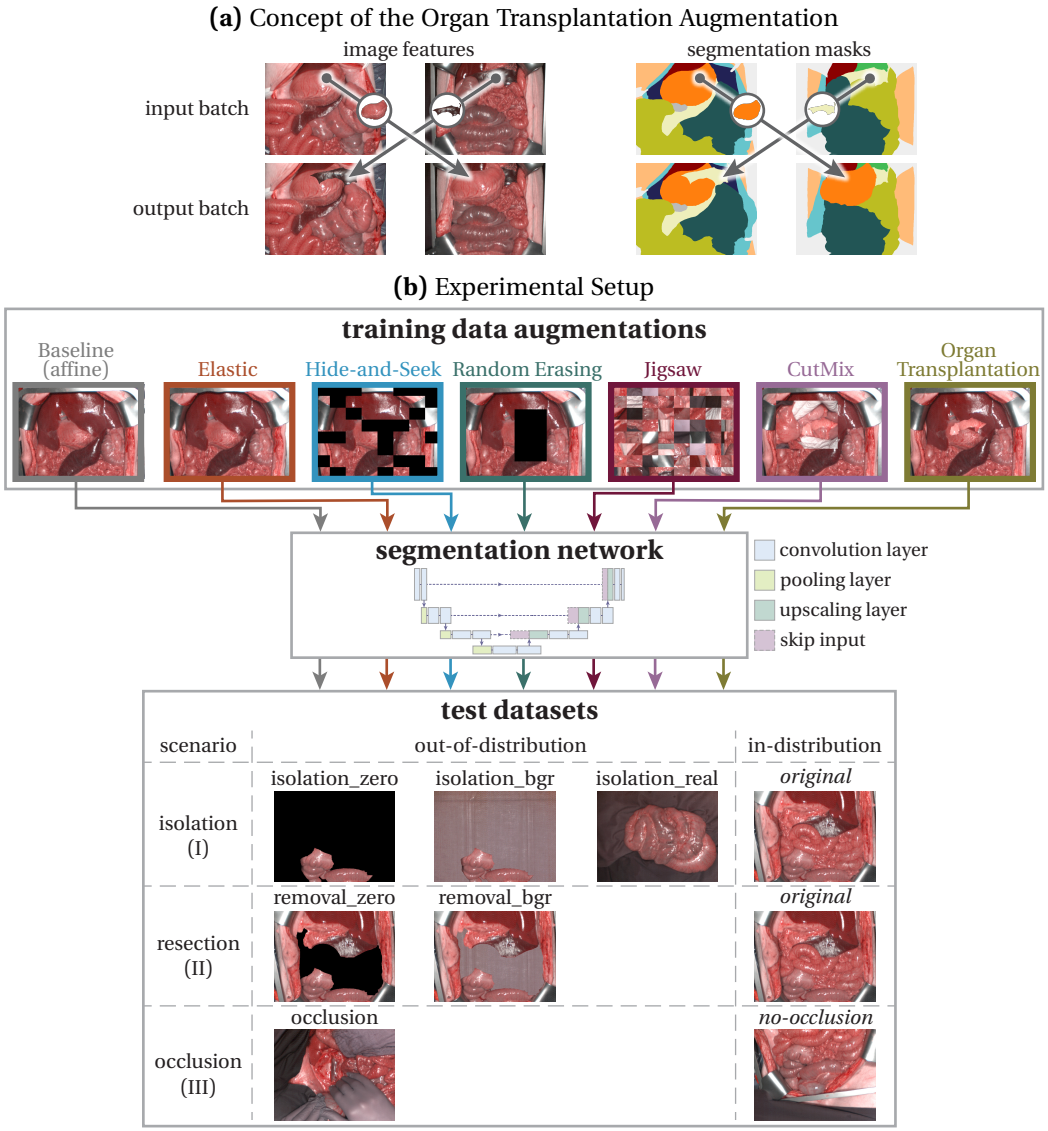


Figure 6.1: Approach and experimental setup to investigate and enhance the generalizability of deep learning-based surgical scene segmentation under geometric domain shifts. (a) We propose to address geometric domain shifts with a surgery-inspired augmentation method, termed Organ Transplantation, which transfers image features and corresponding segmentation masks of randomly selected organs (here: spleen and stomach) between images within the same batch. (b) We evaluate the generalization performance of state-of-the-art surgical scene segmentation models under geometric domain shifts by using either the proposed Organ Transplantation augmentation or one of 6 alternative data augmentation techniques (Affine, Elastic, Hide-and-seek, Random Erasing, Jigsaw, and CutMix augmentations). Our test datasets comprise the 3 geometric out-of-distribution scenarios (I) organs in isolation, (II) organ resections, and (III) situs occlusions, as well as in-distribution data (highlighted in italic). Figure adapted from [314, 309, 311].

corresponding class label c from the original dataset. Pixels in I not belonging to class c were replaced either with zeros, creating the isolation_zero dataset, or with spectra from a background image containing only abdominal linen, forming the isolation_bgr dataset. Similarly, the resection datasets removal_zero and removal_bgr were generated by replacing all pixels in I corresponding to class c with zeros and background spectra, respectively. Example images from these manipulated datasets are shown in Figure 6.1. By generating redundant datasets that differ only in the method of pixel replacement – either with zeros or with background spectra – we can compare scenarios where pixel values are OOD relative to the training data (replacement with zeros) against those where pixel values are in-distribution, originating from a background class seen during training (replacement with abdominal linen spectra).

Real-World Datasets In addition to the manipulated datasets, we collected a real-world isolation dataset, isolation_real, comprising 94 images from 25 pigs. In this dataset, all organs except one were covered with abdominal linen. The same HSI camera, acquisition protocol, and annotation process described in Section 5.2.1 for the original dataset were applied.

To study the impact of occlusions, we divided the original dataset into two subsets: 142 images from 20 pigs containing real-world situs occlusions, such as those caused by gloved hands (dataset occlusion), and 364 images without such occlusions (dataset no-occlusion). Example images from all datasets are provided in Figure 6.1.

6.3 Experiments and Results

The purpose of our experiments was to assess the performance of state-of-the-art DL-based surgical scene segmentation algorithms under geometric domain shifts, considering the input modality (RQ2.3) and spatial granularity (RQ2.4) (Section 6.3.2). Additionally, we sought to evaluate the effectiveness of topology-altering augmentations, particularly our proposed Organ Transplantation augmentation, in mitigating these shifts (RQ2.5, Section 6.3.3). Details of the experimental setup are provided in Section 6.3.1.

6.3.1 Experimental Setup

An overview of our experimental setup is provided in Figure 6.1.

Dataset Splits The same split established in the previous chapter for the dataset original was used, including a training split comprising 340 images from 15 pigs and a hold-out test split containing 166 images from 5 pigs. To ensure a fair comparison across all models and OOD scenarios, the identical train-test split was consistently applied at the pig level. The test splits for isolation_zero, isolation_bgr, removal_zero, and removal_bgr were thus generated by modifying the images in the test split of original. In the occlusion scenario, models were trained on the subset of images in the training split of original that do not contain occlusions, and testing was conducted on both the subset of the test split of original without occlusions (test dataset no-occlusion) and with occlusions (test dataset occlusion).

Model and Training Parameters To study the effect of data augmentations in isolation, we used the same model architectures and training parameters for a given spatial granularity and modality as presented in the previous chapter (cf. Section 5.2.2), with the only divergence between models consisting in the applied data augmentations during model training. The baseline models used geometric data augmentations commonly applied in the state of the art, namely shift, scale and rotate, each applied with a probability of $p = 0.5$. Our competitor models employed one of the augmentations Elastic transformations, Hide-and-Seek, Random Erasing, Jigsaw, CutMix and Organ Transplantation, in addition to the geometric augmentations. Elastic transformations, which apply a random displacement to each pixel, thereby generating tissue deformations that distort the local neighborhood [319], were applied with a displacement magnitude of $\alpha = 0.7$ and a displacement smoothness of $\sigma = 16$. To minimize the need for extensive hyperparameter tuning, instead of determining the optimal grid size in the Hide-and-Seek and Jigsaw augmentations, we randomly sampled grid sizes from a set including grids of 5×5, 8×8, 10×10, 16×16 and 20×20 patches. These sizes were chosen because the image dimensions are divisible by these numbers. Furthermore, the ratio r of grid patches to be blacked out in the Hide-and-Seek augmentation was randomly sampled from the range $r \in [0.2; 0.8]$. Only the probability p of applying one of the 6 competing augmentations was optimized through a grid search with values $p \in \{0.2, 0.4, 0.6, 0.8, 1\}$ – for our image-based models with a batch size of 5, these values correspond to applying the data augmentation on 1, 2, 3, 4 or all 5 images. The optimal p-value was identified using 5-fold cross-validation on the training splits of the original, isolation_zero, and isolation_bgr datasets. This approach ensures that the selected value of p is optimal for both in-distribution and OOD data while preserving the OOD scenarios resection and occlusion, as well as all real-world datasets, as untouched test sets. The resulting optimal p-values are listed in Table 6.2.

Validation Strategy Following the recommendations from [284, 222, 283], we assessed the segmentation performance of our models using both the overlap-based metric DSC

Table 6.2: Optimal setting of the probability hyperparameter according to our grid search. For each data augmentation method, the probability p of applying the augmentation was optimized in a grid search over the values $p \in \{0.2, 0.4, 0.6, 0.8, 1\}$. Table adapted from [311].

data augmentation	optimal probability p
Elastic	0.6
Hide-and-Seek	1
Random Erasing	0.4
Jigsaw	0.8
CutMix	1
Organ Transplantation	0.8

and the boundary-based metric NSD. For the NSD, we used the same class-specific thresholds as derived in the previous chapter (cf. Figure 5.12).

Our datasets maintain a hierarchical structure, with each subject comprising one or more images, and each image containing multiple classes. In the previous chapter, we have accounted for this hierarchy by first aggregating metric values at the image level, yielding image-wise scores, and subsequently aggregating the scores from all images of one subject, resulting in subject-wise scores. Although this approach allowed for analyzing the performance variability across subjects, it introduced the limitation that image-level scores are influenced by the class distribution within an image. Based on the hypothesis that specific classes may be more susceptible to geometric domain shifts than others, we shifted our focus to class-wise scores in this analysis. To achieve this, we calculated the DSC and NSD for each class in every image, and then aggregated the class-wise scores first across all images of one subject, and subsequently across subjects.

In the organ removal scenario, for each class label c in an image I, a set of metric scores $\{M_l(\hat{c})\}$ was obtained by removing each class $\hat{c}$ in I one at a time. To assess the impact of removing the most important neighboring class of c, we selected the minimum score from $\{M_c(\hat{c})\}$ before proceeding with hierarchical aggregation.

We computed performance rankings and assessed their stability with respect to sampling variability in accordance with the guidelines provided by [364]: We generated 1000 bootstrap samples, each comprising 19 class-level scores. For each class label c, the class-level score was derived by randomly selecting N_c subject-level scores belonging to c without replacement, with N_c representing the total number of subjects with images available for class c. These N_c scores were then averaged to obtain the class-level score.

6.3.2 Impact of Geometric Domain Shifts on State-Of-The-Art Surgical Scene Segmentation Models

Our experiments aimed to assess the effect of geometric domain shifts on DL-based surgical scene segmentation as a function of (1) the imaging modalities RGB and HSI, and (2) the choice of input spatial granularity, including pixels, superpixels, patches and entire images.

Performance Degradation as a Function of Modality and Spatial Granularity Figure 6.2 illustrates the segmentation performance, as measured by the DSC, across all studied modalities (HSI and RGB), spatial granularities (pixel, superpixel, patch_32, patch_64 and image), and clinical scenarios involving geometric domain shifts (organs in isolation, resections and situs occlusions). While substantial performance drops between in-distribution and OOD data are observed for both RGB and HSI data, despite the latter's rich spectral information content, the average drop in performance is smaller for HSI, with a decrease of 23 %, compared to a drop of 30 % for RGB.

Consistent with our previous findings, the in-distribution performance is highest for image-based segmentation in both modalities (RGB: DSC of 0.83 (SD 0.10); HSI: DSC of 0.86 (SD 0.10)) and decreases with reduced spatial granularity of the input.

While pixel-based segmentation models exhibit the lowest overall performance, they show no decline when applied to OOD data in the isolation and removal scenarios. In fact, in the isolation scenario, pixel-based models even show improved performance for both manipulated and real-world data. This improvement can likely be attributed to the tendency of pixel-based segmentation models to produce fragmented and scattered boundaries for tissue classes, while background pixels are generally identified with high accuracy (cf. Figure 5.8). In the manipulated isolation data, pixels outside the target organ annotation were replaced with zeros and background pixels. Similarly, in the isolation_real dataset, the entire scene, except for the target organ, was obscured with abdominal linen. These strategies effectively eliminate mispredictions of the target class beyond the annotated region, a challenge often encountered in multi-organ images.

As the spatial granularity increases, the drop in segmentation performance for organs in isolation and removal scenarios becomes more pronounced across both modalities. For image-based segmentation, this performance drop is largest, ranging from 10–46 % for RGB to 5–45 % for HSI, depending on the specific OOD scenario. Although models using smaller input spatial granularities exhibit less performance degradation under geometric domain shifts, none achieve an OOD performance comparable to the in-distribution performance of image-based models. Thus, relying on smaller spatial granularities is not a viable strategy for improving generalizability under geometric domain shifts.

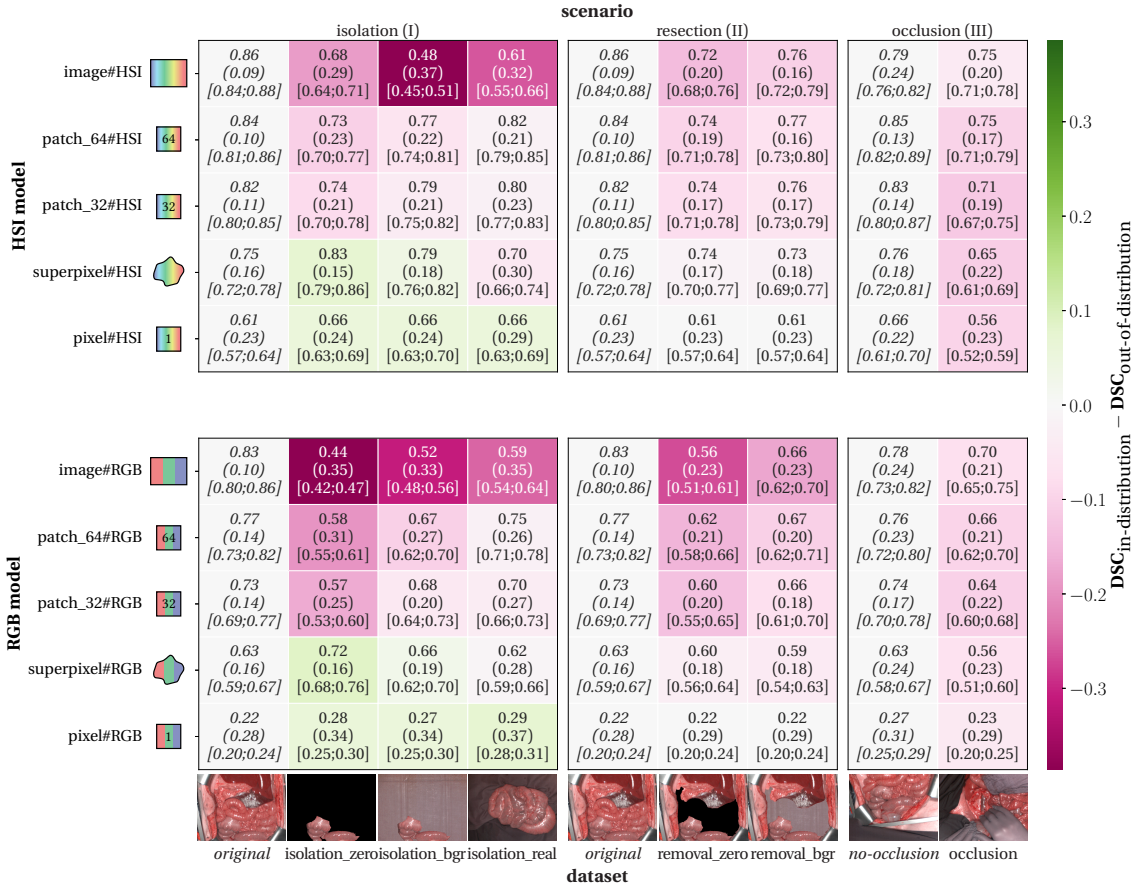


Figure 6.2: Role of the input modality and spatial granularity in segmentation performance degradation under geometric domain shifts as measured by the Dice similarity coefficient (DSC). The segmentation performance is reported for 3 clinical scenarios: organs in isolation (I), organ resections (II), and situs occlusions (III). Columns represent the corresponding in-distribution datasets (highlighted in italic) and out-of-distribution (OOD) datasets. Rows indicate different models, each combining one of two modalities (RGB or hyperspectral imaging (HSI)) with one of 5 spatial granularities: pixel, superpixel, patches of size 32×32 (patch_32) or 64×64 (patch_64), and image. The numbers represent the average DSC across classes, with standard deviations denoted in brackets. The color-coding reflects the difference in DSC relative to the corresponding in-distribution DSC for the same model. Results for the normalized surface Dice (NSD) are shown in Figure B.14. Figure adapted from [309].

Impact of Neighborhood on Segmentation Performance We analyzed which classes are most affected by the removal of neighboring classes in the organ removal scenario. As shown in Figure 6.3, the drop in performance using the image#HSI model is highest with 63 % for the gallbladder upon removal of the liver, and second highest for the major vein after removing the peritoneum. In both cases, the removed organ is a prominent neighbor of the organ under investigation: the liver constitutes 83.9 % of the gallbladder's neighborhood, and the peritoneum makes up 60.1 % of the major vein's neighborhood, on average. Other classes, such as liver, colon, and muscle, do not exhibit a decline in performance upon removal of neighboring classes. These findings indicate that the impact of neighboring classes on segmentation performance is class-specific.

6.3.3 Effectiveness of Data Augmentations

Our experiments aimed to evaluate the effectiveness of targeted data augmentations, particularly our proposed Organ Transplantation augmentation, in mitigating geometric domain shifts in surgical scene segmentation.

Comparison of Our Organ Transplantation Augmentation to the State of the Art As shown in Figure 6.4 for the DSC, and in Figure B.15 for the NSD, equipping the image-based segmentation models with our organ transplantation augmentation effectively mitigates geometric domain shifts for both the HSI and RGB modalities. Notably, small performance improvements are observed even on in-distribution data. For HSI, the performance improvement over the baseline ranges from 9–90 % (DSC) and 16–96 % (NSD), whereas the performance improvement using RGB falls within a lower range of 9–67 % (DSC) and 15–79 % (NSD), underscoring the importance of spectral information in scenarios with limited context.

The largest performance improvement is observed in the isolation scenario, which also exhibits the largest performance drop in the baseline model. In contrast, situs occlusions show the smallest baseline performance drop and the least improvement with the Organ Transplantation augmentation. However, there is a noticeable variation in performance improvement across classes. For instance, on the occlusion dataset, the largest DSC improvement for HSI is obtained for the pancreas (283 %), followed by the stomach (69 %), suggesting that certain classes particularly benefit from the Organ Transplantation augmentation.

The performance improvements observed on manipulated datasets (average DSC improvement of 57 % for HSI and 61 % for RGB across the datasets isolation_zero and isolation_bgr) align with those observed on real data (DSC improvement of 50 % for HSI and 46 % for RGB on the dataset isolation_real). This consistency highlights the

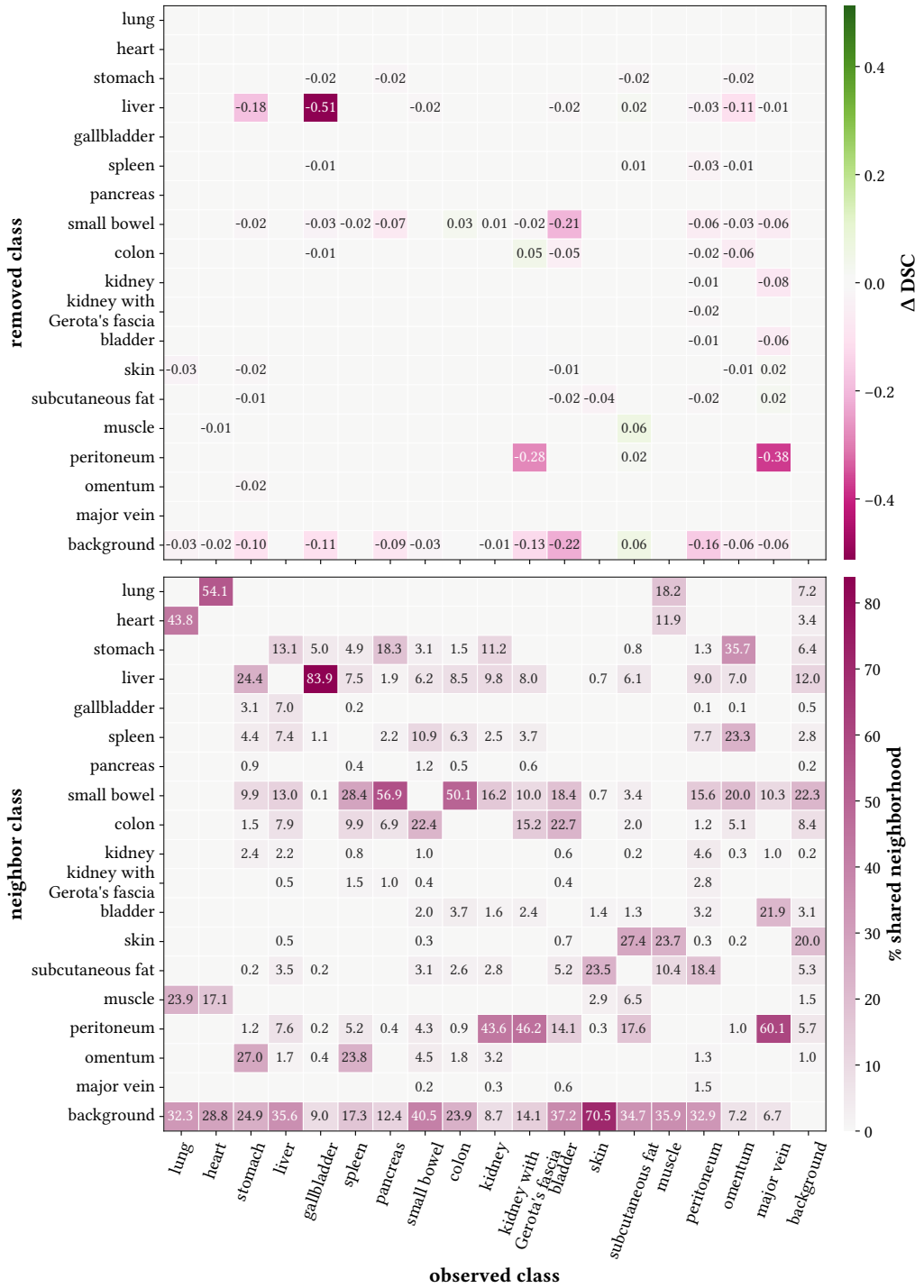


Figure 6.3: Impact of local neighborhood on performance drop following organ removal. The top confusion matrix shows the change in average Dice similarity coefficient (DSC) for class c (columns) when class c' (rows) is replaced with zeros, using the image#HSI model. Changes $|\Delta \text{ DSC} < 0.01|$ were omitted for clarity. The bottom confusion matrix presents the average proportion of boundary pixels in the dataset original that class c (columns) shares with class c' (rows), with values below 0.1 % omitted for readability. Figure adapted from [314, 309, 311].

effectiveness of our manipulated datasets in accurately evaluating segmentation performance under geometric domain shifts.

Visual Assessment of Segmentation Quality Figure 6.5 shows example predictions from the image#HSI baseline model, and the corresponding Organ Transplantation-augmented model, for each of the 6 OOD datasets. The examples were chosen based on the largest difference in DSC performance between the baseline model and the Organ Transplantation model, illustrating cases where the augmentation provides the largest benefit. In all 6 examples, the baseline model's performance drop is mainly due to misclassifying entire organ classes, such as failing to identify the gallbladder and stomach after liver removal or missing the stomach and large portions of the omentum obscured by a gloved hand in the occlusion scenario.

Comparison to Alternative Data Augmentation Techniques Figure 6.6 shows the DSC-based ranking of our Organ Transplantation augmentation compared to the baseline geometric augmentations, Elastic transformations, and 4 other topology-altering augmentations on the 6 geometric OOD test datasets, while the NSD-based ranking is presented in Figure B.16. The Organ Transplantation augmentation consistently ranks first, whereas the baseline augmentations rank last across most OOD scenarios. Although rankings for the other augmentations vary across geometric OOD datasets, the overall ranking reveals that most topology-altering augmentations outperform the Elastic transformations.

Among the topology-altering methods, image-mixing augmentations, such as CutMix and Jigsaw, demonstrate superior performance compared to noise-based augmentations like Random Erasing and Hide-and-Seek. This may be due to image-mixing augmentations introducing unusual neighborhood relationships by copying patches from one surgical scene into another, whereas noise augmentations merely obscure parts of the scene without modifying the existing neighborhood relationships. Additionally, it is notable that Random Erasing and Hide-and-Seek rank better on the datasets isolation_zero and removal_zero compared to the corresponding datasets where tissues were replaced with background, isolation_bgr and removal_bgr. This observation suggests that these augmentations are more effective when the same type of obscuration (i.e., zero values) is present in both the augmented training data and validation data, while their ability to generalize to other types of obscurations, such as those involving background, appears to be limited.

In addition to our Organ Transplantation augmentation consistently ranking first, it is noteworthy that topology-altering augmentations that randomly select patches generally outperform those based on a grid structure (e.g., CutMix vs. Jigsaw, Random Erasing vs. Hide-and-Seek). This may be attributed to the extent of unnatural boundaries introduced by each method: Our Organ Transplantation augmentation preserves

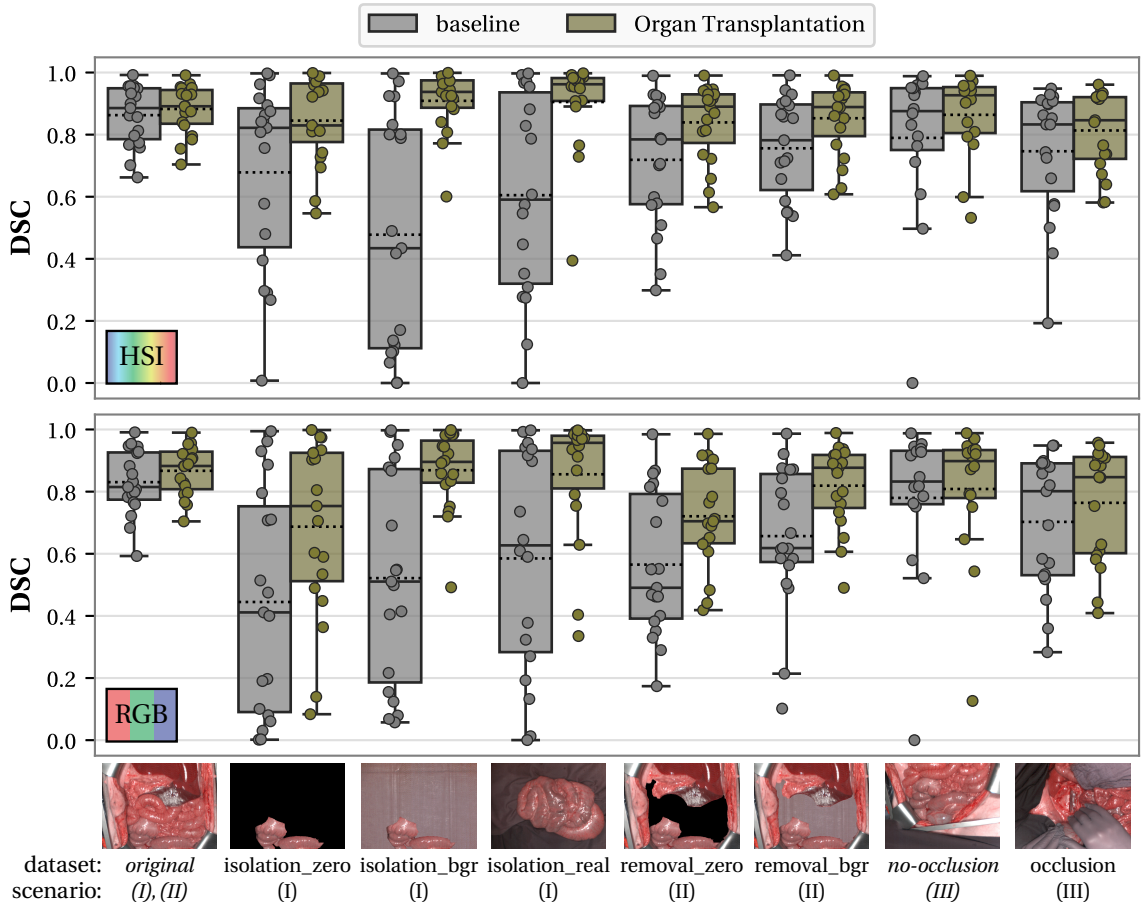


Figure 6.4: Performance comparison of the baseline model and the Organ Transplantation model under geometric domain shifts using the Dice similarity coefficient (DSC). Distributions of class-wise DSC scores are shown for the baseline model and the Organ Transplantation model across the 3 clinical scenarios (I) organs in isolation, (II) organ resections, and (III) situs occlusions, with in-distribution datasets highlighted in italic. The boxplots illustrate the quartiles of the distribution across classes, with whiskers showing the range excluding outliers. The median is shown as a solid line, the mean as a dotted line, and the markers correspond to individual classes. Results for the normalized surface Dice (NSD) are shown in Figure B.15. Figure adapted from [314, 309, 311].

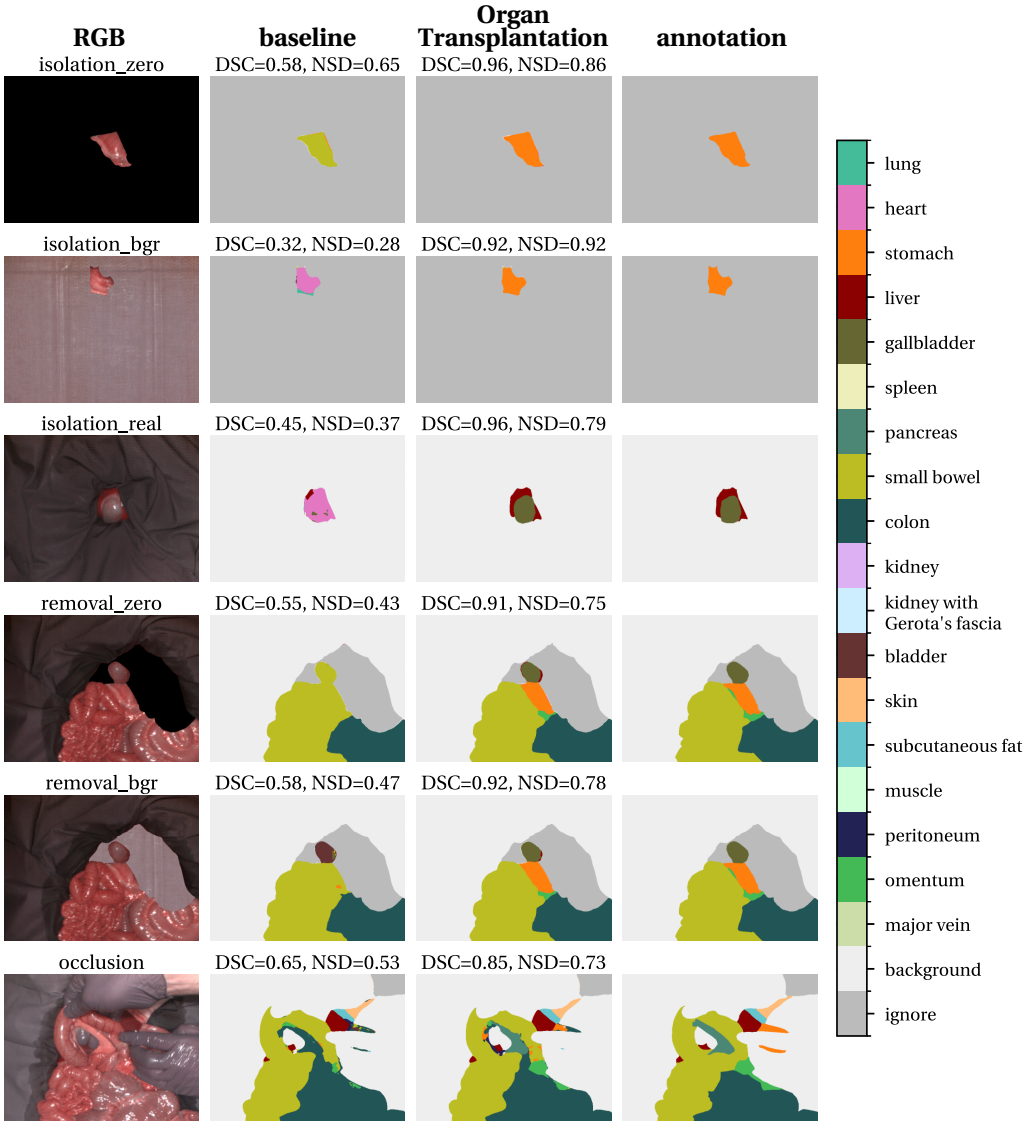


Figure 6.5: Example predictions from the image#HSI baseline model and corresponding Organ Transplantation model on geometric out-of-distribution (OOD) datasets. For each of the 6 OOD datasets (rows), an image was selected to maximize the difference in Dice similarity coefficient (DSC) values between the baseline and Organ Transplantation models. From left to right, the corresponding RGB image, segmentation predictions from the baseline and Organ Transplantation models, and the reference annotation are displayed, along with the image-wise DSC and normalized surface Dice (NSD) scores for the segmentation predictions. Figure adapted from [314, 309, 311].

natural organ boundaries, while CutMix and Random Erasing create artificial boundaries along the edges of a single rectangle. In contrast, Jigsaw and Hide-and-Seek generate even more unnatural boundaries due to their grid-based modifications.

6.4 Discussion and Conclusion

In this study, we demonstrated for the first time that state-of-the-art surgical scene segmentation networks experience substantial performance degradation under geometric domain shifts. Through an extensive validation on 6 geometric OOD datasets, consisting of 600 RGB and HSI cubes from 33 pigs, each annotated with 19 classes, we observed that performance degradation was generally more severe for RGB data compared to HSI. Furthermore, the decline was more pronounced with larger spatial granularities, such as images and patches, compared to smaller spatial granularities like pixels and superpixels. To improve the generalization of state-of-the-art models to OOD geometries, we adapted previously unexplored topology-altering data augmentation methods for surgical scene segmentation. Among these, our proposed Organ Transplantation augmentation outperformed all other topology-altering methods and achieved performance comparable to in-distribution results.

The following sections provide a discussion of key strengths and limitations (Section 6.4.1) and potential future research directions (Section 6.4.2), as well as a conclusion summarizing our findings (Section 6.4.3).

6.4.1 Strengths and Limitations

In the following, we discuss the key strengths and limitations of our manipulated data and our proposed Organ Transplantation augmentation.

Strengths and Limitations of Our Manipulated Data While we validated model performances on real-world data for the isolation and occlusion scenarios, obtaining real-world data was impractical in the resection scenario. In fact, due to the invasive nature of organ resections, covering a wide range of resection scenarios would have required more animals. To minimize animal suffering and reduce research costs, we instead relied on manipulating existing data as a viable alternative for validating our model performance in the resection scenario. We further reduced the amount of real-world data needed by tuning model hyperparameters on the validation splits of the in-distribution dataset original and the manipulated OOD datasets isolation_zero and isolation_bgr, keeping all real-world OOD datasets as untouched test sets. Despite

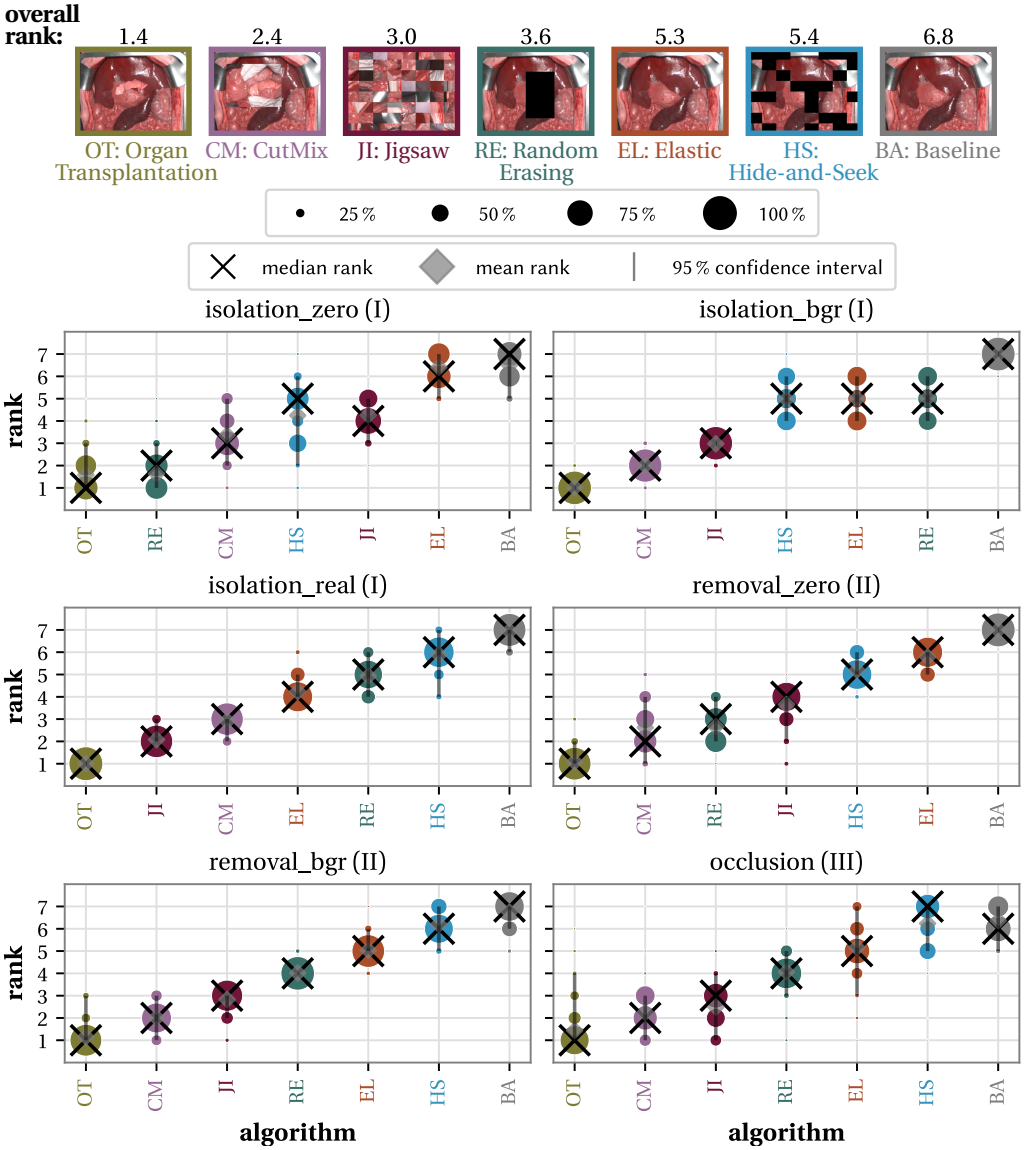


Figure 6.6: Uncertainty-aware Dice similarity coefficient (DSC)-based ranking of different data augmentation methods for addressing geometric domain shifts. Following the concept from [364], bootstrap sampling was performed to assess the ranking stability with respect to sampling variability of our image#HSI models utilizing the data augmentation techniques Organ Transplantation (OT), CutMix (CM), Jigsaw (JI), Random Erasing (RE), Elastic transformations (EL), Hide-and-Seek (HS) and Baseline geometric transformations (BA). For each blob at position (a, r), its area is proportional to the frequency of algorithm a achieving rank r across 1000 bootstrap samples. For each method, black crosses indicate the median rank, gray diamonds show the mean rank, and gray lines represent the 95 % quantile of the bootstrap results. Ranking stability results for the normalized surface Dice (NSD) are shown in Figure B.15. Figure adapted from [314, 309, 311].

these restrictions, (1) our proposed Organ Transplantation augmentation proved effective across all datasets, and (2) we observed similar performance improvements for image-based segmentation on manipulated and real data in the isolation scenario, highlighting the utility of our image manipulations as an effective substitute for real-world data in this context.

However, limitations arise from the oversimplification of our manipulations: First, the removal_zero and removal_bgr datasets do not account for the emergence of other tissues that may be visible in place of the removed organ. Second, our manipulated datasets led to artifacts in the comparison of spatial granularities by biasing the results in favor of the superpixel-based spatial granularity: For instance, as shown in Figure 6.2, superpixel-based models exhibit performance improvements in the manipulated isolation scenarios, which are not observed in the real-world isolation data. This discrepancy can likely be attributed to our manipulation strategy: As shown in Section 5.4.2, superpixel boundaries in real data often do not align precisely with annotation boundaries. In contrast, for the manipulated data, we utilized the reference boundary annotations of the target organ to replace non-target pixels with zeros or background spectra. This approach produced superpixel boundaries that closely matched the annotations, resulting in improved segmentation scores. Despite their limited capability to represent the complexity of real-world geometric OOD scenes, our manipulated datasets provide a valuable tool for evaluating model performance under geometric domain shifts, as they allow for a controlled comparison of the impact of specific geometric changes on segmentation performance.

Strengths of Limitations of Our Organ Transplantation Augmentation Key strengths of our Organ Transplantation augmentation include:

- **Flexibility:** Our Organ Transplantation augmentation is compatible with any existing model architecture, providing the flexibility to select models based on specific requirements while still benefiting from topology-altering augmentations, all without the need for specialized architectures.
- **Effectiveness:** Our augmentation method effectively mitigates geometric domain shifts, as demonstrated by the consistent performance improvements across all OOD scenarios, yielding performance comparable to in-distribution results. Furthermore, the augmentation consistently outperformed competing data augmentation methods, achieving the highest rank across all OOD scenarios.
- **Efficiency:** Our augmentation is computationally efficient for image-based segmentation models, as it can be performed on the GPU [313]. An exception are smaller spatial granularities: Here, the augmentation would introduce substan-

tial computational overhead and increased memory requirements[6], rendering batch-level augmentations impractical in these cases. However, as image-based segmentation models outperformed models using smaller spatial granularities on in-distribution data (cf. Figure 6.2), and our Organ Transplantation augmentation combined with image-based segmentation achieved geometric OOD performance comparable to the baseline in-distribution performance (cf. Figure 6.4), this is not a strong limitation in practice.

A general limitation of image-mixing augmentations, such as our Organ Transplantation method, is the requirement for a minimum batch size of two images to transplant an organ from one scene to another. In our case, this limitation was not problematic due to our batch size of 5 images. However, it may become a concern for applications with limited computational resources or large images.

6.4.2 Future Work

In this study, we investigated geometric OOD scenarios commonly encountered in real-world open surgeries. Similar challenges, such as instrument occlusions and organ removals, also arise in minimally invasive surgeries. However, key differences include more focused views of organs, fewer neighboring organs visible in the image, tissue deformations, and substantial changes in imaging perspectives. With the recent availability of medical device-graded HSI systems for minimally invasive surgery, coupled with the steady increase in such procedures over the past decades [323], exploring and mitigating geometric domain shifts in this context represents a promising avenue for future research.

Furthermore, this work provides a first step towards the broader challenge of investigating and addressing domain shifts in DL-based surgical scene segmentation using RGB and HSI data. Other potentially relevant domain shifts between our training data and real-world human surgeries include for example shifts in illumination and measurement devices (e.g., different spectral channels across MSI and HSI devices, shifts in field of view and image resolution), pathological conditions of tissues instead of physiological data (e.g., presence of tumors, inflammation, malperfusion) and the emergence of artifacts introduced by surgical procedures (e.g., injection of fluorescent dyes, bleeding, cauterization). Since the publication of this study, we have demonstrated and addressed drops in surgical scene segmentation performance under illuminant shifts [34], tissue malperfusion [273, 315] and injection of indocyanine green [315]. Despite

[6] Due to memory and efficiency constraints, extracting smaller input granularities from images must be performed on the central processing unit (CPU). As a result, augmentations would also need to be performed on the CPU, even though this is less efficient compared to GPU-based augmentation [313]. Additionally, the requirement to load at least two images simultaneously would substantially increase both memory usage and computational load when working with lower spatial granularities.

this progress, investigating the impact of additional, so far unexplored domain shifts on the segmentation performance and developing strategies to mitigate them are crucial steps towards enhancing the generalizability of surgical scene segmentation models for real-world human surgeries.

6.4.3 Conclusion

To the best of our knowledge, this work is the first to address surgical scene segmentation under geometric domain shifts. We have demonstrated that state-of-the-art segmentation models experience substantial performance degradation under geometric domain shifts and showed that in-distribution performance can be restored using our Organ Transplantation augmentation. Our method is computationally efficient, effective, and model-independent, making it applicable to image-based surgical scene segmentation for both HSI and RGB data, and across various model architectures. To support further research, we have publicly released our code repository and pretrained models on GitHub[7] [312] and have integrated our Organ Transplantation augmentation into the Kornia library[8] [287], enabling easy access for the broader computer vision community.

[7] https://github.com/IMSY-DKFZ/htc
[8] Kornia `RandomTransplantation` augmentation

Part IV

Robust Sepsis Diagnosis and Mortality Prediction with Hyperspectral Imaging (RQ3)

7 AI-Driven Skin Spectral Imaging for Rapid Sepsis Diagnosis and Mortality Prediction in Critically Ill Patients

As outlined in Section 1.2.3, the early identification of septic patients and individuals at high risk of mortality is of major socioeconomic relevance, as every hour of delayed intervention increases mortality. This chapter presents the first analysis of the potential of DL-based HSI analysis to close this diagnostic gap by enabling rapid, non-invasive sepsis diagnosis and mortality prediction in ICU patients.

Section 7.1 provides an overview of the related work on sepsis diagnosis and mortality prediction, highlighting the limitations of existing approaches. Our DL approach to address the diagnostic and prognostic gaps in current clinical practice, together with our large-scale HSI study conducted in an interdisciplinary surgical ICU, is presented in Section 7.2. Our experimental setup and findings are detailed in Section 7.3, followed by a discussion of the strengths, limitations, and directions for future research in Section 7.4.

The research presented in this chapter was published in the journal Science Advances in 2025 [306], building on earlier work first reported on arXiv in 2021 [85].

7.1 Related Work

Previous research has primarily focused on sepsis diagnosis and mortality prediction through two key avenues: the discovery of diagnostic and prognostic biomarkers, and ML-based prediction using clinical data from electronic health records (EHRs).

Diagnostic and Prognostic Biomarkers Over the past decades, extensive research has explored the identification of biomarkers for sepsis diagnosis and mortality prediction, proposing over 250 molecules as potential candidates. While the investigation of

biomarkers has contributed to the identification of sepsis endotypes and pathways [350], up to date, none of the proposed biomarkers have shown sufficient sensitivity and specificity for reliably detecting sepsis or predicting clinical outcomes [265, 19]. This challenge is likely due to the heterogeneity and complexity of the sepsis pathophysiology, compounded by its non-specific signs and symptoms [136]. For example, C-reactive protein (CRP), an inflammatory marker extensively studied for sepsis diagnosis, is also elevated in sterile inflammations following autoimmune responses or surgical trauma, leading to a high incidence of false positives [362].

Sepsis and Mortality Prediction from Clinical Data More recently, researchers have turned to ML for predicting sepsis and mortality using high-dimensional clinical data from EHRs [186]. Among these approaches, random forest models have emerged as the most frequently employed [375]. While the number and selection of clinical features differ across studies – ranging from as few as two to over 100 – the most commonly employed data for ML models include vital signs, laboratory values, and patient demographics [154]. A 2020 meta-analysis of 28 publications, covering 130 models, reported area under the receiver operating characteristic curve (AUROC) performances between 0.68 and 0.99 for sepsis prediction up to 48 h prior to its onset [104, 154]. Despite these encouraging results, translating EHR-based sepsis and mortality prediction models into clinical practice remains challenging. EHR data, originally designed for clinical documentation and billing, suffers from a lack of standardization, incompleteness, inaccuracies, and inherent biases (e.g., correlations between measurement frequency and disease severity, sample selection biases) [301]. These limitations affect the generalizability of models to external data, as shown in multiple studies on EHR-based sepsis prediction [370, 241]. Moreover, while EHR systems are prevalent in high-income countries, their adoption is much slower in low- and middle-income countries (LMICs), where 85 % of sepsis cases occur [296]. This gap is attributed to inadequate laboratory facilities, lack of infrastructure, training, and implementation frameworks, limited access to expensive monitoring equipment, and insufficient technical support [369].

Leveraging Microcirculatory Dysfunction and Edema Formation for Sepsis Diagnosis Section 2.2.2.2 provides a detailed overview of sepsis pathophysiology. A key pathophysiological process is endothelial and coagulation dysfunction, leading to edema formation and microcirculatory dysfunction, characterized by impaired blood flow in the smallest vessels. Microcirculatory dysfunction emerges early in sepsis [275], plays a critical role in organ failure, and is closely linked to poor outcomes [342, 76]. Because microcirculation is frequently decoupled from systemic hemodynamics, traditional systemic parameters such as blood pressure and cardiac output fail to capture its impairment [275, 355]. Instead, advanced imaging techniques, including NIR spectroscopy, laser speckle contrast imaging, laser Doppler flowmetry, sublingual microscopy, and HSI

could be explored for microcirculatory monitoring. These methods already revealed that sepsis-induced microcirculatory dysfunction is characterized by reduced capillary density and increased microperfusion heterogeneity, resulting in localized hypoxic regions [352, 257].

We thus hypothesize that HSI can support automated sepsis diagnosis and mortality prediction in the ICU by capturing edema formation and microcirculatory dysfunction. HSI offers several key strengths, including its mobility, non-invasiveness, rapidity, objectivity, cost-effectiveness, and standardization. Unlike other imaging techniques that could monitor microcirculation, medical-grade HSI systems are now emerging, setting the stage for HSI to become a standard tool in clinical practice [80, 355].

Previous studies have shown distinct patterns in functional tissue parameters derived from HSI when comparing septic patients to controls [172, 193, 80] and reported promising performance for HSI-based sepsis diagnosis [85, 182]. However, these studies suffer from a major limitation: Septic patients were compared to healthy volunteers or narrowly defined cohorts, such as pancreatic surgery patients. This study design introduces a high risk of shortcut learning, driven by confounders like large age differences, comorbidities, or therapy regimens between groups [85]. As a result, the proposed algorithms are unlikely to generalize to real-world clinical settings, such as automated sepsis diagnosis in ICU patients, where diagnosis is particularly challenging due to disease complexity, high baseline illness severity, and the difficulty of distinguishing sepsis from non-infectious systemic inflammation [42, 212].

Overall, despite substantial research efforts, reliable biomarkers for early sepsis diagnosis and mortality prediction remain elusive. While existing studies suggest that HSI holds promise for automated sepsis diagnosis, their findings are unlikely to generalize to realistic clinical settings. We address this critical gap by presenting the first DL-based HSI analysis for automated, rapid, and non-invasive sepsis diagnosis and mortality prediction in ICU patients. Using data from a prospective study of more than 480 patients – representing, to our knowledge, the largest HSI patient cohort to date – we investigate the following research questions:

RQ3.1 Can DL-based skin HSI analysis enable automated, rapid, and noninvasive sepsis diagnosis and mortality prediction in ICU patients? Which measurement site, imaging modality (HSI vs. TPI vs. RGB images) and spatial granularity (patches vs. median spectra) yields the best performance?

RQ3.2 Do algorithms trained on existing HSI data of selectively chosen cohorts generalize to an ICU population?

RQ3.3 Can structured clinical data further improve diagnostic and predictive performance?

RQ3.4 How does the performance of our approach compare to established clinical biomarkers and scores?

7.2 Materials and Methods

The following sections describe our large-scale HSI study conducted in an interdisciplinary surgical ICU (Section 7.2.1), the external dataset used to evaluate the generalizability of an algorithm trained on existing HSI data of selectively chosen cohorts to an ICU population (Section 7.2.2), and our DL approach for automated sepsis diagnosis and mortality prediction from skin HSI data (Section 7.2.3).

7.2.1 Intensive Care Unit Dataset

In a prospective observational study, we collected HSI and corresponding RGB skin images from patients admitted to the interdisciplinary surgical ICU at the University Hospital Heidelberg, Germany. All adult patients admitted to the ICU between October 24, 2022, and December 15, 2023, were included. The study followed the ethical standards of the 1964 Declaration of Helsinki and its subsequent revisions. Approval was obtained from the Ethics Committee of the Medical Faculty of Heidelberg University (study reference number: S-288/2022), and the trial was prospectively registered in the German Clinical Trials Register (DRKS00029709).

Study Design HSI cubes of the patients' skin, specifically at the measurement sites palm and annular finger, were acquired on the day of admission to the ICU. These measurement sites were selected for their accessibility and lower melanin content compared to other skin areas [373]. For each patient, the decision to image the left or right hand was based on ensuring that the selected hand was not utilized for intravascular access or intra-arterial cannulation. Characteristic spectra for septic and non-septic patients, as well as for survivors and non-survivors, are illustrated in Figure B.17.

Alongside the HSI data, we collected 45 structured clinical data. Of these, 33 parameters are typically available within 1 h of ICU admission, comprising demographics, vital signs, blood gas analysis (BGA) parameters, and therapy details such as organ replacement, ventilation settings and vasopressor or inotrope doses. In addition, 12 laboratory parameters were collected, which are typically available within 10 h of admission. Descriptive statistics are summarized in Table 7.1 (clinical data available within 1 h) and Table 7.2 (laboratory parameters), with more detailed distribution figures for septic and non-septic patients provided in the appendix (Figure B.18 – Figure B.24).

Table 7.1: Descriptive statistics for non-septic and septic patients, as well as survivors and non-survivors. They include clinical data available within 1 h of intensive care unit admission, covering demographics, vital signs, blood gas analysis (BGA) values, organ replacement therapies, ventilation parameters, and vasopressor/inotrope dosing. For ratio-scaled variables, means with standard deviation in brackets are shown. For nominal variables, patient counts per category are listed. For binary therapy variables, the proportion of patients receiving the treatment is reported. Abbreviations: mean arterial pressure (MAP), pulse oxymetrical oxygen saturation (SpO_2), carbon dioxide partial pressure (pCO2), oxygen partial pressure (pO2), oxygen saturation (sO2), hemoglobin (Hb), extracorporeal membrane oxygenation (ECMO), airway pressure release ventilation (APRV), fraction of inspired oxygen (FiO2), positive end-expiratory pressure (PEEP), peak inspiratory pressure (P-peak). Table adapted from [306].

attribute	no sepsis	sepsis	non survivor	survivor
number of subjects	308	129	68	415
		demographics		
age	$6.2 \cdot 10^1$ ($1.5 \cdot 10^1$)	$6.6 \cdot 10^1$ ($1.4 \cdot 10^1$)	$6.9 \cdot 10^1$ ($1.5 \cdot 10^1$)	$6.3 \cdot 10^1$ ($1.4 \cdot 10^1$)
sex	220 male 88 female	90 male 39 female	41 male 27 female	299 male 116 female
weight [kg]	$8.2 \cdot 10^1$ ($2.0 \cdot 10^1$)	$8.2 \cdot 10^1$ ($2.6 \cdot 10^1$)	$7.5 \cdot 10^1$ ($2.3 \cdot 10^1$)	$8.2 \cdot 10^1$ ($2.1 \cdot 10^1$)
type of weight measurement	245 estimated 53 measured	100 estimated 16 measured	52 estimated 8 measured	331 estimated 68 measured
		vital signs		
heart frequency [bpm]	$8.2 \cdot 10^1$ ($1.7 \cdot 10^1$)	$9.9 \cdot 10^1$ ($2.1 \cdot 10^1$)	$9.8 \cdot 10^1$ ($2.4 \cdot 10^1$)	$8.6 \cdot 10^1$ ($1.9 \cdot 10^1$)
sinusrhythm [%]	79	74	60	78
MAP [mmHg]	$8.1 \cdot 10^1$ ($1.4 \cdot 10^1$)	$7.6 \cdot 10^1$ ($1.3 \cdot 10^1$)	$7.7 \cdot 10^1$ ($1.3 \cdot 10^1$)	$8.0 \cdot 10^1$ ($1.4 \cdot 10^1$)
systolic blood pressure	$1.2 \cdot 10^2$ ($2.3 \cdot 10^1$)	$1.2 \cdot 10^2$ ($1.9 \cdot 10^1$)	$1.2 \cdot 10^2$ ($2.3 \cdot 10^1$)	$1.2 \cdot 10^2$ ($2.3 \cdot 10^1$)
temperature [°C]	$3.7 \cdot 10^1$ ($6.7 \cdot 10^{-1}$)	$3.7 \cdot 10^1$ (1.1)	$3.7 \cdot 10^1$ (1.1)	$3.7 \cdot 10^1$ ($7.5 \cdot 10^{-1}$)
SpO2 [%]	$9.7 \cdot 10^1$ (2.2)	$9.7 \cdot 10^1$ (4.0)	$9.6 \cdot 10^1$ (5.1)	$9.7 \cdot 10^1$ (2.3)
		BGA measurements		
pCO2 [mmHg]	$3.9 \cdot 10^1$ (5.8)	$4.4 \cdot 10^1$ (9.8)	$4.3 \cdot 10^1$ (9.8)	$4.0 \cdot 10^1$ (7.0)
pO2 [mmHg]	$9.8 \cdot 10^1$ ($3.4 \cdot 10^1$)	$1.0 \cdot 10^2$ ($2.5 \cdot 10^1$)	$1.0 \cdot 10^2$ ($2.5 \cdot 10^1$)	$9.9 \cdot 10^1$ ($3.2 \cdot 10^1$)
sO2 [%]	$9.7 \cdot 10^1$ (1.6)	$9.6 \cdot 10^1$ (2.8)	$9.6 \cdot 10^1$ (3.5)	$9.7 \cdot 10^1$ (1.6)
Hb (BGA) [g/dl]	9.7 (1.7)	9.4 (1.7)	9.5 (1.6)	9.5 (1.7)
lactate [mg/dl]	$1.6 \cdot 10^1$ ($1.4 \cdot 10^1$)	$2.7 \cdot 10^1$ ($3.4 \cdot 10^1$)	$4.6 \cdot 10^1$ ($5.3 \cdot 10^1$)	$1.5 \cdot 10^1$ ($1.1 \cdot 10^1$)
pH	7.4 ($5.8 \cdot 10^{-2}$)	7.4 ($8.8 \cdot 10^{-2}$)	7.4 ($1.0 \cdot 10^{-1}$)	7.4 ($6.5 \cdot 10^{-2}$)
type BGA	274 arterial 7 venous	104 arterial 1 venous	56 arterial	358 arterial 10 venous
		organ replacement therapies		
renal replacement therapy [%]	4	20	28	7
ECMO [%]	1	2	3	1
impella [%]	0	1	4	0
liver replacement therapy [%]	1	2	4	0
		ventilation parameters		
invasive ventilation [%]	48	95	93	59
ventilation [%]	23	80	78	34
APRV [%]	0	2	3	0
FiO2 [%]	$3.2 \cdot 10^1$ ($1.0 \cdot 10^1$)	$4.4 \cdot 10^1$ ($1.8 \cdot 10^1$)	$4.4 \cdot 10^1$ ($1.7 \cdot 10^1$)	$3.4 \cdot 10^1$ ($1.3 \cdot 10^1$)
PEEP [mbar]	7.0 (2.3)	8.9 (3.2)	8.3 (3.3)	8.1 (2.9)
P-peak [mbar]	$2.0 \cdot 10^1$ (5.5)	$2.1 \cdot 10^1$ (6.1)	$2.2 \cdot 10^1$ (5.7)	$2.0 \cdot 10^1$ (5.8)
respiratory frequency [min^{-1}]	$1.7 \cdot 10^1$ (4.4)	$1.8 \cdot 10^1$ (5.3)	$1.7 \cdot 10^1$ (5.6)	$1.7 \cdot 10^1$ (5.0)
		dose of administered vasopressors and inotropes		
noradrenaline dose [µg/(kg min)]	$4.4 \cdot 10^{-2}$ ($9.4 \cdot 10^{-2}$)	$2.6 \cdot 10^{-1}$ ($2.6 \cdot 10^{-1}$)	$2.7 \cdot 10^{-1}$ ($3.0 \cdot 10^{-1}$)	$7.7 \cdot 10^{-2}$ ($1.4 \cdot 10^{-1}$)
adrenaline dose [µg/(kg min)]	$9.2 \cdot 10^{-4}$ ($1.1 \cdot 10^{-2}$)	$3.7 \cdot 10^{-3}$ ($2.3 \cdot 10^{-2}$)	$8.6 \cdot 10^{-3}$ ($3.2 \cdot 10^{-2}$)	$7.0 \cdot 10^{-4}$ ($1.0 \cdot 10^{-2}$)
vasopressin dose [Unit/(kg min)]	$3.8 \cdot 10^{-6}$ ($3.2 \cdot 10^{-5}$)	$5.4 \cdot 10^{-5}$ ($1.4 \cdot 10^{-4}$)	$5.2 \cdot 10^{-5}$ ($1.1 \cdot 10^{-4}$)	$1.2 \cdot 10^{-5}$ ($7.3 \cdot 10^{-5}$)
dobutamine dose [µg/(kg min)]	$2.0 \cdot 10^{-1}$ ($9.3 \cdot 10^{-1}$)	$6.1 \cdot 10^{-1}$ (1.9)	1.1 (2.4)	$2.7 \cdot 10^{-1}$ (1.2)

Table 7.2: Descriptive statistics for non-septic and septic patients, as well as for survivors and non-survivors (continuation). Table presents descriptive statistics for laboratory parameters obtained within 10 h of intensive care unit admission. Means with standard deviation in brackets are presented. Abbreviations: glomerular filtration rate (GFR), lactate dehydrogenase (LDH), C-reactive protein (CRP), hemoglobin (Hb), procalcitonin (PCT). Table adapted from [306].

attribute	no sepsis	sepsis	non survivor	survivor
creatinine [mg/dl]	1.3 (1.1)	1.9 (1.5)	1.7 ($9.7 \cdot 10^{-1}$)	1.5 (1.3)
GFR [ml/min]	$7.3 \cdot 10^{1}$ ($3.6 \cdot 10^{1}$)	$4.9 \cdot 10^{1}$ ($3.4 \cdot 10^{1}$)	$4.6 \cdot 10^{1}$ ($2.9 \cdot 10^{1}$)	$6.7 \cdot 10^{1}$ ($3.7 \cdot 10^{1}$)
LDH [Unit/l]	$5.4 \cdot 10^{2}$ ($7.8 \cdot 10^{2}$)	$6.8 \cdot 10^{2}$ ($1.6 \cdot 10^{3}$)	$1.3 \cdot 10^{3}$ ($2.3 \cdot 10^{3}$)	$4.8 \cdot 10^{2}$ ($6.7 \cdot 10^{2}$)
bilirubin [mg/dl]	1.9 (2.4)	2.4 (3.5)	2.7 (3.7)	1.9 (2.4)
CRP [mg/l]	$6.6 \cdot 10^{1}$ ($7.4 \cdot 10^{1}$)	$2.0 \cdot 10^{2}$ ($1.1 \cdot 10^{2}$)	$1.2 \cdot 10^{2}$ ($9.5 \cdot 10^{1}$)	$1.1 \cdot 10^{2}$ ($1.1 \cdot 10^{2}$)
leukocytes [nl^{-1}]	$1.1 \cdot 10^{1}$ (5.0)	$1.6 \cdot 10^{1}$ ($1.1 \cdot 10^{1}$)	$1.5 \cdot 10^{1}$ (9.7)	$1.3 \cdot 10^{1}$ (7.3)
Hb (lab) [g/dl]	9.9 (1.9)	9.8 (1.8)	9.6 (1.6)	9.8 (1.9)
platelets [nl^{-1}]	$1.6 \cdot 10^{2}$ ($8.3 \cdot 10^{1}$)	$2.1 \cdot 10^{2}$ ($1.4 \cdot 10^{2}$)	$1.8 \cdot 10^{2}$ ($1.2 \cdot 10^{2}$)	$1.8 \cdot 10^{2}$ ($1.1 \cdot 10^{2}$)
hematocrit [%]	$2.9 \cdot 10^{-1}$ ($5.3 \cdot 10^{-2}$)	$3.0 \cdot 10^{-1}$ ($5.5 \cdot 10^{-2}$)	$2.9 \cdot 10^{-1}$ ($5.1 \cdot 10^{-2}$)	$2.9 \cdot 10^{-1}$ ($5.4 \cdot 10^{-2}$)
sodium [mmol/L]	$1.4 \cdot 10^{2}$ (4.3)	$1.4 \cdot 10^{2}$ (6.0)	$1.4 \cdot 10^{2}$ (5.9)	$1.4 \cdot 10^{2}$ (4.9)
potassium [mmol/L]	4.5 ($5.3 \cdot 10^{-1}$)	4.7 ($6.2 \cdot 10^{-1}$)	4.7 ($6.8 \cdot 10^{-1}$)	4.5 ($5.4 \cdot 10^{-1}$)
PCT [ng/ml]	1.9 (7.7)	$5.2 \cdot 10^{1}$ ($1.6 \cdot 10^{2}$)	$2.3 \cdot 10^{1}$ ($6.3 \cdot 10^{1}$)	$1.6 \cdot 10^{1}$ ($9.6 \cdot 10^{1}$)

To compare our HSI-based models for sepsis diagnosis and mortality prediction against widely used clinical biomarkers and scores, we additionally collected a range of reference measures.

Rapid bedside scores for diagnosing sepsis include the capillary refill time (CRT) [259] and skin mottling score (SMS) [8], both relying on visual assessment of the patient's skin, as well as the national early warning score (NEWS) [280] and qSOFA score [320], which are based on cognitive function and vital signs. Sepsis diagnosis scores available within 10 h of admission include the Systemic Inflammatory Response Syndrome (SIRS) criteria [46], formerly employed for diagnosing sepsis, and the SOFA score, a cornerstone of the current definition of sepsis according to Sepsis-3 [320]. Distributions of these scores among the non-septic and septic patients in our ICU population are shown in the appendix (Figure B.25).

To assess disease severity and mortality risk, the vasoactive inotropic score (VIS) [107] provides a rapidly available bedside score that quantifies the degree of hemodynamic support required from vasopressors and inotropes. Within 10 h of admission, additional scores become available, including the SOFA score, which evaluates organ dysfunction, and the Acute Physiology and Chronic Health Evaluation (APACHE) II score [180], which measures overall disease severity. These scores are calculated from a combination of vital signs, laboratory results, and patient or therapy characteristics. While they are conventionally based on the most abnormal values within the preceding 24 h, we employed modified versions of SOFA and APACHE II that rely on the most recent values at admission, ensuring score availability on the day of ICU admission. Distributions of these scores among the survivors and non-survivors in our ICU population are shown in the appendix (Figure B.26).

Hyperspectral Image Acquisition The HSI data was acquired with the camera system TIVITA® 2.0 Surgery Edition (Diaspective Vision GmbH, Am Salzhaff, Germany). As described in Section 4.2.1, it images an area of approximately 16 cm × 11.5 cm at an imaging distance of approximately 50 cm, ensured by an integrated distance calibration system. The push-broom HSI device captures 100 spectral channels in the range 500–1000 nm at a spectral resolution of approximately 5 nm. The resulting HSI data cubes measure 640 × 480 × 100 (width × height × spectral channels). The acquisition of on image takes approximately 7 s. In addition to the HSI sensor, the system is equipped with an RGB sensor of identical spatial resolution, enabling automatic parallel acquisition of RGB images with each HSI cube.

To ensure that the scene was solely illuminated by the integrated LED lighting unit of the camera, the room lights were turned off and window blinds were lowered. Patient hands were stabilized by the examiner to minimize motion artifacts and standardize positioning, with a uniform background applied across all images.

Hyperspectral Image Annotation Despite the standardized hand positioning and uniform background, images could still contain elements such as wounds, tubes, wires, dressings, or gloved hands of the examiners. To prevent potential shortcut learning from such image elements, analyses were restricted to annotated skin regions. Circular annotations were chosen to ensure consistent measurement sites across patients, independent of the hand's rotation within the imaging plane and avoiding the aforementioned elements. Annotation radii were set to 100 px for the palm and 20 px for the annular finger. Finger annotations were centered on the fingertip, while palm annotations were centered on the palm, defined as the region enclosed by the thumb basal joint, metacarpophalangeal joints, and wrist. The annotations were performed using the built-in annotation software of the HSI system Tivita® Suite.

Definition and Labeling of Sepsis and Mortality Status Sepsis was diagnosed according to the Sepsis-3 criteria, defininig it as "life-threatening organ dysfunction caused by a dysregulated host response to infection" [320]. The SOFA score was used to quantify organ dysfunction, with sepsis identified by an acute increase of two or more points. Distinguishing sepsis-related organ failure from dysfunction due to non-septic inflammation is particularly challenging in a surgical ICU setting after surgical trauma. To ensure label accuracy and reduce ambiguity, a third label, "unsure", was introduced in addition to "sepsis" and "no sepsis". Each patient's sepsis status was independently evaluated by two expert anesthetists, with disagreements resolved by a senior anesthetist (the head of the department of anesthesia and intensive care). Mortality was determined via a follow-up conducted 30 days after inclusion.

Study Population A total of 508 patients were initially included in the study. For 71 patients, the sepsis status could not be determined. Consequently, automated sepsis diagnosis could only be assessed on the remaining 437 patients, of whom 129 (30 %) were diagnosed with sepsis, and 308 (70 %) did not have sepsis at the time of inclusion. Most septic patients had an abdominal focus (53 %), while the focus was respiratory for 17 %, skin or soft tissue for 5 %, and genitourinary for 3 % of the septic patients. Additionally, 8 % of septic patients had multiple infection foci, whereas in 14 % the focus remained unknown.

For 483 of the initial 508 patients, follow-up on mortality 30 days after ICU admission was successful. This cohort, comprising 415 (86 %) survivors and 68 (14 %) non-survivors, was used to evaluate automated mortality prediction. The mortality rate was higher among patients admitted with sepsis (27 % (35/129)), compared to patients without sepsis at admission (6 % (18/308)).

7.2.2 External Dataset

To assess the generalizability of a DL algorithm trained on existing HSI data of selectively chosen cohorts to an ICU population (RQ3.2), we leveraged the dataset from [81, 82, 80, 85]. After approval by the Ethics Committee of the Medical Faculty of Heidelberg University, Heidelberg, Germany (study reference number: S-148/2019) and registration with the German Clinical Trials Register (DRKS00017313), the data was acquired at the Heidelberg University Hospital in 2019. Informed consent was obtained from all participants or their legal guardians.

Study Design and Population A total of 25 septic patients were recruited for the study. Additionally, two control groups were established, comprising 25 healthy volunteers and 25 patients undergoing pancreatic surgery (referred to as pancreas group). Pregnant patients and those under 18 years of age were excluded from the study. Patients in the sepsis subgroup were included upon admission to the interdisciplinary surgical ICU if they met all Sepsis-3 criteria [320], with sepsis onset occurring within the previous 24 h. Healthy volunteers were included if they were free from both acute and chronic diseases. Patients in the pancreas subgroup were included if they were scheduled for an open pancreatic surgery and postoperative ICU admission.

While healthy subjects underwent HSI acquisition at a single time point, pancreas and sepsis patients were imaged at specific intervals over an approximately 72 h observation period.

As illustrated in Figure 7.1, pancreas patients were imaged prior to the induction of anesthesia, after anesthesia induction, before anesthesia emergence, approximately 6 h

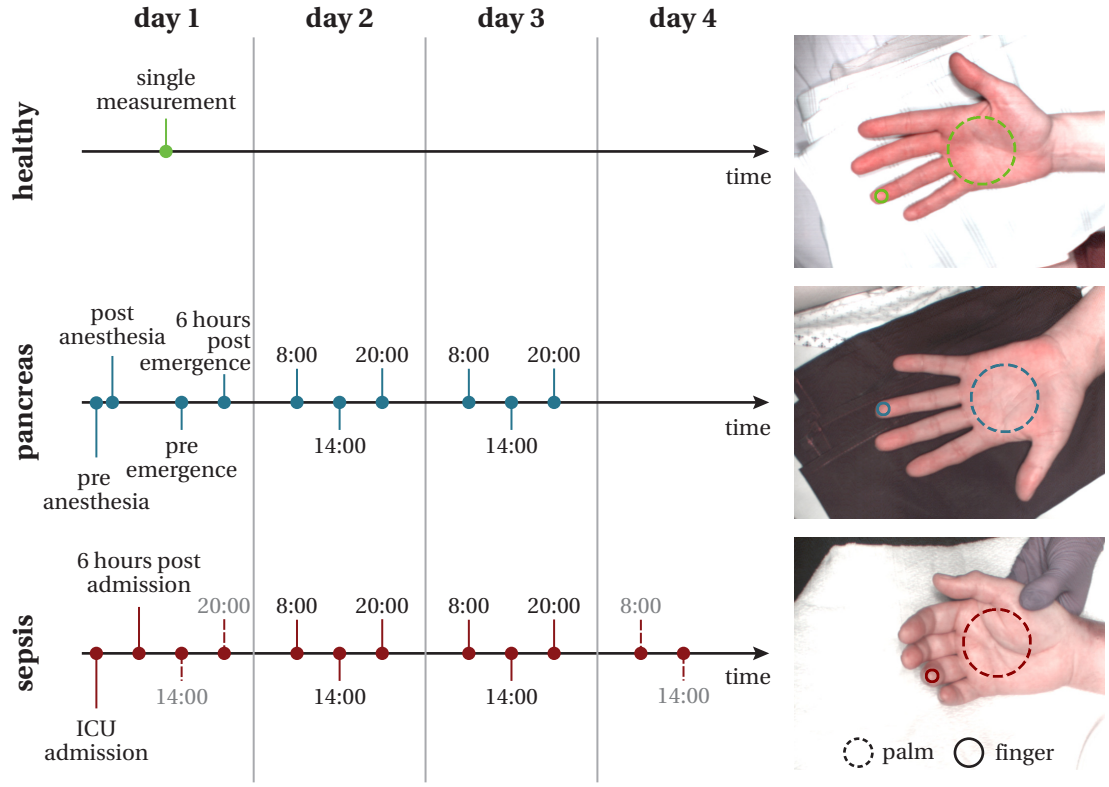


Figure 7.1: Study design of the external dataset. The dataset comprises hyperspectral imaging data of the palm (dashed circle) and annular finger (solid circle) for 25 healthy subjects (top), 25 patients undergoing pancreatic surgery (middle) and 25 septic intensive care unit (ICU) patients (bottom). Scheduled measurement timepoints on up to 4 consecutive days are illustrated on the left, with dashed lines indicating optional measurement timepoints.

after the emergence of anesthesia, and 3 times daily during the first two postoperative days, resulting in 10 consecutive measurements per patient.

For sepsis patients, HSI data was intended to be collected at admission to the ICU, approximately 6 h later, and, if admitted earlier in the day, additional measurements were supposed to be taken in the afternoon and evening. Subsequently, HSI data should be acquired 3 times a day for the following 2–3 days, such that 10 consecutive measurements were achieved. Although HSI measurements were successfully taken for all sepsis patients upon admission to the ICU, subsequent measurements were not always feasible, as 5 patients deceased and one was transferred to another hospital within the 72 h observation period. Additionally, 3 septic patients deviated from the measurement schedule due to undergoing surgery. A total of 563 HSI images were acquired.

Hyperspectral Image Acquisition The HSI data was acquired using the medical device-graded camera system TIVITA® Tissue (Diaspective Vision GmbH, Am Salzhaff, Germany). It is equipped with an 8 mm focal length lens, yielding a field of view of 20 cm × 30 cm at an imaging distance of approximately 50 cm. The spectral specifications and image dimensions match those of the TIVITA® 2.0 Surgery used in our ICU study (Section 7.2.1), although no RGB images were captured alongside the HSI acquisition. A detailed comparison of the devices is provided in Chapter 4. Unlike in the ICU study, where palm and finger had to be imaged separately, the larger field of view of the TIVITA® Tissue allowed both measurement sites to be captured in a single image (cf. Figure 7.1).

To minimize motion artifacts, all subjects were instructed to lie still during image acquisition, with their hand placed next to their body on the bed. If necessary, such as with an unconscious patient, the examiner assisted by gently positioning the hand.

Hyperspectral Image Annotation Circular regions were annotated on the palm and annular finger of each patient, adhering to the same guidelines used for skin annotations in our ICU study. To account for the difference in field of view between the cameras, the annotation radii were set to 70 px for the palm and 13 px for the annular finger.

7.2.3 Hyperspectral Image Analysis

To address RQ3.1, we developed DL classifiers for sepsis diagnosis and mortality prediction using different input modalities and measurement sites: HSI data, TPI cubes and RGB images from both palm and finger sites. We further compared our classification based on HSI patches to a model based on median spectra across the annotated areas

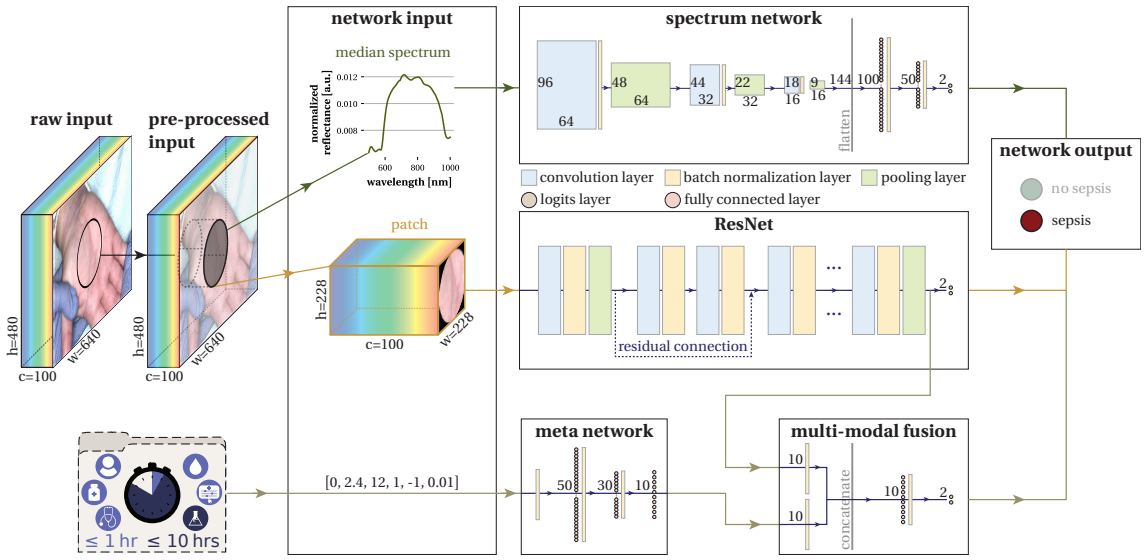


Figure 7.2: Overview of our deep learning pipeline for automated sepsis diagnosis and mortality prediction. After preprocessing the HSI data, including calibration and normalization, circular skin regions were manually annotated to extract both median spectra and image patches. Median spectra were classified using the spectrum#HSI model, a network with 3 convolutional and two fully connected layers, previously established for pixel-based organ segmentation [329] (cf. Chapter 5). Patch-based classification was performed with the patch#HSI model, based on a ResNet14d architecture [133, 365]. To compare modalities, analogous models were trained on TPI and RGB patches (patch#TPI and patch#RGB). To assess the benefit of structured clinical data, we employed the patch#HSI + clinical data model, in which the output of the patch#HSI network is concatenated with the output of a meta network processing clinical features, and the combined representation is further processed by a multimodal fusion network. Clinical data alone was processed with a random forest classifier (clinical data model). Mortality prediction models were trained analogously to the sepsis diagnosis models.

to assess the role of spatial context in HSI. To investigate whether structured clinical data can improve diagnostic and predictive performance (RQ3.3), we extended the patch-based classification model to incorporate clinical features and compared it to a model solely based on clinical data.

An overview of our DL classifiers is provided in Figure 7.2. The subsequent paragraphs detail our data preprocessing steps, model architectures, and training parameters.

Data Preprocessing The HSI cubes were first calibrated using white and dark reference cubes to eliminate sensor noise and convert the spectra from radiance to reflectance [141] (cf. Section 2.1.3). Following our recommendations to address calibration shifts (cf.

Section 4.4), calibration files were captured daily for the ICU dataset. After calibration, ℓ^1-normalization was applied across the spectral dimension.

Based on the HSI cubes, the tissue parameter index images StO_2, NPI, THI, and TWI were derived using the formulas from [141]. A TPI cube with dimensions $640 \times 480 \times 4$ (width $\times$ height $\times$ number of channels) was generated by stacking all 4 index images. Although HSI, TPI and RGB data were in practice generated by the same device in our study, we refer to these input types as different *modalities* to highlight that future applications could use TPI and RGB data from a different device (e.g., an MSI camera, a conventional RGB camera).

To assess the importance of spatial context in the HSI data, we compared the median spectra of the annotated regions with patches generated by cropping the images to a square tightly encompassing the circular annotation. Pixels outside the annotated area were blackened out. The patches were resized using bilinear interpolation to ensure that identical input dimensions were used across palm and finger data. The resulting dimensions were $224 \times 224 \times 100$, $224 \times 224 \times 4$, and $224 \times 224 \times 3$ (width $\times$ height $\times$ number of channels) for the HSI, TPI, and RGB patches, respectively.

The proportion of missing clinical parameter values was low, averaging only 1.6 %. Missing entries were imputed using a value of -1.

Sepsis Diagnosis and Mortality Prediction from Median Spectra For our median spectra-based models, referred to as spectrum#HSI, we adopted a DL architecture that we previously used for organ classification from median spectra [329] and pixel-based organ segmentation [308] (cf. Section 5.2.2). It comprises 3 one-dimensional convolutional layers, using 64 filters in the first, 32 filters in the second, and 16 filters in the third layer. Each convolution uses a kernel size of 5 and after each convolutional layer, an average pooling layer is applied across the spatial dimensions with a kernel size of two. The output from the final convolutional layer is flattened and fed into two fully connected layers, with the first layer containing 100 neurons and the second layer containing 50 neurons. A final linear layer computes the class logits.

This architecture was selected for its simplicity and effectiveness in analyzing spectral information. The convolutional layers capture local spectral patterns, while stacking 3 layers with a small kernel size efficiently expands the receptive field. The fully connected layers make decisions based on the global context, allowing the model to balance local and global information processing while remaining computationally efficient.

The same training setup as in [329] was used: The ELU activation function [67] was employed, with batch normalization applied to all layers except for the pooling layers. The model was optimized using the CE loss function and trained with the AdamW optimizer [213], utilizing an exponential learning rate schedule (initial learning rate: 0.0001,

decay rate γ: 0.9, Adam decay rates β_1: 0.9 and β_2: 0.999). Network regularization was implemented with a weight decay of 0.001. To avoid overfitting, dropout regularization was applied with a rate of 0.5. The model was trained over 10 epochs, with each epoch consisting of 500 000 median spectra using a batch size of 20 000 median spectra. acswa [156] was applied over the final two epochs. Oversampling was applied to ensure equal representation of all classes.

Sepsis Diagnosis and Mortality Prediction from Patches For patch-based classification, we employed a ResNet14d architecture [133, 365], initialized with ImageNet pretrained weights. A CNN architecture was selected due to its widespread use in medical HSI classification and its advantages over traditional ML approaches, such as improved accuracy and computational efficiency through weight sharing and hardware optimization [176]. Employing standardized architectures with pretrained weights further accelerates convergence and typically enhances performance compared to training CNNs from scratch, especially in small medical datasets [335]. Depending on the input modality, this model is referred to as patch#HSI (for HSI data), patch#TPI (for TPI cubes), or patch#RGB (for RGB images).

To ensure a fair comparison across modalities while minimizing computational cost and environmental impact, all patch-based classification models were trained with an identical training setup and fixed hyperparameters instead of modality-specific hyperparameter tuning. During training, data augmentation techniques were applied, including random horizontal and vertical flips, as well as random rotations up to ±180°, each with a probability of 0.5. The CE loss function was used, and the AdamW optimizer [213] was employed with an exponential learning rate schedule (initial learning rate: 0.001, decay rate γ: 0.99, Adam decay rates β_1: 0.9 and β_2: 0.999). A weight decay of 0.001 was applied for network regularization. The model was trained for 10 epochs, with acswa [156] applied during the final two epochs. Each epoch consisted of 500 patches, and the batch size was set to 32 patches. To ensure balanced class distribution within each batch, underrepresented classes were oversampled.

Multimodal Sepsis Diagnosis and Mortality Prediction from Patches and Clinical Data Our multimodal patch#HSI + clinical data model is composed of two submodules: As in the patch#HSI model, the HSI data are processed with a ResNet14d backbone pretrained on ImageNet, up to the bottleneck layer. The clinical data are processed by a dedicated submodel consisting of two fully connected blocks, each comprising a linear, batch normalization, ELU activation, and dropout layer. The first block uses a linear layer of size 50, and the second a size of 30. These are followed by a linear head of size 10, chosen to match the bottleneck dimension of the patch#HSI submodel. After batch normalization, the bottleneck features from both submodels are concatenated and passed through an additional fully connected block, before reaching the final

classification head. The training setup and hyperparameter settings were set identical to those of the patch-based models.

Sepsis Diagnosis and Mortality Prediction from Clinical Data Given the widespread use of random forests for sepsis prediction from EHR data [375], we implemented a 100-tree random forest trained exclusively on clinical data, referred to as the clinical data model. The sklearn implementation [261] was employed with default parameters, except for enabling balanced class weights to account for class imbalance by weighting classes inversely proportional to their frequencies in the training set.

Reduction of Non-determinism in the Network Training Non-determinism in neural network training is undesirable as it results in non-reproducible outcomes [263]. As exact reproducibility would necessitate using slower deterministic operations and result in longer training times, we implemented several measures to improve reproducibility while preserving training efficiency: All models were trained on the same hardware, namely a single NVIDIA® GeForce RTX™ 4090 GPU (Nvidia Corporation, Santa Clara, California, United States of America). Additionally, we set the number of workers as well as a random seed for the training process to ensure consistent random initialization of weights and workers.

7.3 Experiments and Results

The purpose of our experiments was to assess the feasibility of automated sepsis diagnosis using DL-based HSI analysis. Specifically, we aimed to determine the optimal measurement site (palm vs. finger), input modality (HSI vs., TPI vs. RGB data) and spatial granularity (patches vs. median spectra) for this task (RQ3.1, Section 7.3.2). We further assessed the generalizability of an algorithm trained on existing HSI data of selectively chosen cohorts to an ICU population (RQ3.2, Section 7.3.3), as well as the added value of structured clinical data (RQ3.3, Section 7.3.4). Finally, we compared the performance of our models against established clinical biomarkers and scores (RQ3.4, Section 7.3.5). Details of the experimental setup are provided in Section 7.3.1.

7.3.1 Experimental Setup

This section describes the dataset splits and validation metrics, the experimental setup for addressing RQ3.2, and the procedures for hierarchical data aggregation, statistical testing, and feature importance analysis of clinical data.

Dataset Splits To enable a fair model comparison, a consistent training and validation setup was applied across all models. Due to the limited amount of data, we opted against holding out a single untouched test set for validation. Instead, we utilized a nested cross-validation scheme consisting of 5 outer and inner folds, allowing for a more reliable performance estimation using the entire dataset [347].

As described above, we set random seeds during training to enhance reproducibility. To further address the variability introduced by different random seeds and improve the network stability, each training was repeated 3 times with distinct random seed settings. Ensembling was applied to the validation sets by averaging the logits across these 3 repetitions. For the test data, ensembling was performed by averaging logits from networks across all 5 folds and 3 repetitions, resulting in ensembling a total of 15 networks.

Validation Metrics In line with the recommendations in [222], we validated the model performance using the receiver operating characteristic (ROC) curve and AUROC. To account for sampling variability and compute confidence intervals, for each test set T, we generated 1000 bootstrap samples, each consisting of $|T|$ instances randomly drawn with replacement.

Generalizability Experiment Most models were trained and validated on the ICU dataset described in Section 7.2.1. Only for the generalizability experiment (RQ3.2), the patch#HSI model was trained and validated on the external dataset from Section 7.2.2, with the ICU dataset serving as a hold-out test set. Training was performed on the entire external dataset, following the nested cross-validation scheme described above. To reflect the anticipated clinical scenario of early sepsis diagnosis at ICU admission, validation for septic patients was restricted to their first measurement taken at admission. For validation on the ICU dataset, logits were averaged across the 75 networks produced by nested cross-validation (5 outer folds, 5 inner folds, 3 repetitions).

Hierarchical Aggregation To account for the hierarchical structure of the data, we aggregated median spectra, functional tissue parameter indices, and clinical metadata at the subject level. For ratio- and interval-scaled parameters, the averages across all measurements were computed, while for nominal and boolean parameters, the mode was used to represent the subject-level data. These subject-wise aggregates were then used as the foundation for visualizations and descriptive statistics.

Statistical Testing To examine whether significant differences in functional tissue parameter indices exist between septic and non-septic patients, as well as between

survivors and non-survivors, a two-sided Welch's t-test [363] was performed. The analysis encompassed 8 tests in total (4 functional parameters $\times$ 2 datasets). The overall significance level was set at 0.05 and to mitigate the accumulation of alpha errors from multiple testing, the Bonferroni correction [47] was applied, adjusting the significance level to 0.0125 per test.

Feature importance of clinical data We assessed the importance of clinical features by applying recursive feature elimination (RFE) [130] to the clinical data models. To adapt RFE to the 5-fold cross-validation scheme of the inner folds, feature importances were averaged across folds before removing the least important feature from the input set. This procedure was carried out independently within each of the 5 outer folds.

7.3.2 Hyperspectral Imaging-Based Sepsis Diagnosis and Mortality Prediction

Our approach to automated sepsis diagnosis is based on the hypothesis that microcirculatory dysfunction and edema formation in septic patients are reflected in HSI measurements of the skin.

Shifts in Functional Tissue Parameters To test this hypothesis, we compared the distribution of the functional tissue parameters StO_2, NPI, THI and TWI, between non-septic and septic patients, as well as between survivors and non-survivors. The distributions of these functional tissue parameters are displayed in Figure 7.3 for the palm measurement site. Distributions for the finger measurement site are shown in Figure 7.4. Exemplary functional tissue parameter images of palm and finger skin for a non-septic survivor and a septic non-survivor are provided in Figure 7.5.

Statistical analysis showed that, for both palm and finger measurements, septic patients exhibited significantly lower StO_2 and higher THI compared to non-septic patients. Additionally, palm measurements revealed significantly elevated TWI, while finger measurements showed significantly reduced NPI in septic patients. In non-survivors, both palm and finger measurements indicated significantly lower StO_2 and NPI, along with higher THI compared to survivors. Palm measurements further revealed significantly elevated TWI in non-survivors. Detailed results of the statistical analyses are provided in Table B.1.

Optimal Measurement Site, Modality and Spatial Granularity As shown in Figure 7.6, across all modalities and spatial granularities, DL-based sepsis diagnosis achieved higher performance at the palm site compared to the finger site (e.g., AUROC of 0.80

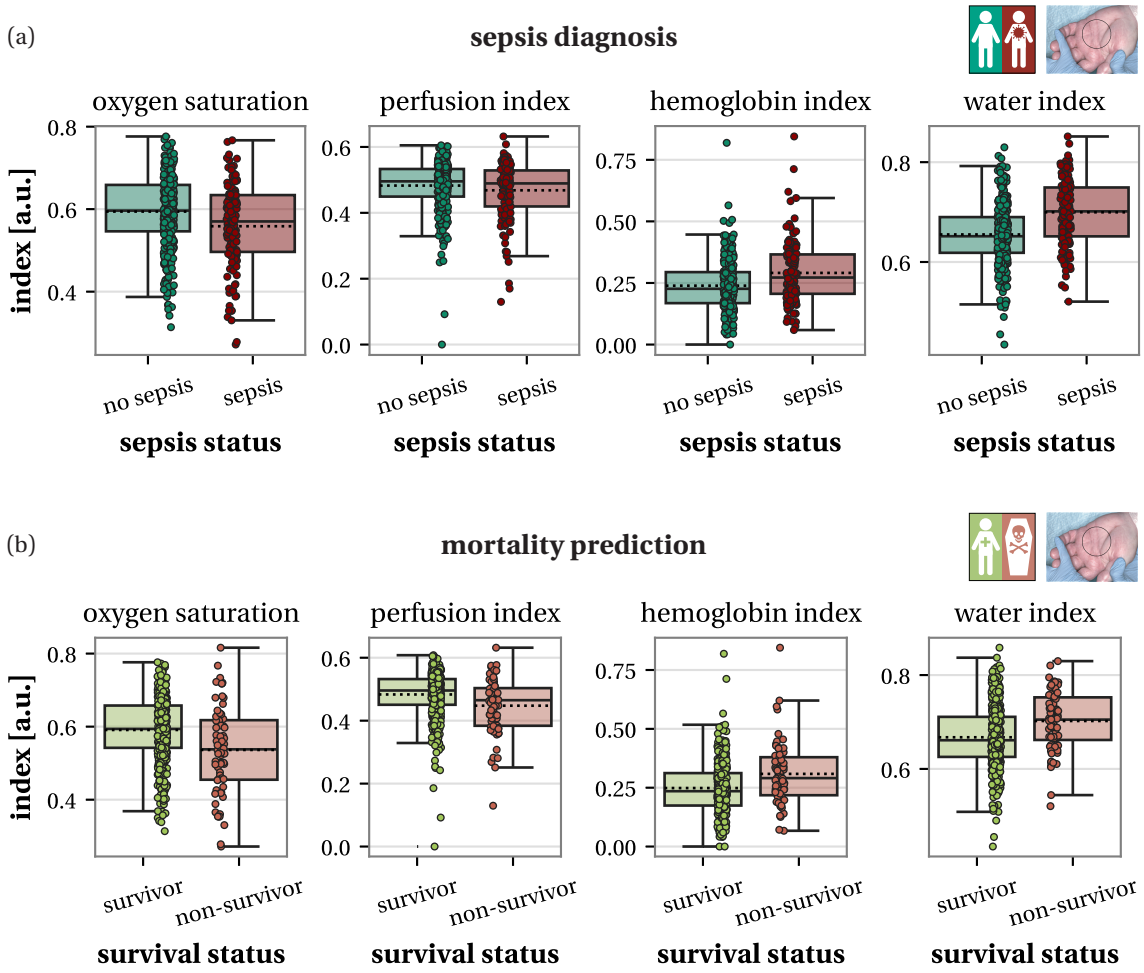


Figure 7.3: Characteristic shifts in functional tissue parameter index distributions of palm skin for (a) non-septic vs. septic patients and (b) survivors vs. non-survivors. Subfigures show the distributions of tissue oxygen saturation, tissue perfusion index, tissue hemoglobin index and tissue water index derived from hyperspectral imaging measurements of the palm skin. Each box displays the interquartile range of the distribution, with whiskers showing the range excluding outliers, and median and mean indicated by a solid and dotted line, respectively. Each marker represents one patient. Figure adapted from [306].

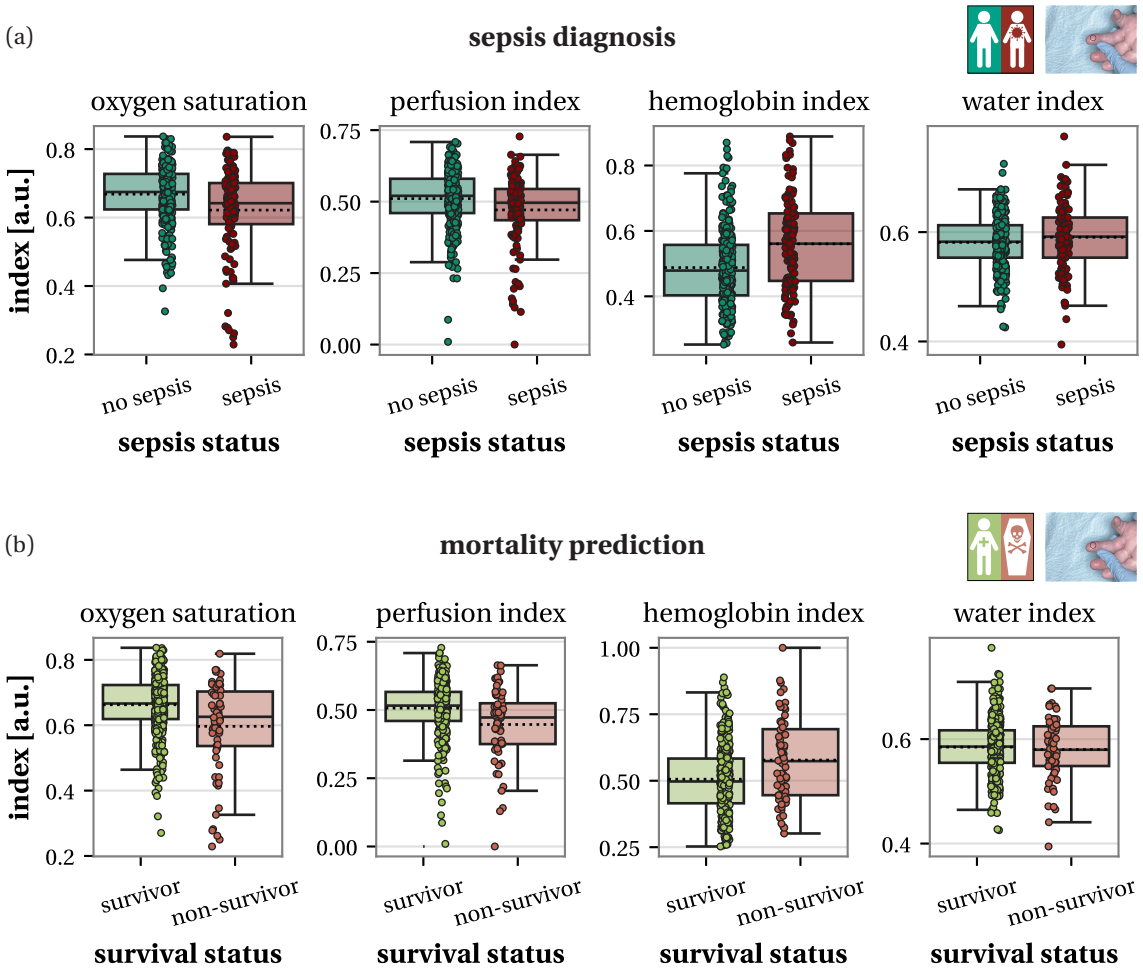


Figure 7.4: Characteristic shifts in functional tissue parameter index distributions of finger skin for (a) non-septic vs. septic patients and (b) survivors vs. non-survivors. Subfigures show the distributions of tissue oxygen saturation, tissue perfusion index, tissue hemoglobin index and tissue water index derived from hyperspectral imaging measurements of the finger skin. Each box displays the interquartile range of the distribution, with whiskers showing the range excluding outliers, and median and mean indicated by a solid and dotted line, respectively. Each marker represents one patient. Figure adapted from [306].

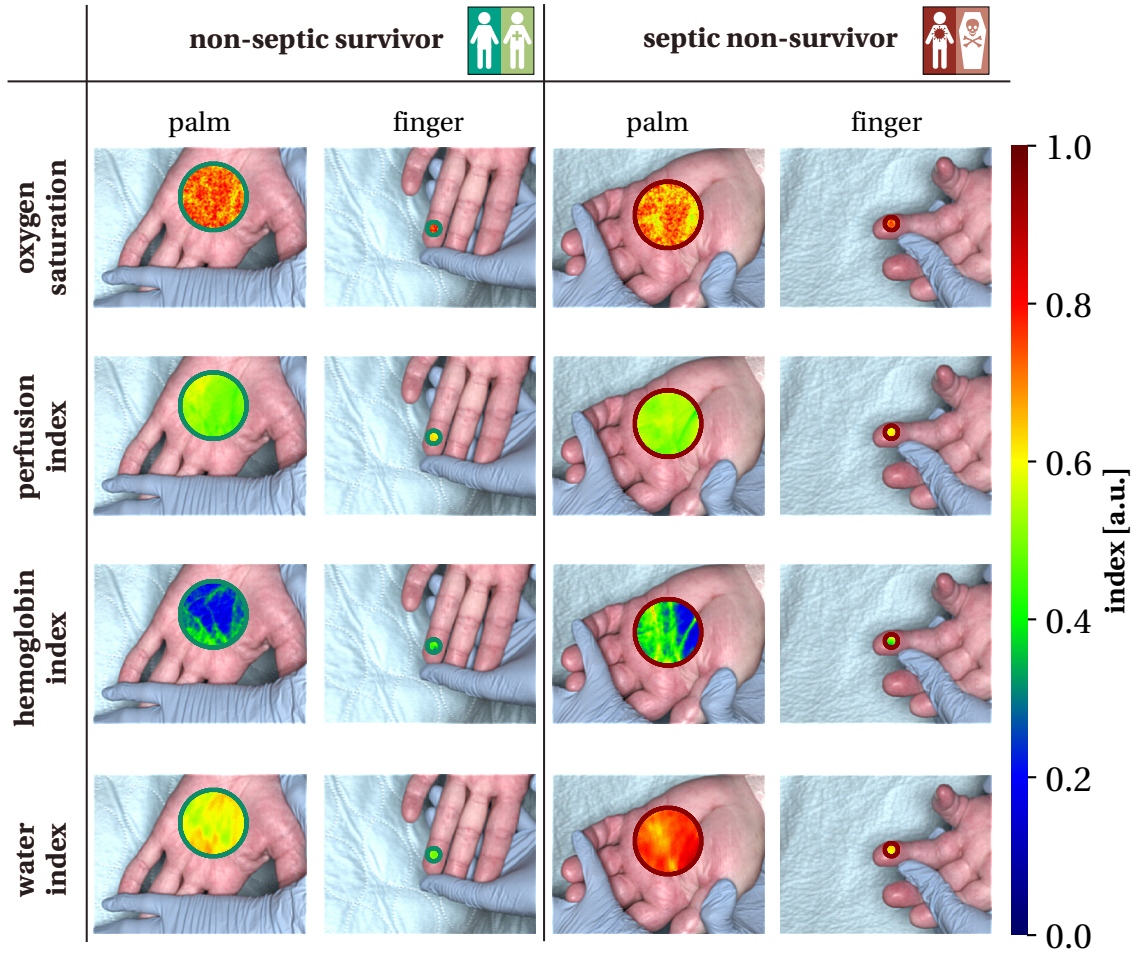


Figure 7.5: Exemplary functional tissue parameter images of palm and finger skin for a non-septic survivor (left) and a septic non-survivor (right). Shown are reconstructed RGB images overlaid with color-coded maps of the 4 functional tissue parameter indices tissue oxygen saturation, tissue perfusion index, tissue hemoglobin index, and tissue water index within the annotated circular skin region.

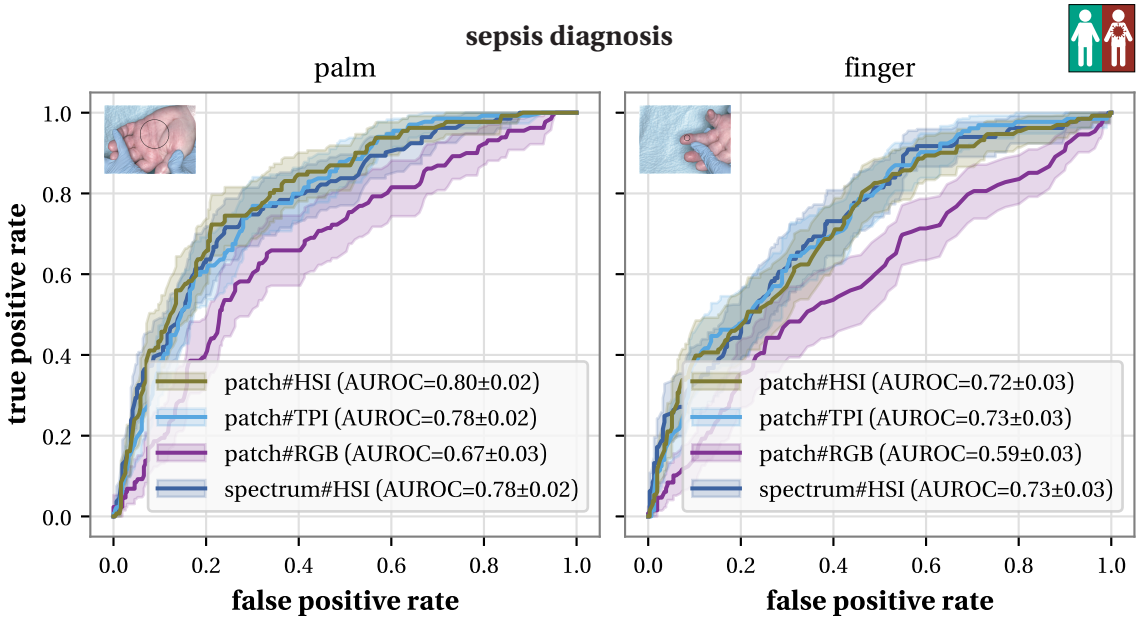


Figure 7.6: Sepsis diagnosis performance across different measurement sites, modalities and spatial granularities. Receiver operating characteristics (ROCs) are shown for models based on hyperspectral imaging (HSI) data (patch#HSI), stacked tissue parameter images (patch#TPI), RGB data (patch#RGB), and median HSI spectra (spectrum#HSI) of the palm (left) and annular finger (right). The 95 % confidence interval derived from 1000 bootstrap samples is shown as shaded area, with mean and standard deviation of the area under the receiver operating characteristic curve (AUROC) reported in the legend. Figure adapted from [306].

(95 % CI [0.76; 0.84]) vs. 0.72 (95 % CI [0.67; 0.78]) for the patch#HSI model). Likewise, mortality prediction performed better at the palm than at the finger site (e.g., AUROC of 0.72 (95 % CI [0.65; 0.79]) vs.0.66 (95 % CI [0.59; 0.73]) for the patch#HSI model), as illustrated in Figure 7.7.

For both sepsis diagnosis and mortality prediction, HSI outperformed conventional RGB imaging, achieving up to a 23 % improvement in classification performance. Models trained directly on HSI data and those using TPI data showed similar performance, indicating that TPI data capture information relevant for sepsis diagnosis and mortality prediction. Also, models based on HSI patches performed similarly to those based on median spectra, suggesting that spatial context within the annotated region plays a minor role in sepsis diagnosis and mortality prediction.

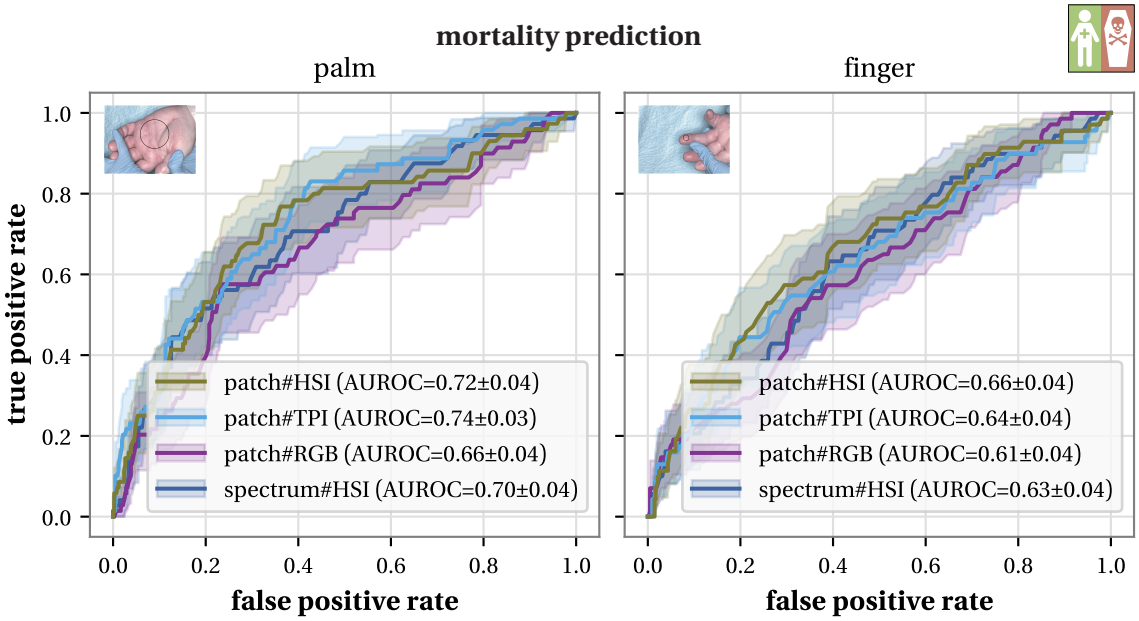


Figure 7.7: Mortality prediction performance across different measurement sites, modalities and spatial granularities. Receiver operating characteristics (ROCs) are shown for models based on hyperspectral imaging (HSI) data (patch#HSI), stacked tissue parameter images (patch#TPI), RGB data (patch#RGB), and median HSI spectra (spectrum#HSI) of the palm (left) and annular finger (right). The 95 % confidence interval derived from 1000 bootstrap samples is shown as shaded area, with mean and standard deviation of the area under the receiver operating characteristic curve (AUROC) reported in the legend. Figure adapted from [306].

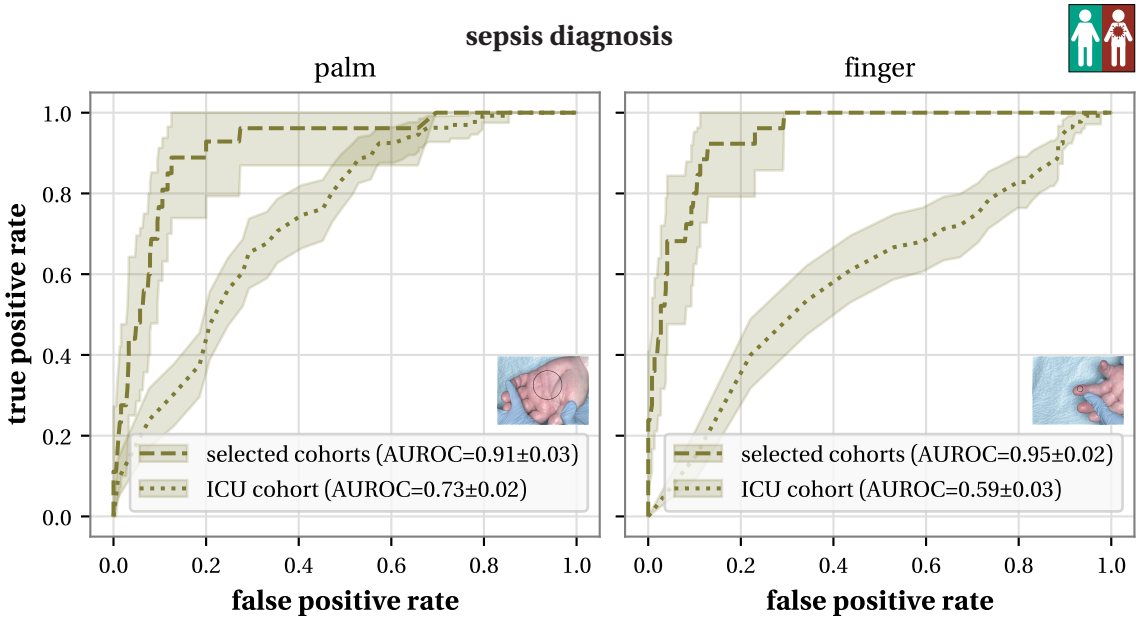


Figure 7.8: Performance drop of a sepsis diagnosis model trained on selected cohorts when tested on the intensive care unit (ICU) cohort. The patch#HSI model was trained on the dataset from [81, 80], comprising 25 septic patients, 25 patients undergoing pancreatic surgery, and 25 healthy volunteers. This dataset was previously used to assess the potential of hyperspectral imaging (HSI) for sepsis diagnosis [85]. Dashed lines indicate in-distribution performance, while dotted lines represent out-of-distribution performance on the ICU cohort. The 95 % confidence interval derived from 1000 bootstrap samples is shown as shaded area, with mean and standard deviation of the area under the receiver operating characteristic curve (AUROC) reported in the legend.

7.3.3 Sepsis Diagnosis Performance under Population Shift

Previous studies compared selectively chosen cohorts, such as septic patients against healthy controls or patients undergoing pancreatic surgery. We evaluated the generalizability of models trained on such data to our ICU population. To this end, we trained the patch#HSI model on the external dataset described in Section 7.2.2 and tested it on our ICU dataset. As shown in Figure 7.8, the palm-based model achieved an AUROC of 0.91 (95 % CI [0.85; 0.96]) on in-distribution data, but its performance dropped markedly to 0.73 (95 % CI [0.69; 0.78]) on the OOD ICU dataset. At the finger measurement site, the performance gap is even more pronounced, with in-distribution performance reaching 0.95 (95 % CI [0.90; 0.97]) but dropping to 0.59 (95 % CI [0.53; 0.65]) in the OOD setting. These findings support our hypothesis that models trained on selectively chosen cohorts fail to generalize to realistic clinical settings.

7.3.4 Performance Boost Through Multimodal Data Fusion

As automated sepsis diagnosis and mortality prediction may be further improved by incorporating structured clinical data, we extended the palm-based patch#HSI model to a multimodal patch#HSI + clinical data model (cf. Figure 7.2). We compared this multimodal model to the patch#HSI model and a clinical data model based solely on structured clinical data.

As shown in Figure 7.9, incorporating all clinical data available within 1 h of admission to the ICU into the patch#HSI + clinical data model yielded better sepsis diagnosis performance, with the AUROC increasing from 0.80 (95 % CI [0.76; 0.84]) to 0.90 (95 % CI [0.87; 0.93]). When additional clinical data available within 10 h from ICU admission were included, the AUROC further rose to 0.94 (95 % CI [0.92; 0.96]). Although the clinical data model performed slightly better than the patch#HSI + clinical data model when using the full dataset available within 10 h, the combined approach proved substantially superior when only limited clinical features were accessible – a situation frequently encountered in emergency care, outpatient settings, and LMICs.

We assessed the importance of clinical features through RFE [130] on the clinical data model, starting from the complete set of features available within one or 10 h of admission, respectively. Results are summarized in Figure B.27 (1 h) and Figure B.28 (10 h). As shown in Figure 7.9, sequentially adding features by importance revealed that combining HSI data with a single readily available bedside parameter – the administered noradrenaline dose – already yielded an AUROC of 0.87 (95 % CI [0.83; 0.90]).

As shown in Figure 7.10, combining HSI and clinical features also enhanced mortality prediction. When including all clinical data available within 1 h of ICU admission, the AUROC increased from 0.72 (95 % CI [0.65; 0.79]) to 0.82 (95 % CI [0.76; 0.88]), and further to 0.83 (95 % CI [0.78; 0.88]) upon incorporating data from the first 10 h. The patch#HSI + clinical data model consistently performed better than the clinical data model when sequentially adding clinical features according to their importance, with the largest advantage observed when only few features were available. The 3 most important features within 1 h of admission – lactate, pH, and noradrenaline dose – combined with palm HSI data yielded an AUROC of 0.80 (95 % CI [0.74; 0.85]).

7.3.5 Comparison to Established Clinical Biomarkers and Scores

To evaluate the clinical relevance of our sepsis diagnosis and mortality prediction models, we compared their performance against established clinical biomarkers and scores. For sepsis diagnosis, we included the NEWS, CRT, SMS, and qSOFA scores for data available within 1 h of ICU admission, and CRP, procalcitonin (PCT), SIRS criteria, and SOFA score for data available within 10 h. For mortality prediction, we compared

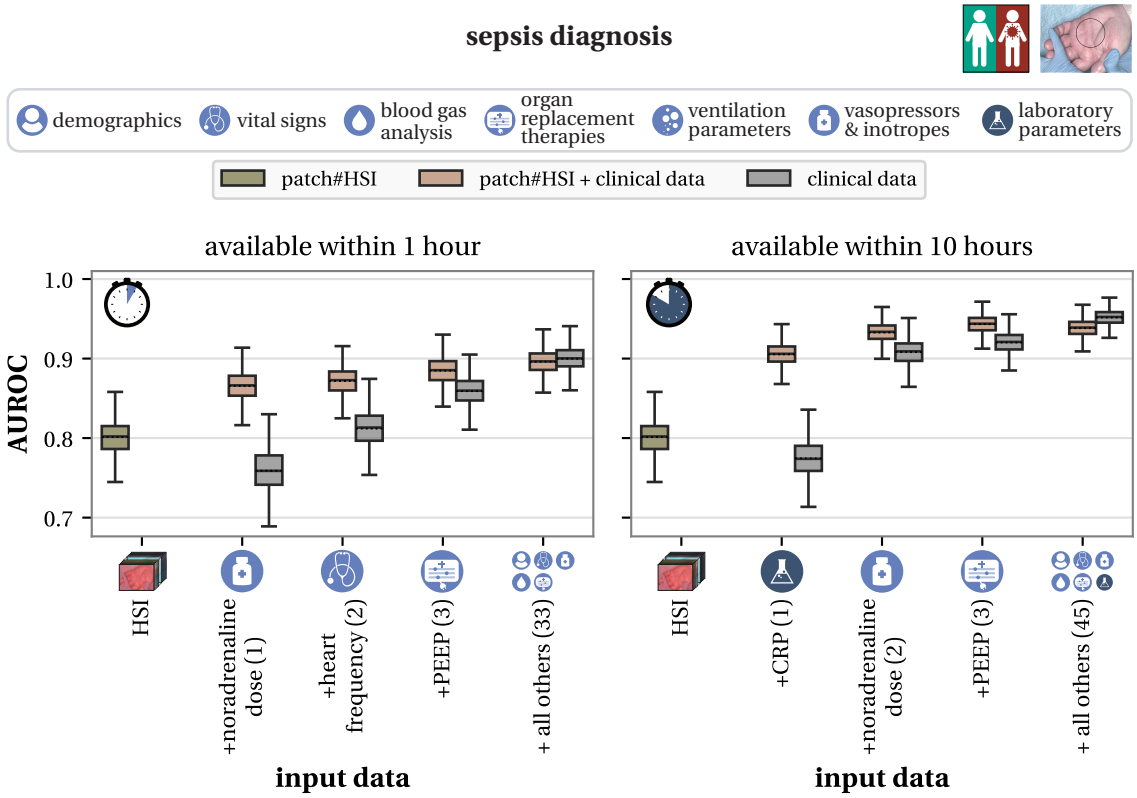


Figure 7.9: Sepsis diagnosis performance with added clinical data. Performance is shown for models based on hyperspectral imaging (HSI) data of the palm skin (patch#HSI), combined HSI and clinical data (patch#HSI + clinical data) and using only clinical data (clinical data), stratified by data availability within 1 h (left) and 10 h (right) after intensive care unit admission. Within each subplot, the patch#HSI model is compared to models utilizing the most important, two most important, 3 most important, or all clinical features available within the respective timeframe. The number of clinical features used is indicated in brackets. The feature importance ranking was derived via recursive feature elimination [130] using the clinical data model, starting from the full set of clinical features available within the given timeframe. Each box plot displays the distribution of the area under the receiver operating characteristic curve (AUROC) across 1000 bootstrap samples, with boxes spanning the interquartile range, whiskers showing the range excluding outliers, and solid and dashed lines marking the median and mean, respectively. Figure adapted from [306].

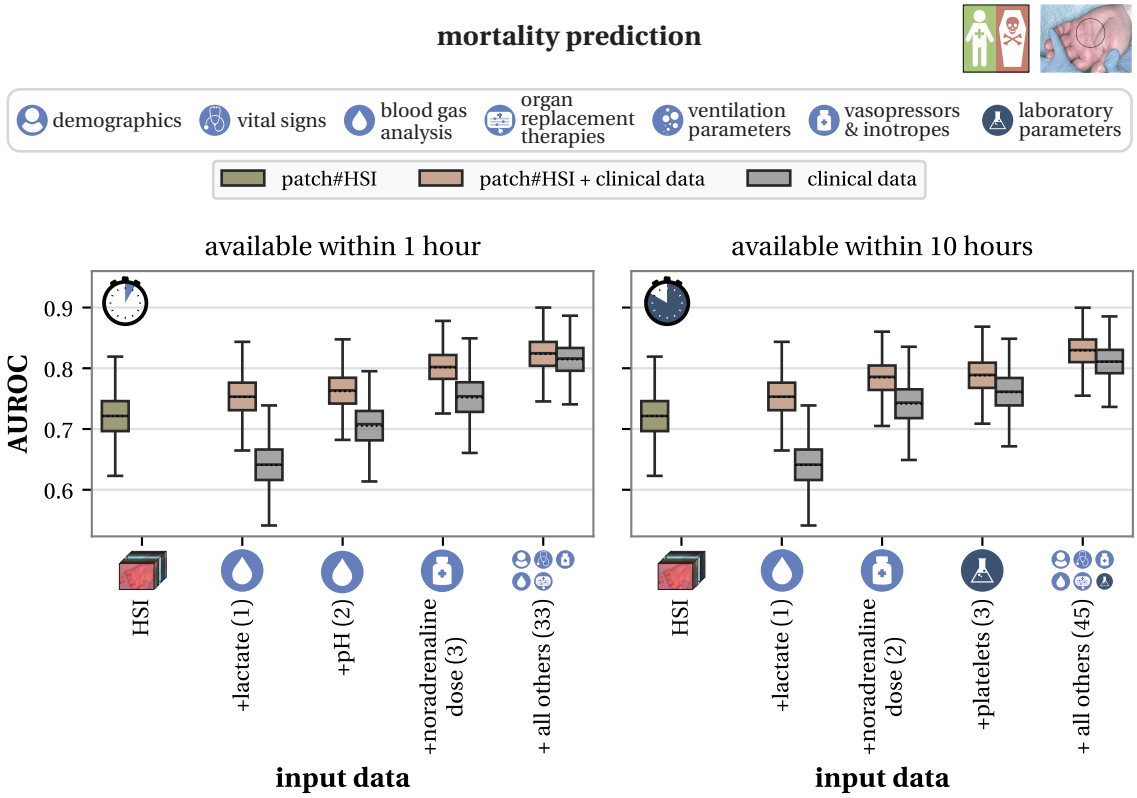


Figure 7.10: Mortality prediction performance with added clinical data. Performance is shown for models based on hyperspectral imaging (HSI) data of the palm skin (patch#HSI), combined HSI and clinical data (patch#HSI + clinical data) and using only clinical data (clinical data), stratified by data availability within 1 h (left) and 10 h (right) after intensive care unit admission. Within each subplot, the patch#HSI model is compared to models utilizing the most important, two most important, 3 most important, or all clinical features available within the respective timeframe. The number of clinical features used is indicated in brackets. The feature importance ranking was derived via recursive feature elimination [130] using the clinical data model, starting from the full set of clinical features available within the given timeframe. Each box plot displays the distribution of the area under the receiver operating characteristic curve (AUROC) across 1000 bootstrap samples, with boxes spanning the interquartile range, whiskers showing the range excluding outliers, and solid and dashed lines marking the median and mean, respectively. Figure adapted from [306].

against the VIS for data available within 1 h, and the SOFA and APACHE II scores for data available within 10 h. As shown in Figure 7.11 and Figure 7.12, our patch#HSI + clinical data model outperformed all clinical scores and biomarkers for both sepsis diagnosis and mortality prediction, regardless of whether data were available within 1 h or 10 h of ICU admission.

7.4 Discussion and Conclusion

In this study, we tackled the urgent need for reliable biomarkers to identify septic patients and those at high risk of mortality. We are the first to show that automated, rapid, and non-invasive sepsis diagnosis and mortality prediction in ICU patients can be achieved through DL-based skin HSI. Drawing on what is, to our knowledge, the largest HSI patient cohort to date, we report the following key findings:

1. **HSI-based prediction:** DL models can accurately predict both sepsis and mortality from HSI data, with palm measurements outperforming the annular finger. HSI provides a clear advantage over conventional RGB imaging. In septic patients and non-survivors, palm skin shows significantly lower StO_2 and higher THI and TWI compared to non-septic patients and survivors.
2. **Generalizability to OOD populations:** Models trained on selectively chosen cohorts from previous studies fail to generalize to an ICU population, underscoring the importance of representative datasets when developing clinically applicable algorithms.
3. **Multimodal data fusion:** Integration of structured clinical data substantially boosts model performance, reaching AUROC scores of up to 0.94 for sepsis diagnosis and 0.83 for mortality prediction. Notably, combining HSI with just a few clinical features available at bedside already outperforms models based solely on clinical data.
4. **Clinical relevance:** Our multimodal models outperform established clinical biomarkers and scores.

The following sections provide a discussion of key strengths and limitations of our HSI-based approach to automated sepsis diagnosis and mortality prediction (Section 7.4.1), outline directions for future research (Section 7.4.2), and conclude with a summary of our findings (Section 7.4.3).

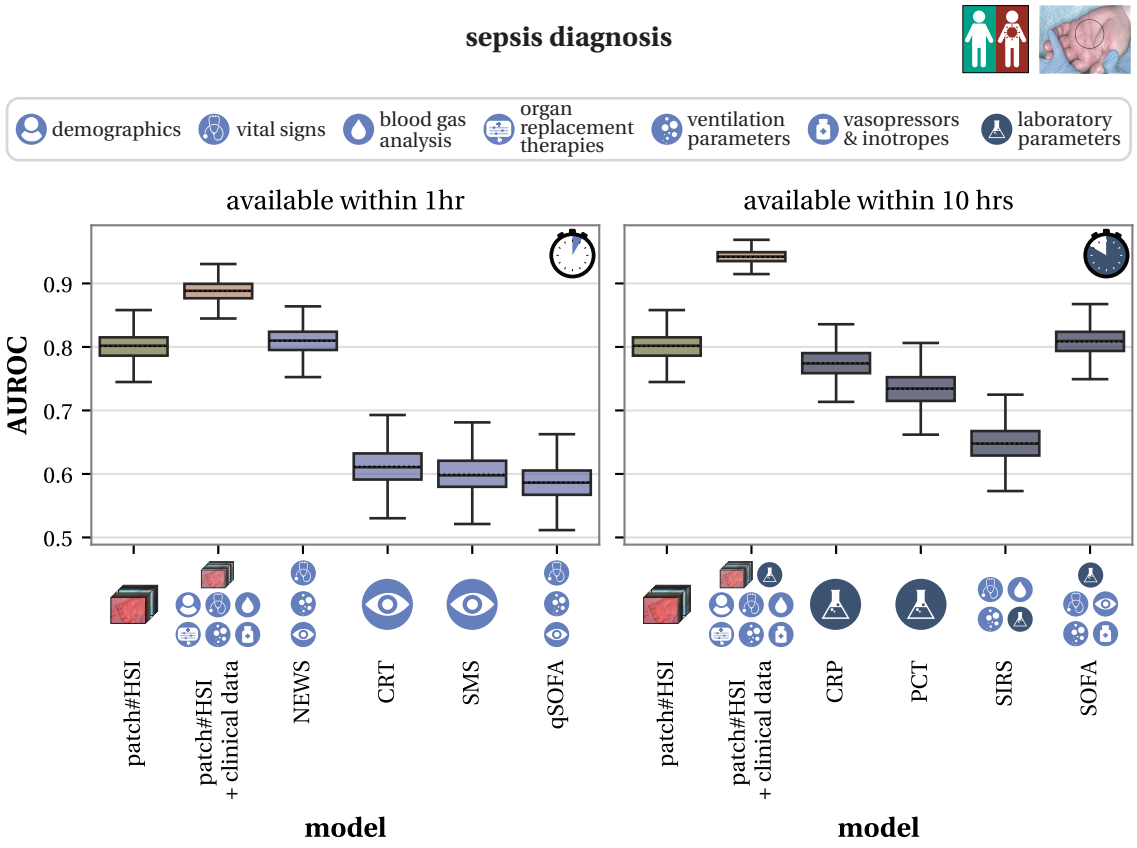


Figure 7.11: Comparison of our sepsis diagnosis models with established clinical biomarkers and scores. Performance is shown for models based on hyperspectral imaging (HSI) data of the palm skin (patch#HSI) and a combination of HSI data with all clinical data available within 1 h (left) and 10 h (right) from intensive care unit admission (patch#HSI + clinical data), compared against widely used clinical biomarkers and scores. For data available within 1 h, the comparison comprises national early warning score (NEWS), capillary refill time (CRT), skin mottling score (SMS) and quick Sequential Organ Failure Assessment (qSOFA) score. For data available within 10 h, it includes C-reactive protein (CRP), procalcitonin (PCT), Systemic Inflammatory Response Syndrome (SIRS) criteria and Sequential Organ Failure Assessment (SOFA) score. Each box plot displays the distribution of the area under the receiver operating characteristic curve (AUROC) across 1000 bootstrap samples, with boxes spanning the interquartile range, whiskers showing the range excluding outliers, and solid and dashed lines marking the median and mean, respectively. Figure adapted from [306].

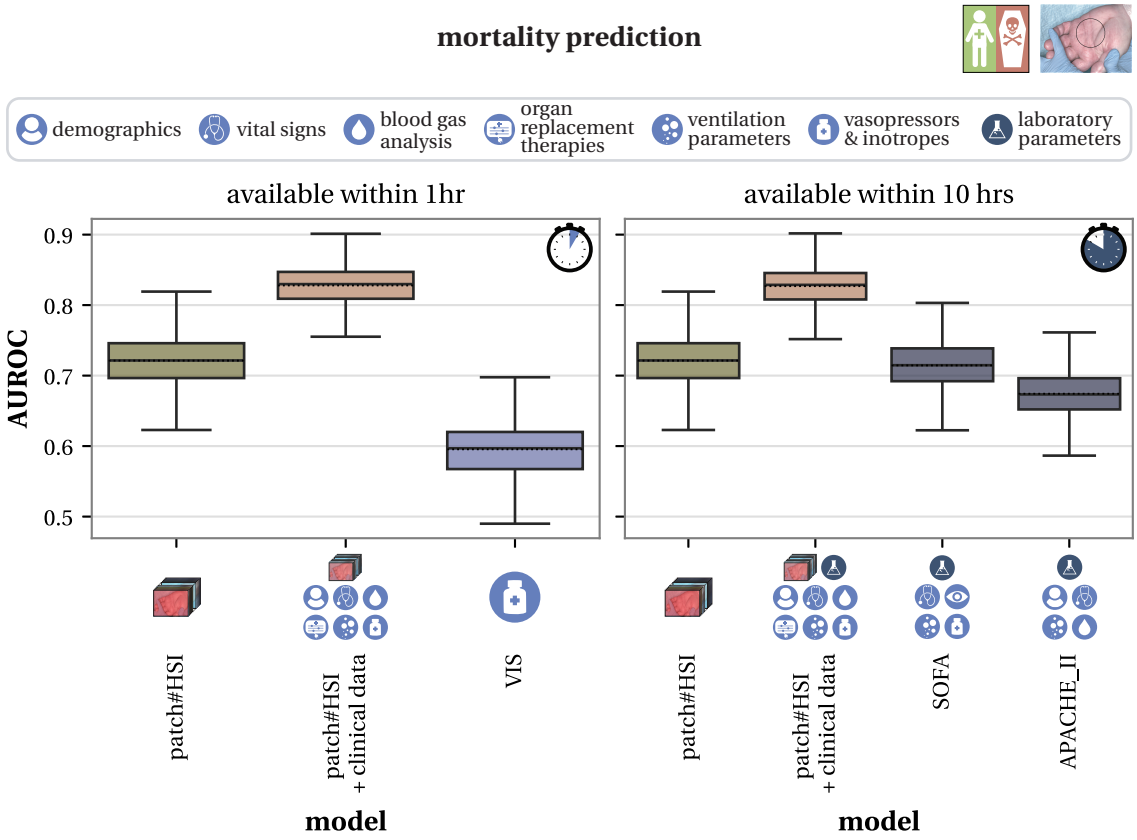


Figure 7.12: Comparison of our mortality prediction models with established clinical biomarkers and scores. Performance is shown for models based on hyperspectral imaging (HSI) data of the palm skin (patch#HSI) and a combination of HSI data with all clinical data available within 1 h (left) and 10 h (right) from intensive care unit admission (patch#HSI + clinical data), compared against widely used clinical biomarkers and scores. For data available within 1 h, the comparison comprises the vasoactive inotropic score (VIS). For data available within 10 h, it includes Sequential Organ Failure Assessment (SOFA) score and Acute Physiology and Chronic Health Evaluation (APACHE) II score. Each box plot displays the distribution of the area under the receiver operating characteristic curve (AUROC) across 1000 bootstrap samples, with boxes spanning the interquartile range, whiskers showing the range excluding outliers, and solid and dashed lines marking the median and mean, respectively. Figure adapted from [306].

7.4.1 Strengths and Limitations

We consider the main strengths of our HSI-based approach for sepsis diagnosis and mortality prediction to be its objectivity, non-invasiveness, cost-effectiveness, and speed, as predictions are derived from a single HSI cube acquired within few seconds directly at the bedside. These advantages suggest that our method could serve as a screening tool for all critically ill ICU patients, enabling timely and objective identification of those at high risk for sepsis and mortality, thereby supporting rapid initiation of further diagnostics and therapeutic interventions. Moreover, HSI systems are portable, allowing measurements to be performed in various clinical settings, including emergency departments and even ambulances.

We recognize that our classification models based solely on HSI data may not achieve sufficient accuracy to serve as standalone diagnostic or prognostic tools. Nevertheless, we see high potential for HSI as a pre-screening method to identify patients who would benefit from more time-consuming and costly assessments, such as laboratory tests and intensive monitoring. This approach is especially valuable in resource-limited settings, including LMICs, where roughly half of critical care interventions occur outside the ICU [32], and in scenarios requiring rapid decision-making, such as emergency care.

We demonstrated that integrating just a few clinical parameters readily available at the bedside can lead to substantial performance gains. For example, combining HSI with the administered noradrenaline dose increased the AUROC for sepsis prediction from 0.80 (95 % CI [0.76; 0.84]) to 0.87 (95 % CI [0.83; 0.90]). However, it is important to note that incorporating clinical data can introduce potential biases and limit generalizability. Treatment decisions, such as the administered noradrenaline dose, are influenced by local clinical guidelines, which can vary over time and between healthcare systems.

While our patch#HSI + clinical data models achieved excellent sepsis diagnosis and mortality prediction, and substantially outperformed widely used clinical biomarkers and scores, a key limitation of models requiring clinical data is the considerable effort needed for prospective data collection. We opted against the more convenient extraction of EHR data because many clinical features are either are captured at insufficient temporal resolution or not reliably recorded. For instance, vital signs and ventilation parameters in the EHR often fail to reflect the status of the patient at the time of HSI measurement. Since collecting clinical involves substantial effort, it is desirable to minimize the number of required clinical features. Importantly, our results show that, for both sepsis diagnosis and mortality prediction, the multimodal models outperform clinical data models when only a limited set of clinical parameters is available.

7.4.2 Future Work

A major limitation of our study is that all data were obtained from a single surgical ICU, where the majority of septic patients had an abdominal infection focus, with other infection sites being less represented. Additionally, different clinical sites manage critically ill patients differently. At some hospitals, patients are initially managed in emergency wards before ICU transfer, whereas at our site, critically ill patients – whether newly admitted from the emergency department or experiencing postoperative complications on general wards – are transferred directly to the ICU. Consequently, in our cohort, septic patients may be at earlier stages of the syndrome compared to other ICUs populations. Due to these differences in patient populations, external validation of our models is required to determine their generalizability across various ICU settings and clinical sites.

Considering the primary advantages of our HSI-based classification models, which allow for a non-invasive, mobile, rapid, and cost-effective diagnosis of sepsis and prediction of mortality, a promising avenue for future research is to evaluate their performance in resource-limited and time-critical settings, such as ambulances, emergency wards, and LMICs. It would also be valuable to investigate whether HSI can detect sepsis earlier in the course of the disease, potentially hours or even days before organ dysfunction occurs. Furthermore, considering that approximately 40 % of sepsis cases in 2017 involved children under 5 years old [296], including infants in the study cohort represents another important direction for future work.

While our observational study demonstrated high accuracy for automated HSI-based sepsis diagnosis and mortality prediction and highlighted potential applications, future interventional studies should evaluate the clinical impact of integrating such an automated alert system into routine care, comparing it to the standard of care and assessing its effects on key outcomes, such as mortality, morbidity, and hospital length of stay. To date, only a limited number of studies have systematically examined the clinical effectiveness of automated sepsis and mortality alert systems [383].

Although we view our single timepoint measurements as advantageous for enabling immediate diagnosis with minimal resource requirements, future studies could further expand the potential of HSI by incorporating longitudinal measurements. Such longitudinal measurements may enhance understanding of disease progression by revealing features linked to clinical improvement or deterioration.

Beyond its applications in disease diagnosis and prognosis, HSI holds promise for guiding novel therapeutic strategies by continuously monitoring tissue microcirculation, assessing treatment responses, and informing interventions. Recently, studies in animal models have explored the use of HSI to monitor the impact of vasopressor and fluid administration in hemorrhagic shock [83, 327]. Likewise, future research should

explore the potential of HSI to guide therapy in septic patients and other high-risk conditions, with the goal of improving clinical outcomes.

7.4.3 Conclusion

In this study, we addressed the urgent need for reliable sepsis diagnosis and mortality prediction. Based on a prospective study of over 480 patients – the largest HSI patient cohort to date – we present the first investigation of HSI for automated sepsis diagnosis and mortality prediction in the ICU. Our models utilizing single HSI cubes acquired within seconds achieved high predictive performance, which was further enhanced by combining HSI with a minimal set of clinical data. Our HSI + clinical data models outperformed established clinical scores and biomarkers. The non-invasive, mobile, rapid, and cost-effective nature of our HSI-based predictions makes them suitable for a wide range of clinical settings, including resource-limited environments (e.g., LMICs) and time-critical scenarios (e.g., ambulances, emergency wards). In addition to the demonstrated value in sepsis diagnosis and mortality prediction, monitoring microcirculation with HSI may deepen our understanding of disease progression and contribute to the development of novel therapeutic strategies. To support further research, we have publicly released our framework, together with our pretrained models[1] [312].

[1] https://github.com/IMSY-DKFZ/htc

Part V

Closing

8 Discussion and Conclusion

While the individual research questions are discussed and concluded within their respective chapters, this chapter provides a high-level summary of the main contributions of this thesis (Section 8.1). The work presented here has pioneered previously underexplored applications of SI in perioperative care and addressed key technical challenges hindering the clinical translation of ML-based SI analysis. Nevertheless, numerous challenges remain before such methods can be routinely integrated into clinical practice. To inspire future research, several of these open challenges are outlined in Section 8.2. The chapter concludes with a brief summary in Section 8.3.

8.1 Summary of Contributions

This thesis has made several contributions to advance the field of ML-based SI analysis in perioperative care. The answers to the core clinical and technical research questions of this thesis, along with related publications and international conference presentations, are summarized in Section 8.1.1. The broader impact of this thesis is highlighted in Section 8.1.2, including open-source contributions to foster future research in the field.

A more comprehensive overview of publications, international conference contributions, awards and patents is provided in Appendix A. This overview extends beyond the core outcomes of this thesis to include results from my supervision of fellow SI researchers as group lead, as well as from additional data science and clinical collaborations.

8.1.1 Answers to Research Questions

A structured overview of the individual contributions and their relation to the research questions is provided in Figure 8.1.

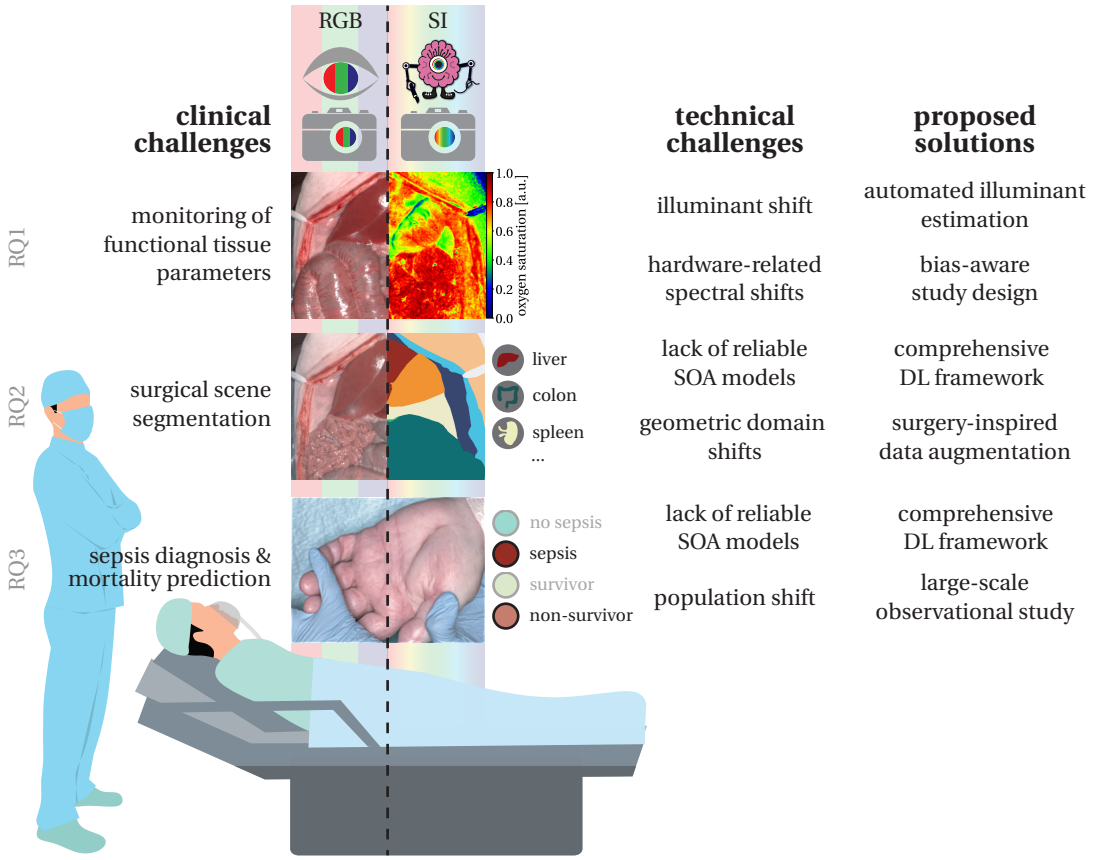


Figure 8.1: This thesis has advanced the field of spectral imaging (SI) analysis in perioperative care by addressing key clinical and technical challenges. Clinically, it tackled the need for robust functional parameter estimation (research question (RQ) 1), robust surgical scene segmentation (RQ2), and reliable sepsis diagnosis and mortality prediction (RQ3). Technically, we provided the first evidence of substantial inaccuracies in functional tissue parameter estimation caused by illuminant shifts during real-world surgeries and introduced a surgical workflow-compatible method for automated illuminant estimation. We presented the first systematic analysis of hardware-related spectral shifts and their impact on parameter estimates, and proposed strategies for bias-aware study design. Through a comprehensive deep learning (DL) framework, we addressed key knowledge gaps on the value of SI data for surgical scene segmentation and its optimal representation. We further demonstrated, for the first time, that real-world geometric domain shifts, such as situs occlusions, cause substantial performance drops which can be effectively mitigated by our proposed surgery-inspired data augmentation method. Through our DL framework and a large-scale observational study, we pioneered SI-based sepsis diagnosis and mortality prediction in intensive care unit (ICU) patients and identified strategies for optimal performance. Moreover, we showed that prior work relying on selectively chosen cohorts fails to generalize to a real-world ICU population.

RQ1: How can we achieve robust functional tissue parameter estimation with spectral imaging under real-world imaging conditions?

We were the first to demonstrate that illuminant shifts during real-world open surgeries can cause substantial inaccuracies in estimating tissue oxygenation. To address this challenge, we proposed an automated illuminant estimation method that can be seamlessly integrated into the surgical workflow. Our approach outperforms state-of-the-art color constancy methods and achieves accuracy close to the ideal scenario in which the illuminant spectrum is perfectly known. These findings were presented orally to an international expert audience at the IPCAI 2020, with a full article published in the conference proceedings [24]. Furthermore, the proposed automated illuminant estimation method was included in two filed patents [225, 226].

We conducted the first systematic analysis of hardware-related variability in HSI measurements and found that the widely used medical-grade TIVITA® cameras (Diaspective Vision GmbH, Am Salzhaff, Germany) are affected by spectral shifts across device generations and instances, sensor temperatures, and calibration schemes. In particular, rises in sensor temperature over minutes to hours of measuring led to substantial drifts in functional tissue parameter index values. Based on these findings, we proposed recommendations for mitigating hardware-induced variability in HSI study design. An abstract summarizing parts of this work was accepted at IEEE's 13th WHISPERS in 2023, where I delivered an oral presentation to an international audience of SI researchers [310].

Both studies represent important milestones toward reliable functional tissue parameter estimation under real-world imaging conditions, paving the way for safe integration of SI-based functional imaging into clinical practice to support informed intraoperative decision-making.

RQ2: How can we achieve robust surgical scene segmentation under geometric domain shifts?

Using the largest semantically annotated intraoperative SI dataset to date, we conducted the first systematic analysis of the optimal spectral and spatial granularity for automated surgical scene segmentation. Based on developing an extensive DL framework encompassing segmentation at different spatial granularities (pixels, superpixels, patches, and full images), as well as multiple imaging modalities (RGB, HSI, and derived functional tissue parameters), we demonstrated that HSI data outperforms both RGB and processed HSI data, particularly at small spatial granularities. At the same time, image-based segmentation consistently outperformed smaller granularities, independent of the number of training subjects. Notably, segmentation performance based on full HSI cubes matched the accuracy of annotations by a second medical expert. This work was published in the journal Medical Image Analysis in 2022 [308], and following submission of a long abstract, I orally presented the research to an international expert audience at the IPCAI 2022 [307].

Given the well-known vulnerability of DL models to domain shifts, we addressed the lack of research on the generalizability of surgical scene segmentation models under such conditions. We were the first to show that geometric domain shifts – frequently arising in real surgeries due to variations in procedures or situs occlusions, yet typically overlooked during model development – cause substantial performance degradation. To mitigate these effects, we introduced a surgery-inspired data augmentation technique that is computationally efficient, model-independent, and capable of restoring in-distribution performance. This work was accepted at MICCAI 2023, where I presented it to an international audience of medical image computing experts. The full article was subsequently published in the conference proceedings [314]. The contribution was recognized with the MICCAI Student-Author Registration (STAR) award and the MICCAI Young Scientist Award.

Our work represents a key advance toward robust surgical scene segmentation under real-world domain shifts, forming a critical foundation for surgical data science applications that enable intraoperative decision support and context-aware assistance.

RQ3: Can we reliably diagnose sepsis and predict mortality in an intensive care unit population using skin spectral images?

Based on a large-scale observational study of over 480 patients – representing the so far largest HSI patient cohort – we were the first to demonstrate that DL-based analysis of skin HSI enables rapid, mobile and non-invasive sepsis diagnosis and mortality prediction in ICU patients. Through an extensive DL framework encompassing classification at different spatial granularities (median spectra and patches) and imaging modalities (RGB, HSI, and derived functional tissue parameters), as well as multimodal fusion of HSI with clinical data, we showed that HSI is superior to conventional RGB imaging, with the palm identified as the optimal measurement site. The predictive performance was further enhanced when HSI data were combined with a minimal set of clinical parameters, outperforming established clinical biomarkers and scoring systems. Moreover, we showed that previous approaches, which relied on selectively chosen cohorts and were therefore susceptible to shortcut learning, fail to generalize to a real-world ICU population. This work was published in Science Advances in 2025 [306] and an abstract was accepted for oral presentation at the 5th Conference on Clinical Translation of Medical Image Computing & Computer Assisted Intervention (CLINICCAI, part of the MICCAI) in 2025 [84].

Our work tackles the critical shortage of reliable biomarkers for timely sepsis diagnosis and mortality prediction in ICU patients – a challenge of high socioeconomic importance – and lays the groundwork for future research on the potential of HSI-based sepsis diagnosis and mortality prediction in other clinical settings (e.g., LMICs) and for therapy guidance.

8.1.2 Broader Impact of this Thesis

Maier-Hein et al. identified the lack of data, and in particular representative annotated data, as a major obstacle in the field of surgical data science [221]. Medical SI is no exception: Although the increasing adoption of medically certified SI devices has led to the creation of larger datasets – the biggest ones now encompassing several tens to hundreds of patients [308, 30, 315, 306] – only few datasets in the field have been made publicly available [73]. This scarcity represents a substantial barrier to the clinical translation of ML-based SI analysis, as developing and validating robust ML models requires large, diverse datasets to capture the full variability of the underlying data distribution and to evaluate and improve generalizability. While our ICU dataset cannot be shared due to the lack of patient consent, we have released the majority of our open surgery datasets along with their annotations [328, 315]. These include a total of 10 818 images from 89 subjects across two species and different perfusion states, establishing them as the largest publicly available medical SI resource to date in terms of number of images.

In addition to releasing open datasets, we contributed to established open-source frameworks for medical image analysis such as MONAI, where we integrated our implementation of the NSD [56], and the Kornia library [287], where we integrated our Organ Transplantation augmentation. Furthermore, we advanced the field of medical SI analysis by open-sourcing all our code as a comprehensive framework[1] [312]. The framework provides modules for data loading, preprocessing, augmentation, visualization, and the training and validation of DL models specifically tailored to SI data, and includes pretrained models. It enables benchmarking against our models and allows researchers to easily train or fine-tune segmentation and classification models for other applications. Since models trained on SI data often suffer from data-loading bottlenecks due to large image sizes, leading to delayed training runs, low GPU utilization, and long inference times, we developed several strategies to optimize data-loading efficiency [313]. These strategies are transferable to any application where data loading is a bottleneck, making them broadly useful even beyond medical SI.

8.2 Open Challenges

Beyond pioneering underexplored applications of SI in perioperative care, the overarching goal of this thesis was to advance the clinical translation of SI analysis by enhancing the robustness of ML models to domain shifts between training and real-world application data. However, several challenges related to SI hardware (Section 8.2.1) and

[1] https://github.com/IMSY-DKFZ/htc

ML-based SI analysis (Section 8.2.2) still impede a widespread clinical adoption of perioperative SI. This section highlights these challenges with the aim of guiding future research.

8.2.1 Challenges Related to Spectral Imaging Hardware

The clinical translation of SI remains challenged by limitations in the available hardware:

Trade-off Between Spectral, Spatial, and Temporal Resolution SI devices typically involve trade-offs between spectral, spatial, and temporal resolution [65]. For example, the HSI system used in this thesis captures a cube with 100 spectral bands at spatial dimensions of 640 px×480 px, but the acquisition of a single image takes approximately 7 s. In contrast, our MSI system achieves a frame rate of 25 Hz, but comes with a reduced number of only 16 spectral bands and lower spatial dimensions of 512 px × 272 px. These limitations introduce several challenges:

The slow acquisition speed of our HSI device restricts its use to stationary objects in a mounted setup, rather than handheld use, and prevents real-time feedback on rapidly evolving scenes. In time-critical clinical settings where rapid decision-making is essential (e.g., during intraoperative tissue clamping [21]), such delays can pose a substantial barrier. Additionally, object movements – such as tissue movements due to respiration, heartbeat, surgical manipulation or patient movements – inevitably introduce motion artifacts (cf. Figure 5.8 for an example).

Compared to modern RGB cameras, which routinely provide videos at spatial dimensions of 3840 px × 2160 px [221], both our HSI and MSI devices provide only limited pixel resolution. Particularly for our HSI systems that need to be operated at a fixed imaging distance with a fixed focal length, the resulting spatial resolution can be a limiting factor. For example, the HSI system used in this thesis for automated surgical scene segmentation provides a spatial resolution of only about 0.5 mm, which limits the visibility of fine structures. For instance, in our surgical scene segmentation datasets, the class major vein was often represented by only a few pixels, while smaller structures such as nerves and lymph nodes were not distinguishable at all. Both the limited spatial resolution and motion artifacts resulting from low temporal resolution complicate the annotation process, as poorly defined structure boundaries can lead to imprecise annotations.

Variability and Limited Availability of Spectral Imaging Hardware The trade-off between temporal, spatial and spectral resolution has contributed to existing SI devices being

highly diverse regarding number of spectral channels, spectral range, bandwidths, spatial resolution, and acquisition speed [205]. Furthermore, a range of different light sources, including halogen, LED, or tunable lights, are used across studies [21, 218]. Most devices are custom research prototypes tailored to a single specific application, and oftentimes equipped with additional optical components, such as laparoscopes, endoscopes or microscopes [377, 25, 272]. The high diversity across devices and setups poses a major challenge: algorithms developed on data from one device may fail to generalize to another device. Unless the community converges on a few standardized setups, developing robust algorithms that are invariant to hardware differences will be essential.

The recent introduction of the first commercial, medically certified SI systems offers potential for convergence toward fewer, more standardized setups. Medical certification reduces bureaucratic hurdles in clinical studies and facilitates clinical translation by ensuring regulatory compliance for safe integration into workflows. However, the high cost of these systems remains a substantial barrier – particularly in LMICs, where limited access to advanced imaging technology further exacerbates global disparities in perioperative care [299]. Moreover, the SI hardware market is still small, with only a handful of manufacturers producing devices in limited quantities, further impeding widespread clinical adoption.

Reliability of Spectral Imaging Devices We showed that medically certified HSI systems are subject to hardware-related spectral variability, which can cause substantial drifts in functional tissue parameter estimates (cf. Chapter 4). In addition to these inaccuracies, the HSI devices used in this thesis exhibited data acquisition and storage failures in about 1 % of cases, resulting in corrupted images or metadata. Such errors can delay clinical workflows when measurements must be repeated and, if unnoticed, may lead to the irreversible loss of valuable data.

To mitigate hardware-related measurement inaccuracies, we implemented several measures in our studies, such as frequent device calibration and extending the interval between acquisitions to limit sensor heating. However, these remain only temporary workarounds. Long-term solutions require manufacturers to enhance device reliability and reproducibility, for example by improving sensor cooling, correcting systematic errors, and developing algorithms to compensate for hardware-related shifts. In parallel, the field urgently needs systematic investigations of additional bias sources in SI measurements, along with community-wide standardized recording protocols (e.g., timing and methods for calibration, acquisition of imaging metadata such as sensor temepratures) and quality control procedures (e.g., regular phantom measurements) to ensure data accuracy and comparability.

In summary, the clinical translation of SI remains constrained by limitations in hardware reliability, availability, and standardization, as well as the need to balance spatial and temporal resolution against spectral information content. Nonetheless, the field has made substantial progress in recent years, with key manufacturers in the nonmedical domain (e.g., imec, Leuven, Belgium; HAIC, Hanover, Germany) and researchers developing more compact, real-time SI systems and scaling production toward high-volume, cost-effective availability [35, 387, 75, 148, 277, 341]. Future research should support these efforts by exploring emerging technologies and software advances. Promising directions include ML-based methods to enhance spatial and spectral resolution [73, 148, 235, 218], band-selection strategies that improve temporal resolution by discarding uninformative channels [22, 381], and tunable-band devices that allow adaptive spectral sampling [272, 218]. In light of these advances, future generations of SI hardware can be expected to overcome many of the current shortcomings and move closer to widespread clinical adoption.

8.2.2 Challenges Related to Machine Learning-Based Spectral Image Analysis

To date, the potential of ML in medicine has only been partially realized, with relatively few tools developed in academia successfully implemented in clinical practice [266]. The translation of ML models into real-world healthcare is hindered by multiple factors: some are inherent to the broader medical ML community, such as limitations in the generalizability and trustworthiness, while others are specific to SI, particularly the challenges of data availability and rigorous validation.

Generalizability and Trustworthiness of Machine Learning Models This thesis has focused on improving the robustness of ML models under domain shifts between training and real-world application data – one of the key barriers to clinical translation. We effectively addressed shifts in illumination and surgical scene geometry and investigated the impact of hardware and population shifts. However, important domain shifts remain unexplored. For example, population distributions can differ substantially across hospitals and countries [57], and in the case of SI, may be further compounded by variations in devices, acquisition, and annotation procedures. While such technical shifts should be minimized through community-wide efforts to establish standardized hardware, acquisition, and annotation protocols (cf. Section 8.2.1), residual shifts must be systematically investigated and addressed within the ML community (cf. Section 6.4.2 for a detailed discussion of further potential domain shifts in the context of surgical scene segmentation). A particular challenge lies in the inherently dynamic nature of healthcare: evolving clinical practices and shifting patient populations con-

tinuously alter data distributions, requiring strategies for drift detection and regular model updates to ensure sustained performance [174].

ML models are often considered as "black boxes", as it is not directly accessible how they arrive at their predictions. This property conflicts with the demand for accountability and result transparency in medical scenarios [334], and raises the need for models that are not only performing well, but also trustworthy and explainable [142]. In the context of SI, this is particularly important, as the rich spectral information captured by SI devices is not immediately interpretable by humans. Explainable ML models hold the potential to provide novel insights into the underlying biological processes. However, only few studies have yet investigated the explainability of ML models for spectral data [71], and we are not aware of a single study investigating explainable ML for medical SI analysis. Despite substantial progress in recent years, developing explainable ML remains a major challenge across the entire ML community. Many existing approaches lack end-to-end interpretability and computational efficiency, and trust in their explanations is limited by inconsistencies between methods and the lack of objective validation [334]. Future research should aim to address these challenges by developing explainable ML models that provide consistent, reliable, and interpretable explanations for SI data.

Further trust in ML prediction could be strengthened through the quantification of model uncertainty. Assessing the confidence of model predictions could enable more principled decision-making by allowing uncertain outputs to be discarded, thereby reducing the risk of potentially harmful misclassifications [187]. However, in the broader field of ML, uncertainty quantification remains underexplored, with no consensus on optimal methods and persistent shortcomings in the evaluation of methods on real-world data [211, 111]. Future research should explore uncertainty quantification in medical SI analysis to establish reliable measures of model confidence.

Sparsity of Annotated Data Large, high-quality datasets are essential for developing and validating robust ML models, as they allow capturing the full diversity of the underlying data distribution and improve generalizability [15]. However, the availability of such datasets remains a major challenge in medical SI. Compared to widely used datasets from general computer vision, such as ImageNet (1 281 167 training examples) [78], annotated SI datasets are comparably small. While the size of SI datasets has grown over the past years – for instance, surgical scene segmentation datasets have increased from several tens to several hundreds of patients (cf. Table 5.2) – they remain orders of magnitude smaller than standard computer vision benchmarks. This is partly due to the substantial complexity of conducting clinical studies with devices that are not yet standard of care, have limited availability, and are costly (cf. Section 8.2.1). Additionally, generating extensive, high-quality annotations for SI data poses its own challenges.

In fact, our continuously growing intraoperative SI database currently comprises 46 831 images from 388 subjects across 3 species (cf. [315]), yet only about 6 % of these images are semantically annotated. Manual annotation is time-consuming, taking approximately 30 min per image for our semantic surgical scene annotation. Future efforts could leverage active learning strategies to prioritize the most informative images for annotation, thereby maximizing the value of each annotated sample. Such approaches have already proven successful in surgical workflow analysis [44]. Moreover, given the abundance of unlabeled SI data – with our entire perioperative SI database even encompassing 69 747 images from 1000 subjects – semi-supervised and unsupervised learning methods represent a promising avenue to exploit this data and potentially improve model performance [281].

Another challenge arises from uncertainties in manual annotations. For example, in our porcine surgical scene segmentation dataset, 2 % of the scene on average could not be confidently assigned a class label, and in our sepsis diagnosis dataset, 14 % of patients had an ambiguous sepsis status. While some of these uncertainties could potentially be mitigated by improving the annotation process (cf. Section 5.4.2 for suggestions in the context of surgical scene segmentation), a large portion reflects the inherent difficulty of clinical practice: even experienced medical experts may struggle to differentiate certain tissue types or determine sepsis status in critically ill patients. In our analyses, instances with uncertain labels were excluded. Although this approach avoids introducing potentially incorrect labels, it reduces the amount of usable data and may bias ML models, as the remaining dataset may not fully represent the underlying population. Although data with uncertain labels could, in principle, be incorporated into training using approaches such as soft labels, evaluating algorithm performance on samples with ambiguous labels remains an open challenge across the entire ML community.

Lack of External Validation External Validation is essential for the clinical translation of ML models, as it demonstrates reliable generalization to data that may differ in population, device or other factors [266]. However, due to the considerable effort required to acquire and annotate SI data, we were unable to perform multi-center studies. To our knowledge, only a single study to date has conducted external validation of an SI-based ML model on data from a different clinic [30].

External validation of ML models could be facilitated through the availability of open-source datasets. However, in medical SI, publicly available datasets are scarce, often small, and frequently lack high-quality data calibration [73]. To advance the field, we have released large portions of our data [328, 23, 315]. Nevertheless, to reduce inconsistencies between datasets and enable meaningful model comparisons, the establishment of community-wide benchmarks, along with standardized acquisition, annotation, and evaluation protocols, is essential.

Lack of Clinical Value Demonstration Metrics commonly used during model development and validation often fail to capture the ultimate clinical value of a model, such as improvements in patient outcomes or cost-effectiveness. To date, only a few randomized controlled trials have evaluated medical ML algorithms (e.g., [359, 371]), and, to the best of our knowledge, no interventional studies have been conducted using SI. Future research should therefore focus on demonstrating the clinical benefit of ML-based SI analysis.

8.3 Conclusion

This thesis represents a pioneering step toward the clinical translation of ML-based SI analysis in perioperative care. We demonstrated the potential of SI in underexplored applications, including semantic scene segmentation in open surgeries, as well as sepsis diagnosis and mortality prediction among critically ill ICU patients. Importantly, we investigated and addressed critical domain shifts previously overlooked by the community, including variations in illumination, hardware, surgical scene geometry and population between training and real-world application data.

In addition, we contributed to the advancement of the field by releasing open-source data, annotations and pretrained models, encompassing both RGB and SI data [328, 315]. These could serve as a benchmark for future research. Furthermore, we developed and publicly released a comprehensive framework for efficient SI data management and DL model development [312].

Beyond the perioperative applications studied in this thesis, many of the concepts, tools, and datasets introduced are broadly applicable: For example, our framework includes strategies to improve training efficiency on high-dimensional HSI data, transferable to other domains facing data-loading bottlenecks, and our approaches to investigate and mitigate domain shifts may inspire similar efforts in SI and other medical imaging fields.

We acknowledge that, despite representing important first steps, several challenges remain for the clinical translation of ML-based SI analysis in perioperative care, including shortcomings of spectral imaging hardware and limitations of the ML models. Future research should aim to address these challenges by developing more standardized and reliable hardware, advancing generalizable, explainable, and uncertainty-aware ML methods for SI analysis, and establishing community-wide benchmarks and standardized protocols through collaborative efforts. In the light of recent promising advancements (e.g., in imaging technologies and data availability), it is expected that the clinical translation of SI will continue to progress, ultimately integrating ML-based SI analysis as a routine component of perioperative care.

Own Contributions, Publications, Conferences, Awards and Patents

A

A.1 Own Contributions

This thesis was carried out in the Division of Intelligent Medical Systems (IMSY), formerly Computer Assisted Medical Interventions (CAMI), at the German Cancer Research Center (DKFZ) under the supervision of Prof. Dr. Lena Maier-Hein, and supported by the Helmholtz Information & Data Science School for Health (HIDSS4Health). Close collaboration with IMSY group members and clinical partners was integral to the work. To distinguish own contributions from team efforts, I highlight my key contributions to each research question addressed in this thesis in the following.

RQ1: How can we achieve robust functional tissue parameter estimation with spectral imaging under real-world imaging conditions?

Regarding the research on robust oxygenation estimation under illuminant shifts (Chapter 3), I designed the experiments, acquired and curated the data, conducted the data analyses, and designed, developed and validated the illuminant estimation methodology. I also created the manuscript figures and contributed to writing the resulting publication [24]. In addition, I contributed to the filing of two patents covering the automated illuminant estimation method by conducting the prior art search and providing technical input during the patent drafting process [225, 226].

Regarding the research on robust functional tissue parameter estimation under hardware variations (Chapter 4), I planned and conducted the experiments, acquired, annotated and curated the data, performed the data analysis, and developed the recommendations for hardware bias-aware study design. I also generated the figures and wrote the corresponding abstract for conference presentation [310].

RQ2: How can we achieve robust surgical scene segmentation under geometric domain shifts?

Regarding the research on optimal spectral and spatial granularity for robust surgical scene segmentation (Chapter 5), I designed the experiments, curated the data, and

supervised the annotation process. I designed, developed and validated the pixel- and superpixel-based models, conducted the training size experiments, generated figures, and wrote the corresponding journal manuscript [308] and long abstract [307].

Regarding the research on robust surgical scene segmentation under geometric domain shifts (Chapter 6), I designed the experiments, curated the OOD data and supervised its annotation. I designed, developed and validated the Organ Transplantation augmentation. I conducted the experiments comparing the effects of geometric domain shifts across different spatial granularities and input modalities and co-supervised the bachelor thesis of Alessandro Motta including the local neighborhood experiment. I generated figures, and wrote the manuscript published in the MICCAI proceedings [314], as well as the extended version on arXiv [309].

RQ3: Can we reliably diagnose sepsis and predict mortality in an intensive care unit population using skin spectral images?

Regarding our research on automated sepsis diagnosis and mortality prediction, I co-designed the clinical study in ICU patients, supervised data acquisition and annotation, and curated the dataset. I performed the data analysis, including statistical testing, and designed, developed and validated the machine learning models. I generated the figures and wrote the manuscript of the resulting publication [306].

A.2 Publications

My research has led to several publications in peer-reviewed journals and conference proceedings. In addition to the 6 first-author publications directly related to this thesis, my role as group lead supervising a team of SI researchers, together with close collaborations with further data scientists and clinical partners, has enabled me to contribute to a wider range of research projects beyond visceral surgery and anesthesiology, extending into gastroenterology, urology, oncology and hematology. As a result, I have authored further publications in SI and medical image computing, both as last author and as co-author. In the following, these publications are categorized according to my role as first, last, or co-author.

First-Author Publications

1. **Silvia Seidlitz**, Katharina Hölzl, Ayca von Garrel, Jan Sellner, Stephan Katzenschlager, Tobias Hölle, Dania Fischer, Maik von der Forst, Felix C. F. Schmitt, Alexander Studier-Fischer, Markus A. Weigand, Lena Maier-Hein, and Maximilian Dietrich. “AI-powered skin spectral imaging enables instant sepsis diagnosis and outcome prediction in critically ill patients”. In: *Science Advances* 11.29 (2025), eadw1968. DOI: 10.1126/sciadv.adw1968

2. **Silvia Seidlitz**, Jan Sellner, Alexander Studier-Fischer, Alessandro Motta, Berkin Özdemir, Beat P. Müller-Stich, Felix Nickel, and Lena Maier-Hein. "Handling Geometric Domain Shifts in Semantic Segmentation of Surgical RGB and Hyperspectral Images". In: *arXiv preprint arXiv:2408.15373* (2024). DOI: 10.48550/arXiv.2408.15373

3. Jan Sellner, **Silvia Seidlitz**, Alexander Studier-Fischer, Alessandro Motta, Berkin Özdemir, Beat Peter Müller-Stich, Felix Nickel, and Lena Maier-Hein. "Semantic Segmentation of Surgical Hyperspectral Images Under Geometric Domain Shifts". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Cham: Springer Nature Switzerland, 2023, pp. 618–627. ISBN: 978-3-031-43996-4. DOI: 10.1007/978-3-031-43996-4_59

4. **Silvia Seidlitz**, Jan Sellner, Jan Odenthal, Berkin Özdemir, Alexander Studier-Fischer, Samuel Knödler, Leonardo Ayala, Tim J. Adler, Hannes G. Kenngott, Minu Tizabi, Martin Wagner, Felix Nickel, Beat P. Müller-Stich, and Lena Maier-Hein. "Robust deep learning-based semantic organ segmentation in hyperspectral images". In: *Medical Image Analysis* 80 (Aug. 2022), p. 102488. ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102488

5. Maximilian Dietrich, **Silvia Seidlitz**, Nicholas Schreck, Manuel Wiesenfarth, Patrick Godau, Minu Tizabi, Jan Sellner, Sebastian Marx, Samuel Knödler, Michael M. Allers, Leonardo Ayala, Karsten Schmidt, Thorsten Brenner, Alexander Studier-Fischer, Felix Nickel, Beat P. Müller-Stich, Annette Kopp-Schneider, Markus A. Weigand, and Lena Maier-Hein. "Machine learning-based analysis of hyperspectral images for automated sepsis diagnosis". In: *arXiv preprint arXiv:2106.08445* (2021). DOI: 10.48550/arXiv.2106.08445

6. Leonardo Ayala, **Silvia Seidlitz**, Anant Vemuri, Sebastian J Wirkert, Thomas Kirchner, Tim J Adler, Christina Engels, Dogu Teber, and Lena Maier-Hein. "Light source calibration for multispectral imaging in surgery". In: *Int J Comput Assist Radiol Surg* 15.7 (June 2020), pp. 1117–1125. DOI: 10.1007/s11548-020-02195-y

Last-Author Publications

1. Alexander Baumann, Leonardo Ayala, Alexander Studier-Fischer, Jan Sellner, Berkin Özdemir, Karl-Friedrich Kowalewski, Slobodan Ilic, **Silvia Seidlitz**, and Lena Maier-Hein. "Neural Illumination Calibration for Surgical Workflow-Optimized Spectral Imaging". In: *Under review at the International Journal of Computer Assisted Radiology and Surgery (IJCARS) for the IJCARS-MICCAI 2024 Special Issue* (2025)

2. Alexander Baumann, Leonardo Ayala, Alexander Studier-Fischer, Jan Sellner, Berkin Özdemir, Karl-Friedrich Kowalewski, Slobodan Ilic, **Silvia Seidlitz**, and Lena Maier-Hein. "Deep intra-operative illumination calibration of hyperspectral cameras". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Cham: Springer Nature Switzerland, 2024. DOI: 10.1007/978-3-031-72089-5_12

3. Ahmad Bin Qasim, Alessandro Motta, Alexander Studier-Fischer, Jan Sellner, Leonardo Ayala, Marco Hübner, Marc Bressan, Berkin Özdemir, Karl Friedrich Kowalewski, Felix Nickel, **Silvia Seidlitz**, and Lena Maier-Hein. "Test-time augmentation with synthetic data addresses distribution shifts in spectral imaging". In: *International Journal of Computer Assisted Radiology and Surgery* (Mar. 14, 2024). ISSN: 1861-6429. DOI: 10.1007/s11548-024-03085-3

Co-Author Publications

1. Jan Sellner, Alexander Studier-Fischer, Ahmad Bin Qasim, **Silvia Seidlitz**, Nicholas Schreck, Minu Tizabi, Manuel Wiesenfarth, Annette Kopp-Schneider, Samuel Knödler, Caelan Max Haney, Gabriel Salg, Berkin Özdemir, Maximilian Dietrich, Maurice Stephan Michel, Felix Nickel, Karl-Friedrich Kowalewski, and Lena Maier-Hein. "Xeno-learning: knowledge transfer across species in deep learning-based spectral image analysis". In: *Accepted at Nature Biomedical Engineering, arXiv preprint arXiv:2410.19789* (2024). DOI: 10.48550/arXiv.2410.19789

2. Gabriel Szydlo Shein, Elisa Bannone, **Silvia Seidlitz**, Mohamed Hassouna, Luca Baratelli, Arturo Pardo, Frédéric Triponez, Manish Chand, Sylvain Gioux, Lena Maier-Hein, and Michele Diana. "Surgical Optomics: A new science towards surgical precision". In: *Accepted at npj Gut and Liver* (2025)

3. Luisa Egen, Moritz Hommel, Caelan Max Haney, Berkin Özdemir, Samuel Knoedler, Jan Sellner, **Silvia Seidlitz**, Maximilian Dietrich, Gabriel Alexander Salg, Felix Nickel, Lena Maier-Hein, Maurice Stephan Michel, Alexander Studier-Fischer, and Karl-Friedrich Kowalewski. "Hyperspectral Imaging Accurately Detects Renal Malperfusion Due to High Intrarenal Pressure". In: *European Urology Open Science* 78 (2025), pp. 16–27. ISSN: 2666-1683. DOI: 10.1016/j.euros.2025.06.007

4. Viet Tran Ba, Marco Hübner, Ahmad Bin Qasim, Maike Rees, Jan Sellner, **Silvia Seidlitz**, Evangelia Christodoulou, Berkin Özdemir, Alexander Studier-Fischer, Felix Nickel, Leonardo Ayala, and Lena Maier-Hein. "Semantic hyperspectral image synthesis for cross-modality knowledge transfer in surgical data science".

In: *International Journal of Computer Assisted Radiology and Surgery* 20.6 (June 2025), pp. 1205–1213. DOI: 10.1007/s11548-025-03364-7

5. F. Nickel, A. Studier-Fischer, B. Özdemir, J. Odenthal, L.R. Müller, S. Knoedler, K.F. Kowalewski, I. Camplisson, M.M. Allers, M. Dietrich, K. Schmidt, G.A. Salg, H.G. Kenngott, A.T. Billeter, I. Gockel, C. Sagiv, O.E. Hadar, J. Gildenblat, L. Ayala, S. Seidlitz, L. Maier-Hein, and B.P. Müller-Stich. "Optimization of anastomotic technique and gastric conduit perfusion with hyperspectral imaging and machine learning in an experimental model for minimally invasive esophagectomy". In: *European Journal of Surgical Oncology* 51.1 (2025), p. 106908. ISSN: 0748-7983. DOI: 10.1016/j.ejso.2023.04.007

6. Alexander Baumann, Leonardo Ayala, **Silvia Seidlitz**, Jan Sellner, Alexander Studier-Fischer, Berkin Özdemir, Lena Maier-Hein, and Slobodan Ilic. "CARL: Camera-Agnostic Representation Learning for Spectral Image Analysis". In: *Submitted to the 14th International Conference on Learning Representations (ICLR), arXiv preprint arXiv:2504.19223* (2025). DOI: 10.48550/arXiv.2504.19223

7. Marco Hübner, Ahmad Bin Qasim, Alexander Studier-Fischer, Viet Tran Ba, Maike Rees, Jan-Hinrich Nölke, **Silvia Seidlitz**, Jan Sellner, Janne Heinecke, Jule Brandt, Berkin Özdemir, Kris Dreher, Alexander Seitel, Felix Nickel, Caelan Max Haney, Karl-Friedrich Kowalewski, Leonardo Ayala, and Lena Maier-Hein. "Learning to Simulate Realistic Human Diffuse Reflectance Spectra". In: *Under review at the Journal of Biomedical Optics (JBO)* (2025)

8. Laura Simons, Lina Alasfar, Muath Qadoura, Franziska Sunderer, Felix Korell, Ignatios Ikonomidis, Maximilian Dietrich, **Silvia Seidlitz**, Hans Vink, Lena Maier-Hein, Michael Schmitt, Richard F. Schlenk, Carsten Müller-Tidow, Peter Dreger, and Thomas Luft. "Comprehensive assessment of endothelial dysfunction before cellular therapy: EASIX, local imaging and systemic biomarkers". In: *Accepted at Blood Vessels, Thrombosis & Hemostasis* (2025)

9. Alexander Studier-Fischer, Marc Bressan, Ahmad bin Qasim, Berkin Özdemir, Jan Sellner, **Silvia Seidlitz**, Caelán Haney, Luisa Egen, Maurice Michel, Maximilian Dietrich, Gabriel Alexander Salg, Franck Billmann, Henrik Nienhüser, Thilo Hackert, Beat Müller-Stich, Lena Maier-Hein, Felix Nickel, and Karl-Friedrich Kowalewski. "Spectral characterization of intraoperative renal perfusion using hyperspectral imaging and artificial intelligence". In: *Scientific Reports* 14.1 (July 27, 2024), p. 17262. ISSN: 2045-2322. DOI: 10.1038/s41598-024-68280-3

10. Alexander Studier-Fischer, Berkin Özdemir, Maike Rees, Leonardo Ayala, **Silvia Seidlitz**, Jan Sellner, Karl-Friedrich Kowalewski, Caelán Max Haney, Jan Odentahl, Samuel Knödler, Maximilian Dietrich, Daniel Gruneberg, Thorsten Brenner, Karsten Schmidt, Felix Carl Fabian Schmitt, Markus A. Weigand, Gabriel Alexander Salg, Anna Dupree, Henrik Nienhüser, Arianeb Mehrabi, Thilo Hackert, Beat

Müller-Stich, Lena Maier-Hein, and Felix Nickel. "Crystalloid volume versus catecholamines for management of hemorrhagic shock during esophagectomy – assessment of microcirculatory tissue oxygenation of the gastric conduit in a porcine model using hyperspectral imaging – an experimental study". In: *International Journal of Surgery* (9900). DOI: 10.1097/JS9.0000000000001849

11. Dominik Rivoir, Martin Wagner, Sebastian Bodenstedt, Keno März, Fiona Kolbinger, Lena Maier-Hein, **Silvia Seidlitz**, Johanna Brandenburg, Beat Peter Müller-Stich, Marius Distler, Jürgen Weitz, and Stefanie Speidel. "Importance of the Data in the Surgical Environment". In: *Artificial Intelligence and the Perspective of Autonomous Surgery*. Ed. by Konrad Karcz, Zbigniew Nawrat, and Andrew A. Gumbs. Cham: Springer Nature Switzerland, 2024, pp. 29–43. ISBN: 978-3-031-68574-3. DOI: 10.1007/978-3-031-68574-3_2

12. Laura Simons, Muath Qadoura, Jule Buhl, Franziska Sunderer, Felix Korell, Ignatios Ikonomidis, Maximilian Dietrich, **Silvia Seidlitz**, Hans Vink, Lena Maier-Hein, Richard F. Schlenk, Carsten Müller-Tidow, Peter Dreger, and Thomas Luft. "Endothelial Dysfunction in Hematological Patients: Sublingual Microscopy, Hyperspectral Imaging, and Serum Endothelial Markers". In: *Blood* 144 (2024). 66th ASH Annual Meeting Abstracts, p. 5519. ISSN: 0006-4971. DOI: 10.1182/blood-2024-205216

13. Leonardo Ayala, Diana Mindroc-Filimon, Maike Rees, Marco Hübner, Jan Sellner, **Silvia Seidlitz**, Minu Tizabi, Sebastian Wirkert, Alexander Seitel, and Lena Maier-Hein. "The SPECTRAL Perfusion Arm Clamping dAtaset (SPECTRALPACA) for video-rate functional imaging of the skin". In: *Scientific Data* 11.1 (May 25, 2024), p. 536. ISSN: 2052-4463. DOI: 10.1038/s41597-024-03307-y

14. Leonardo Ayala, Tim J. Adler, **Silvia Seidlitz**, Sebastian Wirkert, Christina Engels, Alexander Seitel, Jan Sellner, Alexey Aksenov, Matthias Bodenbach, Pia Bader, Sebastian Baron, Anant Vemuri, Manuel Wiesenfarth, Nicholas Schreck, Diana Mindroc, Minu Tizabi, Sebastian Pirmann, Brittaney Everitt, Annette Kopp-Schneider, Dogu Teber, and Lena Maier-Hein. "Spectral imaging enables contrast agent-free real-time ischemia monitoring in laparoscopic surgery". In: *Science Advances* 9.10 (2023), eadd6778. DOI: 10.1126/sciadv.add6778

15. Alexander Studier-Fischer, **Silvia Seidlitz**, Jan Sellner, Marc Bressan, Berkin Özdemir, Leonardo Ayala, Jan Odenthal, Samuel Knoedler, Karl-Friedrich Kowalewski, Caelán Max Haney, Gabriel Salg, Maximilian Dietrich, Hannes Kenngott, Ines Gockel, Thilo Hackert, Beat Peter Müller-Stich, Lena Maier-Hein, and Felix Nickel. "HeiPorSPECTRAL - the Heidelberg Porcine HyperSPECTRAL Imaging Dataset of 20 Physiological Organs". In: *Scientific Data* 10.1 (June 24, 2023), p. 414. ISSN: 2052-4463. DOI: 10.1038/s41597-023-02315-8. URL: https://heiporspectral.org

16. Kris K. Dreher, Leonardo Ayala, Melanie Schellenberg, Marco Hübner, Jan-Hinrich Nölke, Tim J. Adler, **Silvia Seidlitz**, Jan Sellner, Alexander Studier-Fischer, Janek Gröhl, Felix Nickel, Ullrich Köthe, Alexander Seitel, and Lena Maier-Hein. "Unsupervised Domain Transfer withăConditional Invertible Neural Networks". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Cham: Springer Nature Switzerland, 2023, pp. 770–780. ISBN: 978-3-031-43907-0. DOI: 10.1007/978-3-031-43907-0_73

17. Marco Hübner, Leonardo Ayala, Maike Rees, Tim J. Adler, Kris Dreher, **Silvia Seidlitz**, Jan Sellner, Ahmad Bin Qasim, Alexander Seitel, Alexander Studier-Fischer, Alexey Aksenov, Christina Engels, Dogu Teber, Beat Müller-Stich, Felix Nickel, and Lena Maier-Hein. "How to assess the realism of synthetic spectral images". In: *Molecular-Guided Surgery: Molecules, Devices, and Applications IX*. ed. by Sylvain Gioux, Summer L. Gibbs, and Brian W. Pogue. Vol. PC12361. International Society for Optics and Photonics. SPIE, 2023, PC1236104. DOI: 10.1117/12.2648461

18. Alexander Studier-Fischer, Florian Marc Schwab, Maike Rees, **Silvia Seidlitz**, Jan Sellner, Berkin Özdemir, Leonardo Ayala, Jan Odenthal, Samuel Knoedler, Karl-Friedrich Kowalewski, Caelán Max Haney, Maximilian Dietrich, Gabriel Alexander Salg, Hannes Götz Kenngott, Beat Peter Müller-Stich, Lena Maier-Hein, and Felix Nickel. "ICG-augmented hyperspectral imaging for visualization of intestinal perfusion compared to conventional ICG fluorescence imaging: an experimental study". In: *International Journal of Surgery* 109.12 (2023). DOI: 10.1097/JS9.0000000000000706

19. Alexander Studier-Fischer, **Silvia Seidlitz**, Jan Sellner, Berkin Özdemir, Manuel Wiesenfarth, Leonardo Ayala, Jan Odenthal, Samuel Knödler, Karl Friedrich Kowalewski, Caelán Max Haney, Isabella Camplisson, Maximilian Dietrich, Karsten Schmidt, Gabriel Alexander Salg, Hannes Götz Kenngott, Tim Julian Adler, Nicholas Schreck, Annette Kopp-Schneider, Klaus Maier-Hein, Lena Maier-Hein, Beat Peter Müller-Stich, and Felix Nickel. "Spectral organ fingerprints for machine learning-based intraoperative tissue classification with hyperspectral imaging in a porcine model". In: *Scientific Reports* 12.1 (June 30, 2022), p. 11028. ISSN: 2045-2322. DOI: 10.1038/s41598-022-15040-w

20. Leonardo Ayala, Sebastian J. Wirkert, Anant Vemuri, Tim Adler, **Silvia Seidlitz**, Sebastian Pirmann, Christina Engels, Dogu Teber, and Lena Maier-Hein. "Video-rate multispectral imaging in laparoscopic surgery: First-in-human application". In: (2021). DOI: 10.48550/arXiv.2105.13901

21. Tom Rix, Marco Hübner, Kris K. Dreher, Jan-Hinrich Nölke, Leonardo Ayala, Melanie Schellenberg, Jan Sellner, **Silvia Seidlitz**, Alexander Studier-Fischer, Beat Müller-Stich, Felix Nickel, Alexander Seitel, and Lena Maier-Hein. "Deep learning for spectral image synthesis". In: *Multimodal Biomedical Imaging XVII*. ed. by Fred S. Azar, Xavier Intes, and Qianqian Fang. Vol. PC11952. International Society for Optics and Photonics. SPIE, 2022, PC119520I. DOI: 10.1117/12.2608622

22. Claire Chalopin, Felix Nickel, Annekatrin Pfahl, Hannes Köhler, Marianne Maktabi, René Thieme, Robert Sucher, Boris Jansen-Winkeln, Alexander Studier-Fischer, **Silvia Seidlitz**, Lena Maier-Hein, Thomas Neumuth, Andreas Melzer, Beat Peter Müller-Stich, and Ines Gockel. "Artificial intelligence and hyperspectral imaging for image-guided assistance in minimally invasive surgery". In: *Chirurgie (Heidelberg, Germany)* 93.10 (Oct. 2022), pp. 940–947. ISSN: 2731-6971. DOI: 10.1007/s00104-022-01677-w

23. Lena Maier-Hein, Martin Wagner, Tobias Ross, Annika Reinke, Sebastian Bodenstedt, Peter M. Full, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Anna Kisilenko, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Minu D. Tizabi, Matthias Eisenmann, Tim J. Adler, Janek Gröhl, Melanie Schellenberg, **Silvia Seidlitz**, T. Y. Emmy Lai, Bünyamin Pekdemir, Veith Roethlingshoefer, Fabian Both, Sebastian Bittel, Marc Mengler, Lars Mündermann, Martin Apitz, Annette Kopp-Schneider, Stefanie Speidel, Felix Nickel, Pascal Probst, Hannes G. Kenngott, and Beat P. Müller-Stich. "Heidelberg colorectal data set for surgical data science in the sensor operating room". In: *Scientific Data* 8.1 (Apr. 12, 2021), p. 101. ISSN: 2052-4463. DOI: 10.1038/s41597-021-00882-2

A.3 Contributions at International Conferences

I have contributed to several international conferences in the fields of SI, medical image computing, and computer-assisted interventions. Notably, I served as chair of the *Biomedical Applications* session at WHISPERS 2023. In addition, I presented my research through oral and poster presentations and supported collaborators in delivering joint work. The following section lists my conference contributions – limited to those related to the core research questions of this thesis – and categorizes them by presentations delivered personally and those presented by collaborators.

Presentations Delivered Personally

1. Alexander Baumann, Leonardo Ayala, Alexander Studier-Fischer, Jan Sellner, Berkin Özdemir, Karl-Friedrich Kowalewski, Slobodan Ilic, **Silvia Seidlitz**, and Lena Maier-Hein. *Deep intra-operative illumination calibration of hyperspectral*

cameras. Online presentation at the 27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Marrakesh, Morocco. Oct. 9, 2024

2. **Silvia Seidlitz**, Alexander Studier-Fischer, Maximilian Dietrich, Ayca von Garrel, Katharina Hölzl, Felix Nickel, Markus A. Weigand, and Lena Maier-Hein. *Shedding light on hidden factors: Unvealing biases in medical hyperspectral images*. Oral presentation at the 13th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Athens, Greece. Nov. 2, 2023

3. Jan Sellner, **Silvia Seidlitz**, Alexander Studier-Fischer, Alessandro Motta, Berkin Özdemir, Beat Peter Müller-Stich, Felix Nickel, and Lena Maier-Hein. *Semantic Segmentation of Surgical Hyperspectral Images Under Geometric Domain Shifts*. Poster presentation at the 26th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Vancouver, Canada. Oct. 9, 2023

4. **Silvia Seidlitz**, Jan Sellner, Jan Odenthal, Berkin Özdemir, Alexander Studier-Fischer, Samuel Knödler, Leonardo Ayala, Tim J. Adler, Hannes G. Kenngott, Minu Tizabi, Martin Wagner, Felix Nickel, Beat P. Müller-Stich, and Lena Maier-Hein. *Robust deep learning-based semantic organ segmentation in hyperspectral images*. Oral presentation at the 13th International Conference on Information Processing in Computer-Assisted Interventions (IPCAI), Tokyo, Japan. June 7, 2022

Presentations Delivered by Collaborators

1. Maximilian Dietrich, **Silvia Seidlitz**, Katharina Hölzl, Ayca von Garrel, Jan Sellner, Stephan Katzenschlager, Tobias Hölle, Dania Fischer, Maik von der Forst, Felix C. F. Schmitt, Alexander Studier-Fischer, Markus A. Weigand, and Lena Maier-Hein. *AI-Based Spectral Imaging Biomarkers for Recognizing Sepsis and Predicting Mortality in Intensive Care: A Prospective Study of 483 Critically Ill Patients*. Oral presentation at the 5th Conference on Clinical Translation of Medical Image Computing and Computer Assisted Intervention (CLINICCAI), Daejeon, South Korea. Sept. 25, 2025

2. Alexander Baumann, Leonardo Ayala, Alexander Studier-Fischer, Jan Sellner, Berkin Özdemir, Karl-Friedrich Kowalewski, Slobodan Ilic, **Silvia Seidlitz**, and Lena Maier-Hein. *Deep intra-operative illumination calibration of hyperspectral cameras*. Oral presentation at the 27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Marrakesh, Morocco. Oct. 8, 2024

3. Ahmad Bin Qasim, Alessandro Motta, Alexander Studier-Fischer, Jan Sellner, Leonardo Ayala, Marco Hübner, Marc Bressan, Berkin Özdemir, Karl Friedrich

Kowalewski, Felix Nickel, **Silvia Seidlitz**, and Lena Maier-Hein. *Test-time augmentation with synthetic data addresses distribution shifts in spectral imaging*. Oral and poster presentation at the 15th International Conference on Information Processing in Computer-Assisted Interventions (IPCAI), Barcelona, Spain. June 19, 2024

4. Jan Sellner, **Silvia Seidlitz**, and Lena Maier-Hein. *Dealing with I/O bottlenecks in high-throughput model training*. Poster presentation at the PyTorch Conference 2023, San Francisco, United States of America. Oct. 16, 2023. URL: https://e130-hyperspectal-tissue-classification.s3.dkfz.de/figures/PyTorchConference_Poster.pdf

5. Leonardo Ayala, **Silvia Seidlitz**, Anant Vemuri, Sebastian J Wirkert, Thomas Kirchner, Tim J Adler, Christina Engels, Dogu Teber, and Lena Maier-Hein. *Light source calibration for multispectral imaging in surgery*. Oral presentation at the 11th International Conference on Information Processing in Computer-Assisted Interventions (IPCAI), Munich, Germany. June 23, 2020

A.4 Awards

My research was recognized with the following awards:

1. The Cascination & Zeiss Machine Learning in CAI Award 2024
2. Women in MICCAI (WiM) Best Oral Presentation Award 2024, 3^{d} place
3. MICCAI Young Scientist Award 2023
4. MICCAI Student-Author Registration (STAR) Award 2023
5. GIANA Polyp Classification Challenge 2021, 1st place

A.5 Patents

My research has led to the filing of the following patents:

1. Lena Maier-Hein, Sebastian Josef Wirkert, Anant Suraj Vemuri, Leonardo Antonio Ayala Menjivar, **Silvia Seidlitz**, Thomas Kirchner, and Tim Adler. “Method and system for augmented imaging in open treatment using multispectral information”. EP3829416B1. June 9, 2021

2. Lena Maier-Hein, Sebastian Josef Wirkert, Anant Suraj Vemuri, Leonardo Antonio Ayala Menjivar, **Silvia Seidlitz**, Thomas Kirchner, and Tim Adler. "Method and system for augmented imaging using multispectral information". US20220012874A1. Jan. 13, 2022

Additional Results

B

B.1 Hardware-Related Sources of Variation in Hyperspectral Imaging

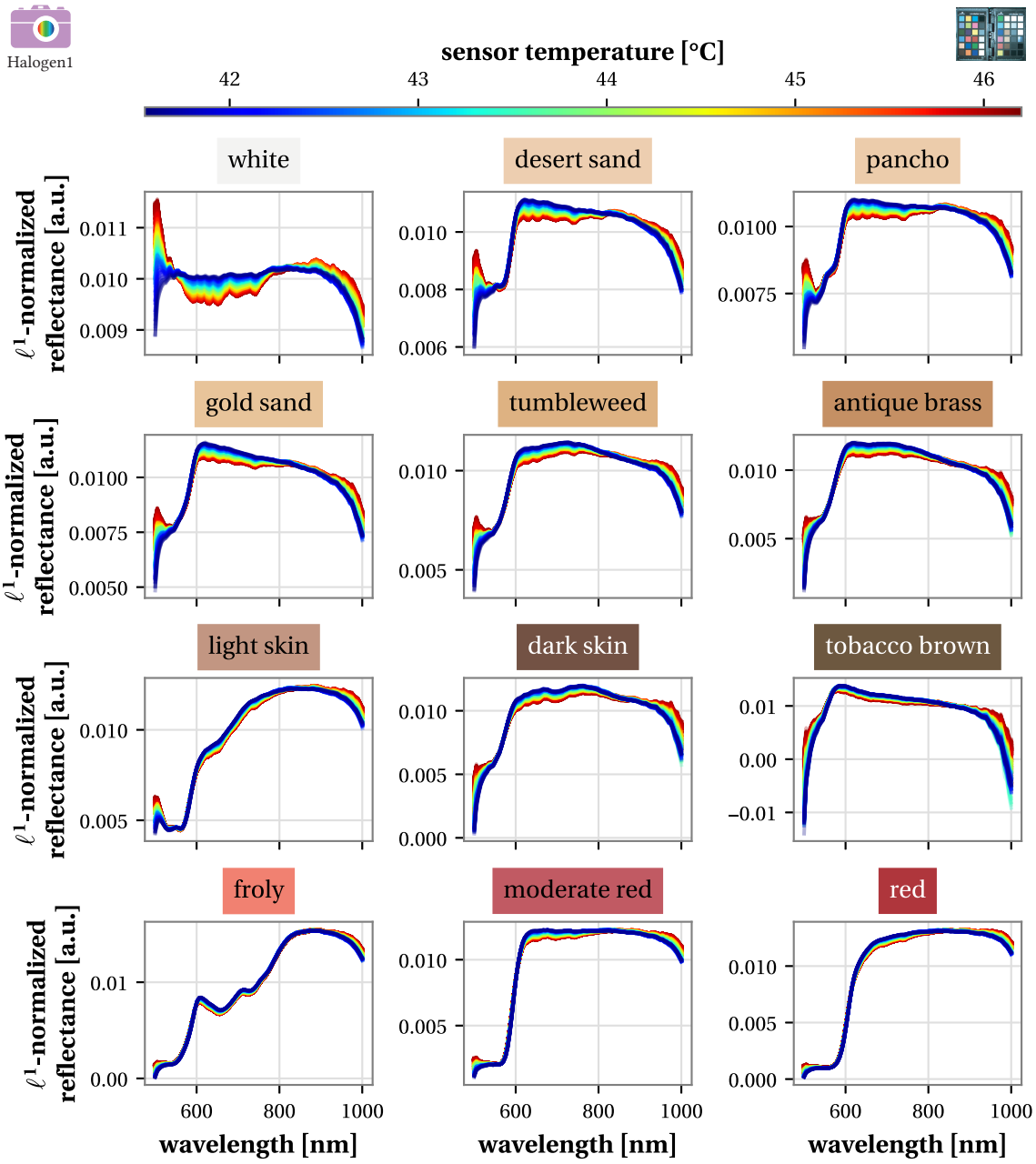


Figure B.1: Shift in colorchecker board spectra measured with the device Halogen1 as a function of sensor temperature. ℓ^1-normalized spectra are shown for 12 color fields of the colorchecker board, with curves color-coded according to the sensor temperature at the time of measurement. For clarity, only data from one of the 3 repetitions is displayed.

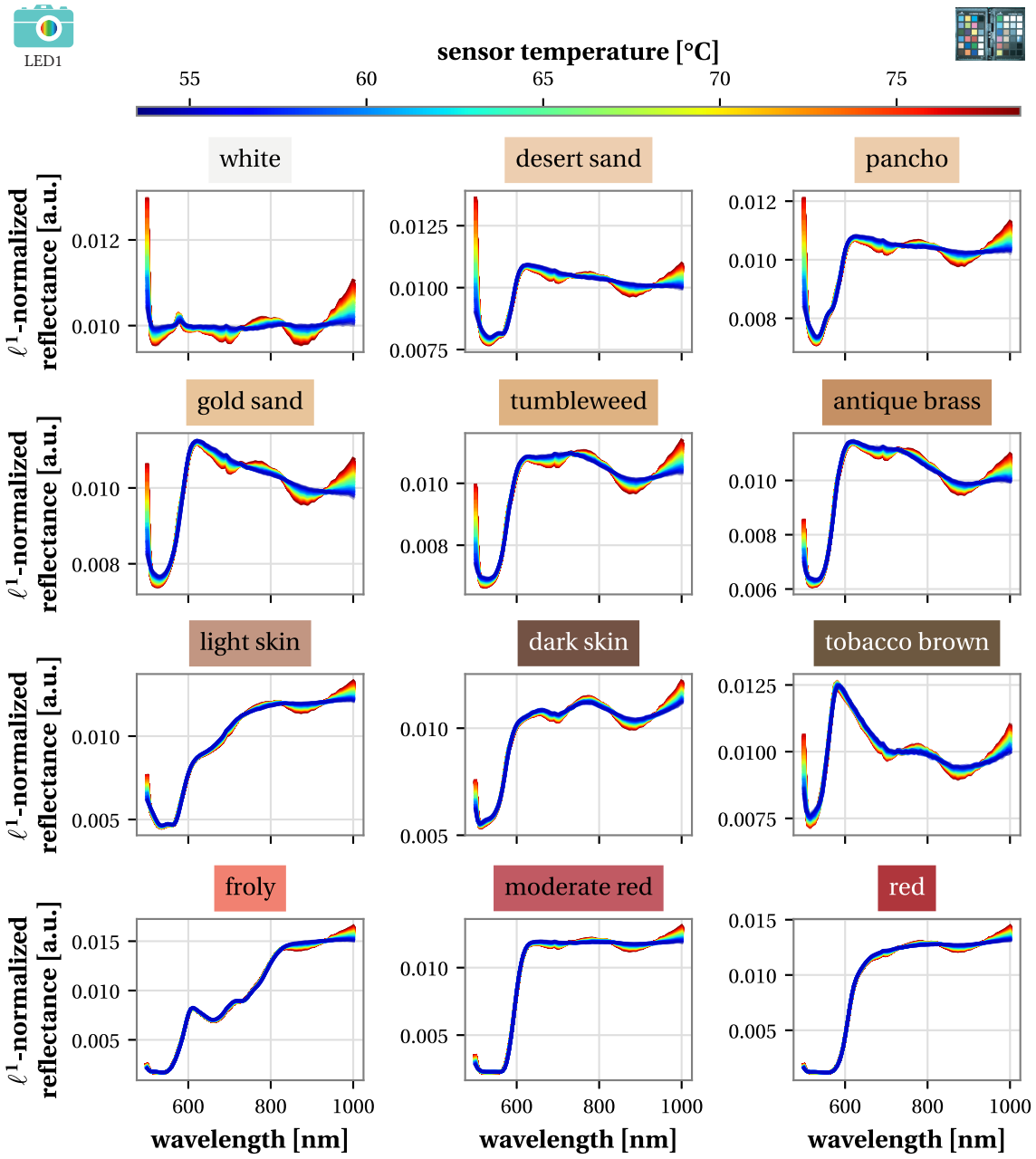


Figure B.2: Shift in colorchecker board spectra measured with the device LED1 as a function of sensor temperature. ℓ^1-normalized spectra are shown for 12 color fields of the colorchecker board, with curves color-coded according to the sensor temperature at the time of measurement. For clarity, only data from one of the 3 repetitions is displayed.

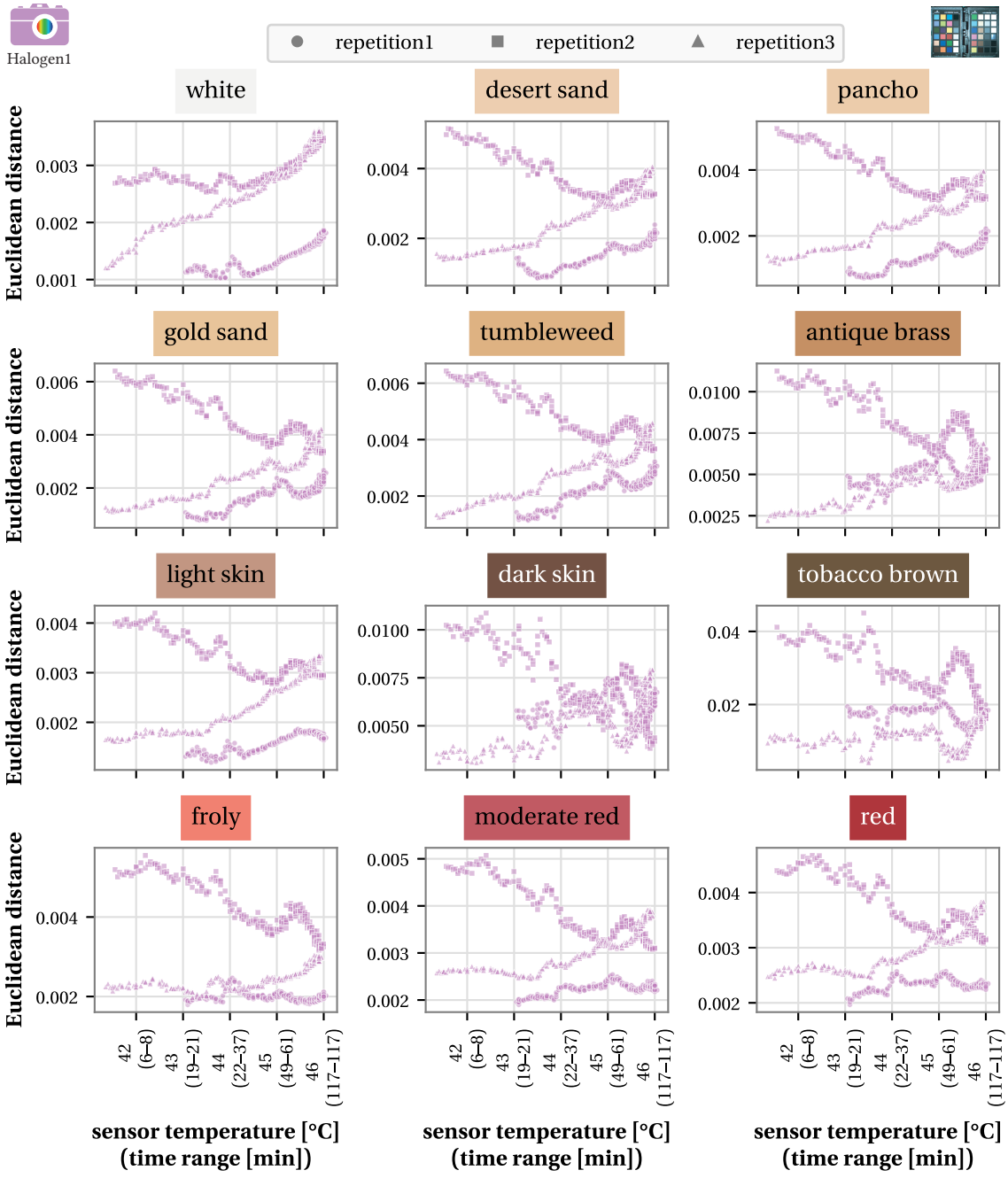


Figure B.3: Shift in Euclidean distance between spectra measured with device Halogen1 and a reference spectrometer as a function of sensor temperature. Euclidean distance is shown as a function of sensor temperature for 12 color fields of the colorchecker board. Measurements from the 3 repetitions are distinguished by different markers.

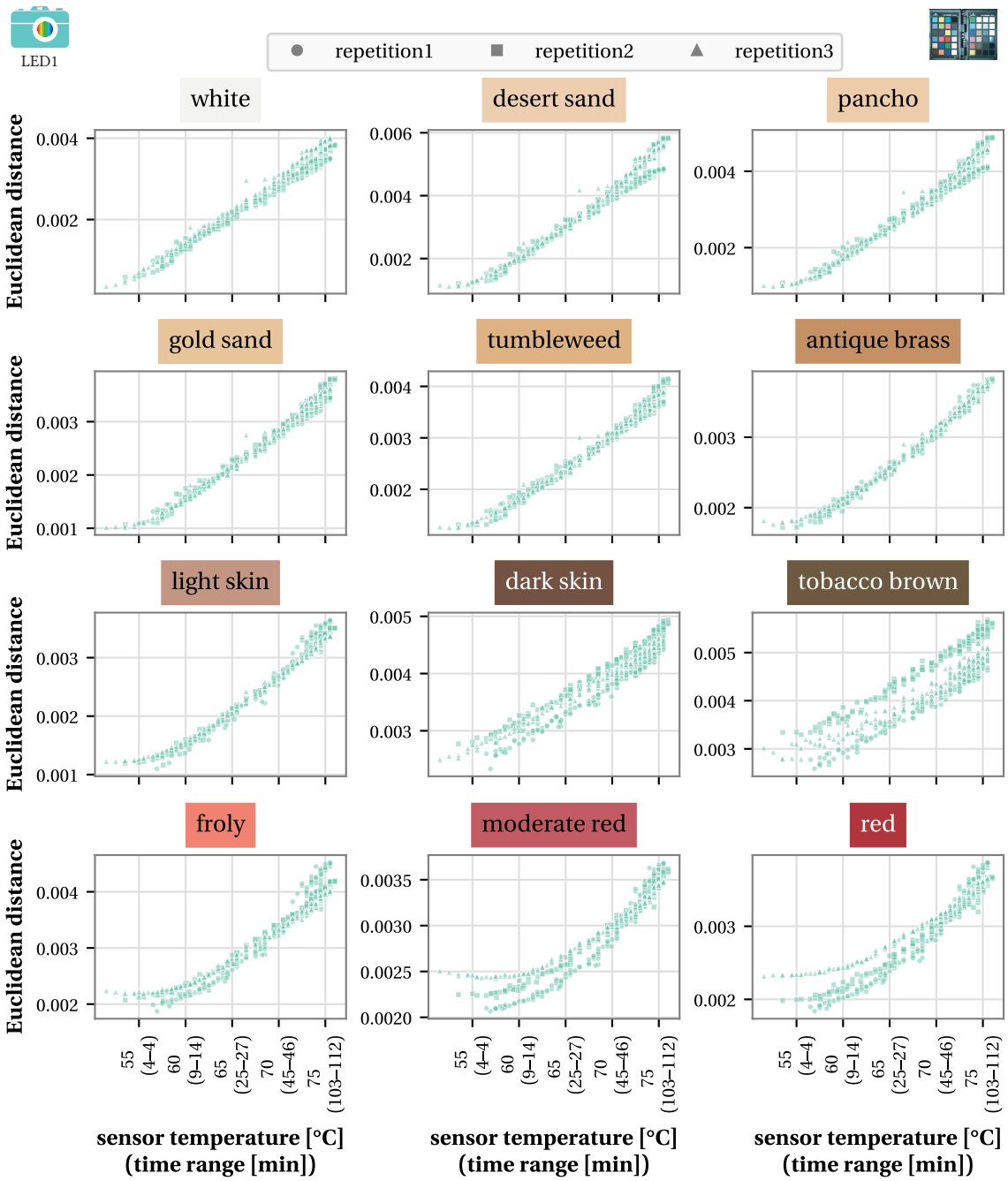


Figure B.4: Shift in Euclidean distance between spectra measured with device LED1 and a reference spectrometer as a function of sensor temperature. Euclidean distance is shown as a function of sensor temperature for 12 color fields of the colorchecker board. Measurements from the 3 repetitions are distinguished by different markers.

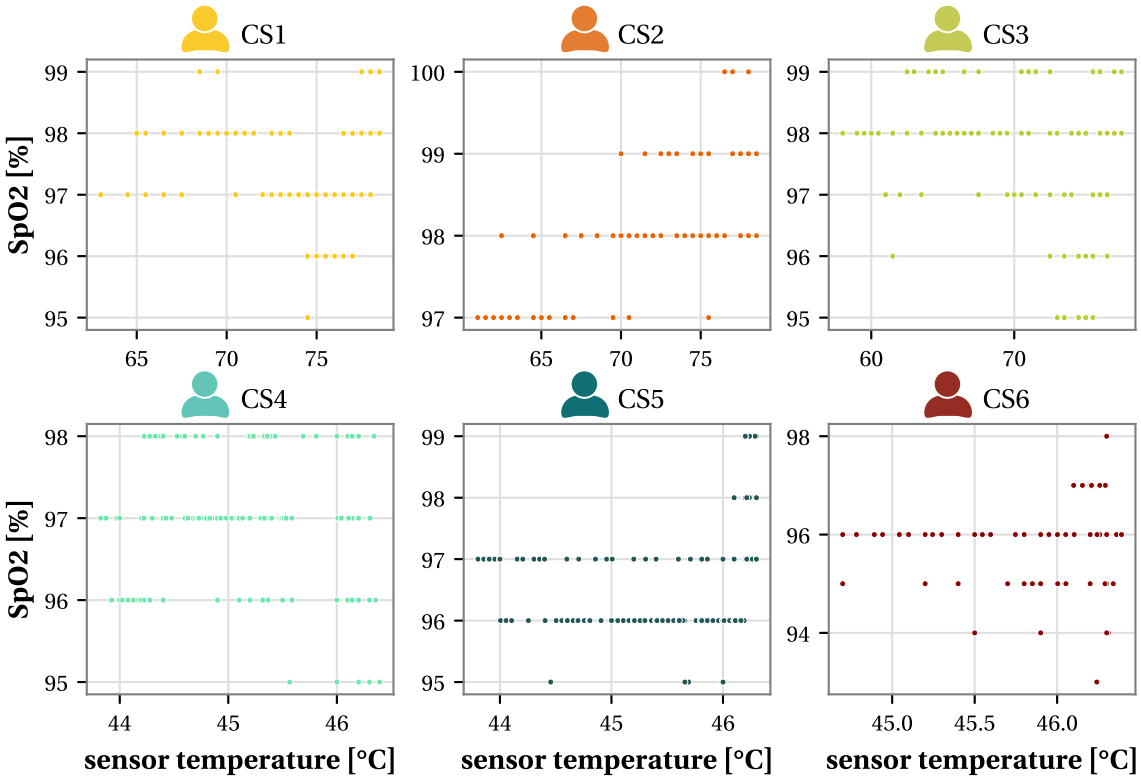


Figure B.5: Recordings of pulse oxymetrical oxygen saturation (SpO_2) in human probands during sensor temperature experiments. SpO_2 was measured in 6 healthy volunteers using a pulse oximeter on the same hand examined with HSI. Scatter plots show SpO_2 values against the corresponding sensor temperature of the HSI device used in parallel (LED2 for probands CS1 to CS3 and Halogen2 for probands CS4 to CS6).

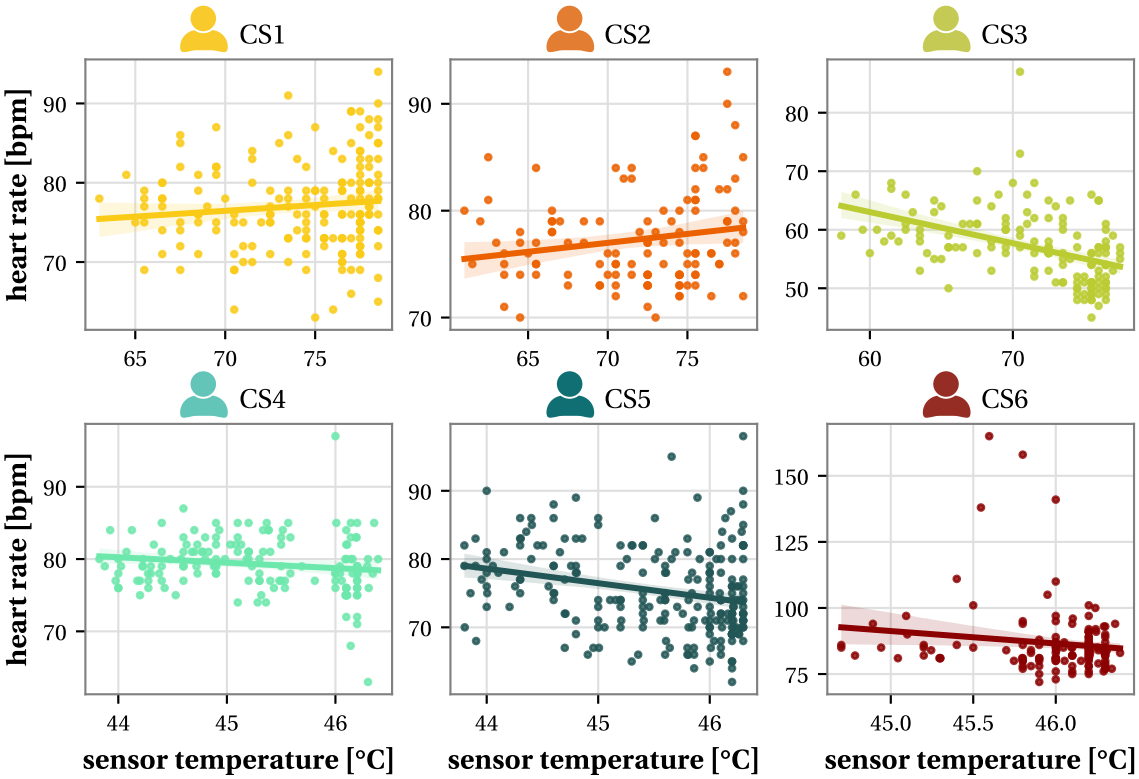


Figure B.6: Recordings of heart rate in human probands during sensor temperature experiments. Heart rate was measured in 6 healthy volunteers using a pulse oximeter on the same hand examined with HSI. Scatter plots show heart rate values against the corresponding sensor temperature of the HSI device used in parallel (LED2 for probands CS1 to CS3 and Halogen2 for probands CS4 to CS6). Linear regression fits are shown as lines, with the 95 % confidence interval derived from 1000 bootstrap samples indicated by shaded areas.

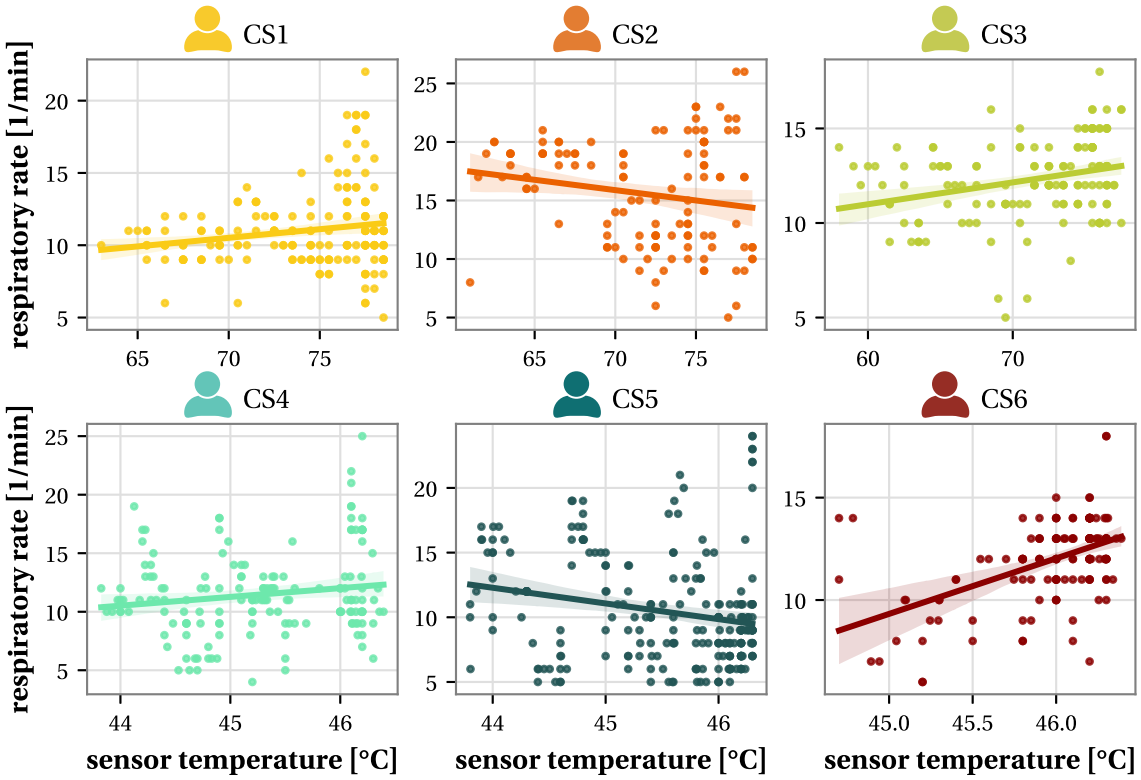


Figure B.7: Recordings of respiratory rate in human probands during sensor temperature experiments. Respiratory rate was measured in 6 healthy volunteers using a pulse oximeter on the same hand examined with HSI. Scatter plots show respiratory rate values against the corresponding sensor temperature of the HSI device used in parallel (LED2 for probands CS1 to CS3 and Halogen2 for probands CS4 to CS6). Linear regression fits are shown as lines, with the 95 % confidence interval derived from 1000 bootstrap samples indicated by shaded areas.

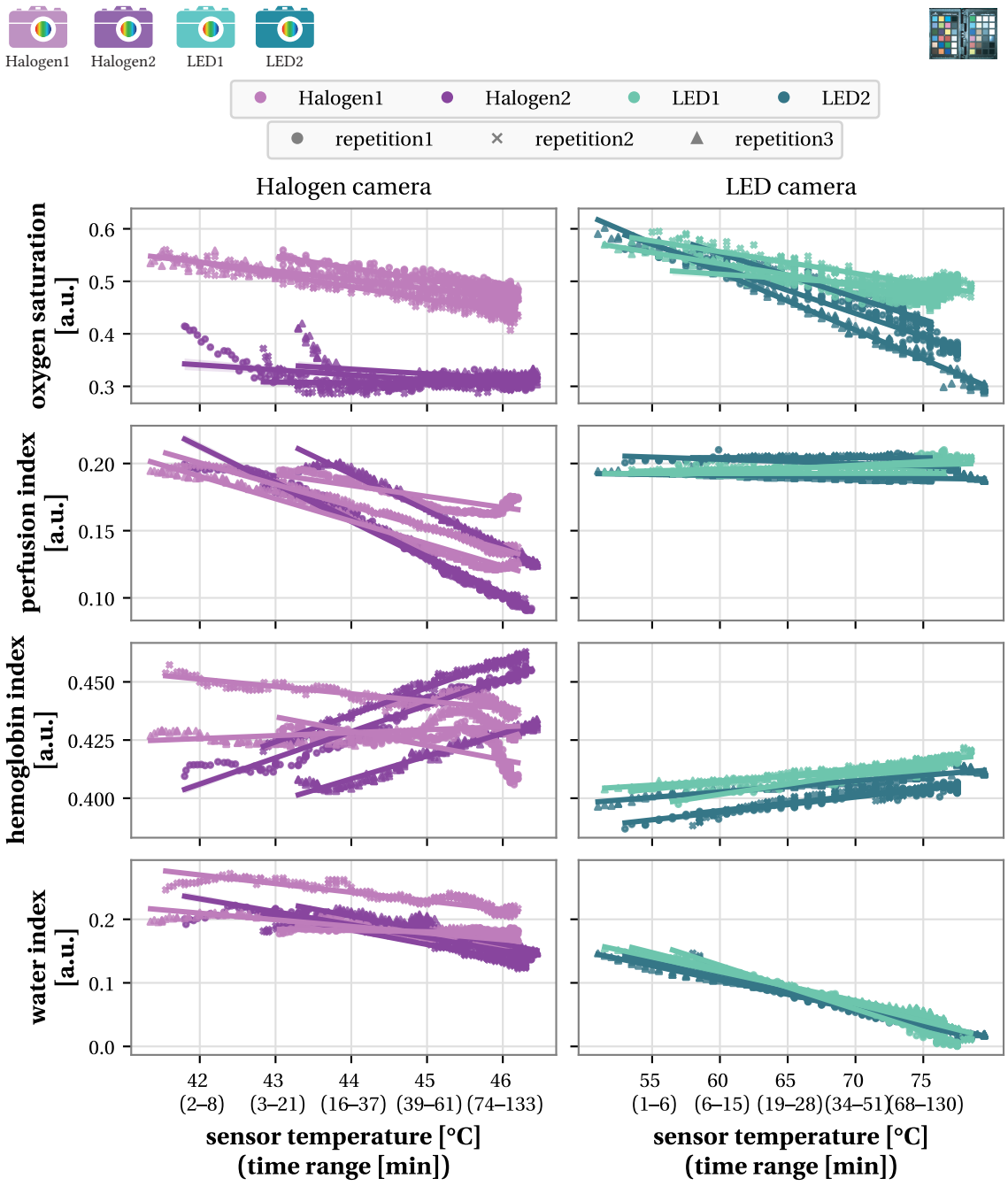


Figure B.8: Impact of sensor temperature increase on functional tissue parameter indices. Scatter plots show measurements of the tissue parameter indices oxygen saturation, perfusion index, hemoglobin index and water index on the "light skin" color field of a colorchecker board, plotted against sensor temperature for the devices Halogen1, Halogen2, LED1 and LED2. Linear regression fits are shown as solid lines, with shaded areas indicating the 95 % confidence interval from 1000 bootstrap samples.

B.2 Impact of Spatial Granularity And Modality on Surgical Scene Segmentation

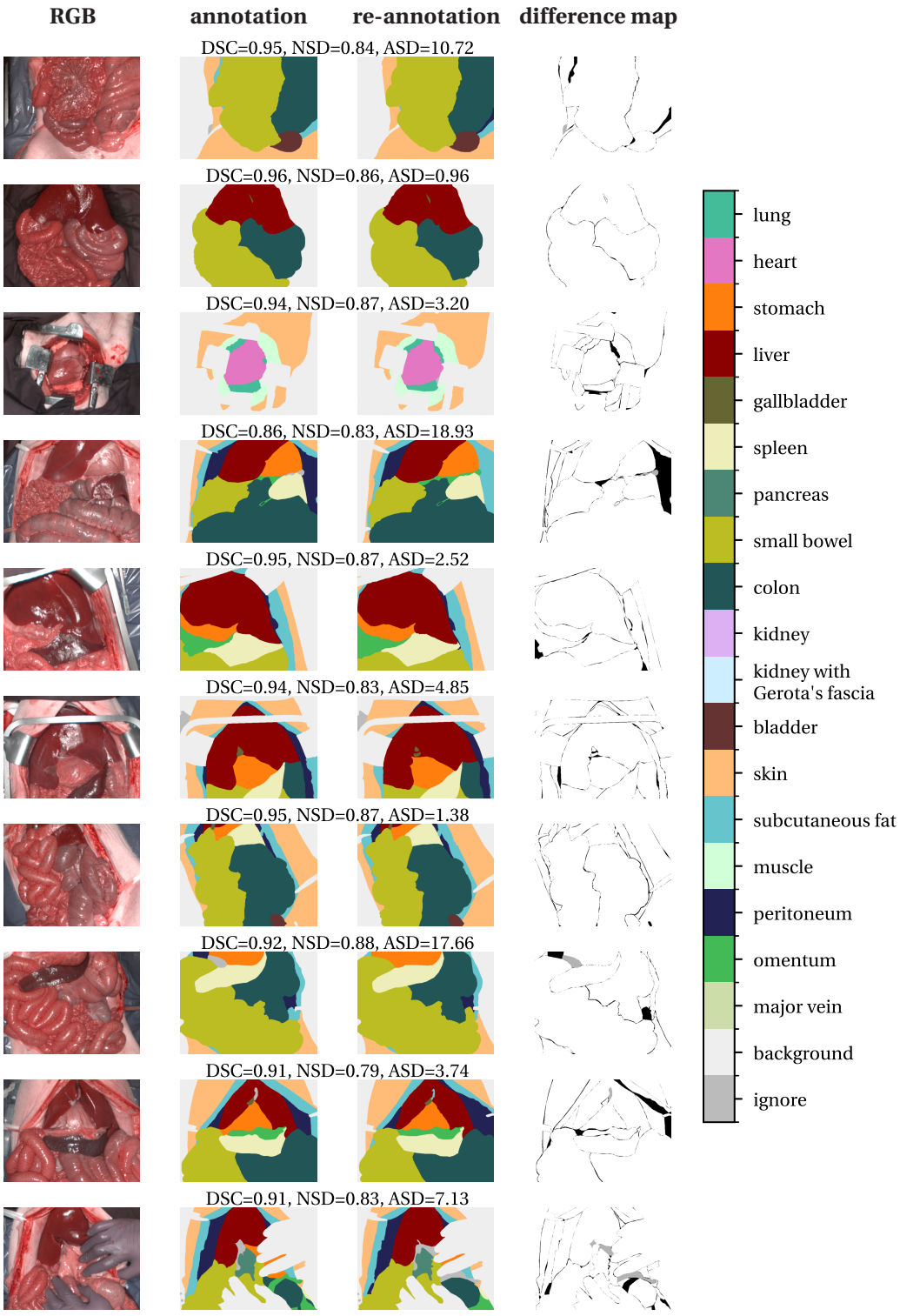
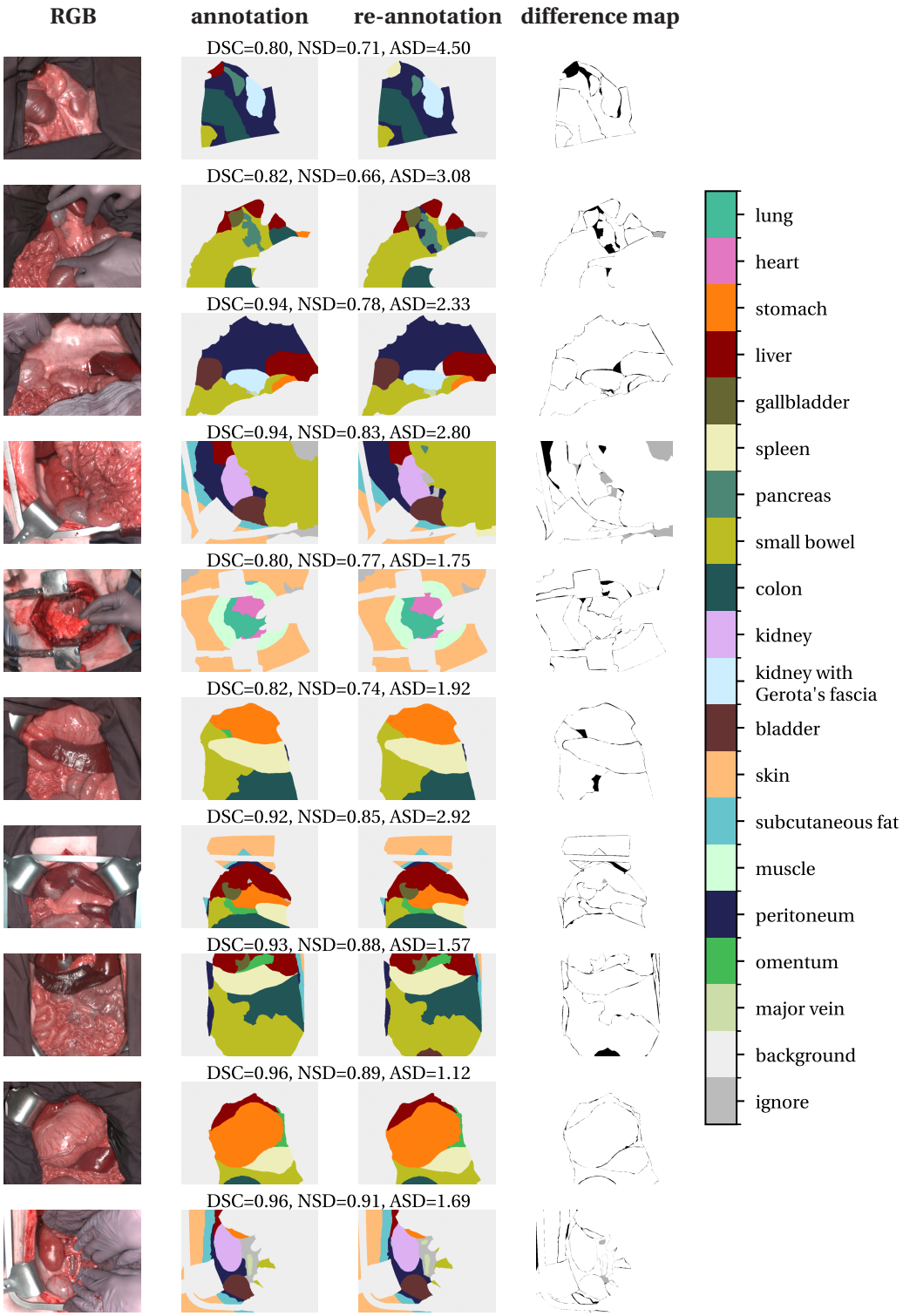


Figure B.9: Intra-rater agreement of reference annotations. Re-annotations of the twenty selected images are shown with their RGB images, original annotations, and difference maps between the annotations. Figure continued on the next page.



Continued Figure B.9: Intra-rater agreement of reference annotations (continuation). Mismatches between the "ignore" class and a valid class are highlighted in gray, while discrepancies between valid classes are marked in black.

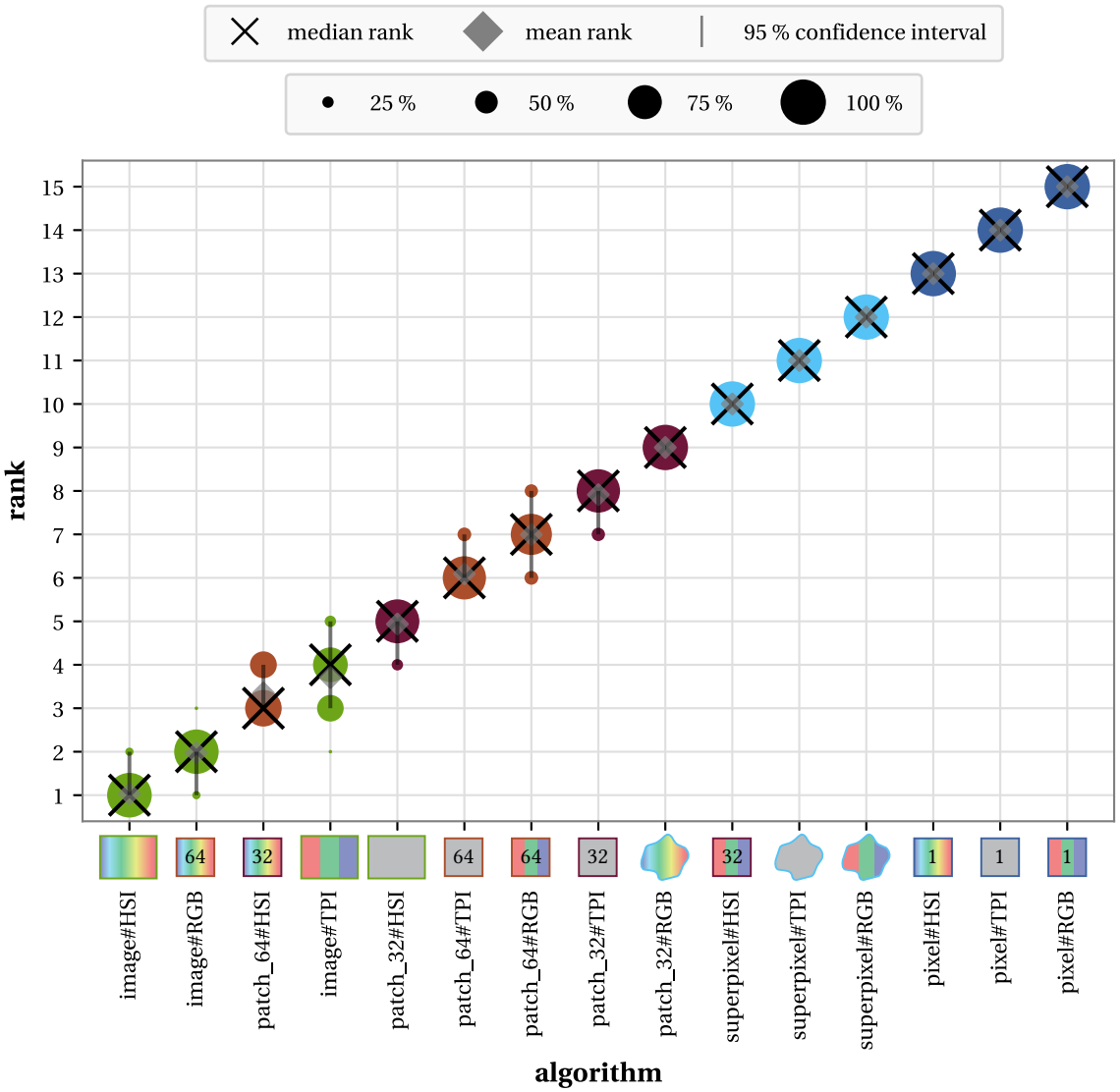


Figure B.10: Ranking stability of our segmentation algorithms with respect to sampling variability using the normalized surface Dice (NSD). Following the concept from [364], bootstrap sampling was performed to assess the ranking stability of our segmentation algorithms across different spatial granularities (pixel, superpixel, patch_32, patch_64 and image) and modalities (RGB, tissue parameter images (TPI), and hyperspectral imaging (HSI)) and modalities (RGB, tissue parameter images (TPI), and hyperspectral imaging (HSI)). For each blob at position (a, rank r), its area is proportional to the frequency of algorithm a achieving rank r across 1000 bootstrap samples. Each sample comprises 5 subject-level NSD values. For each method, black crosses indicate the median rank, gray diamonds show the mean rank, and gray lines represent the 95 % quantile of the bootstrap results. Ranking stability figures for the Dice similarity coefficient (DSC) and average surface distance (ASD) are available in Figure 5.6 and Figure B.11, respectively. Figure adapted from [308, 311].

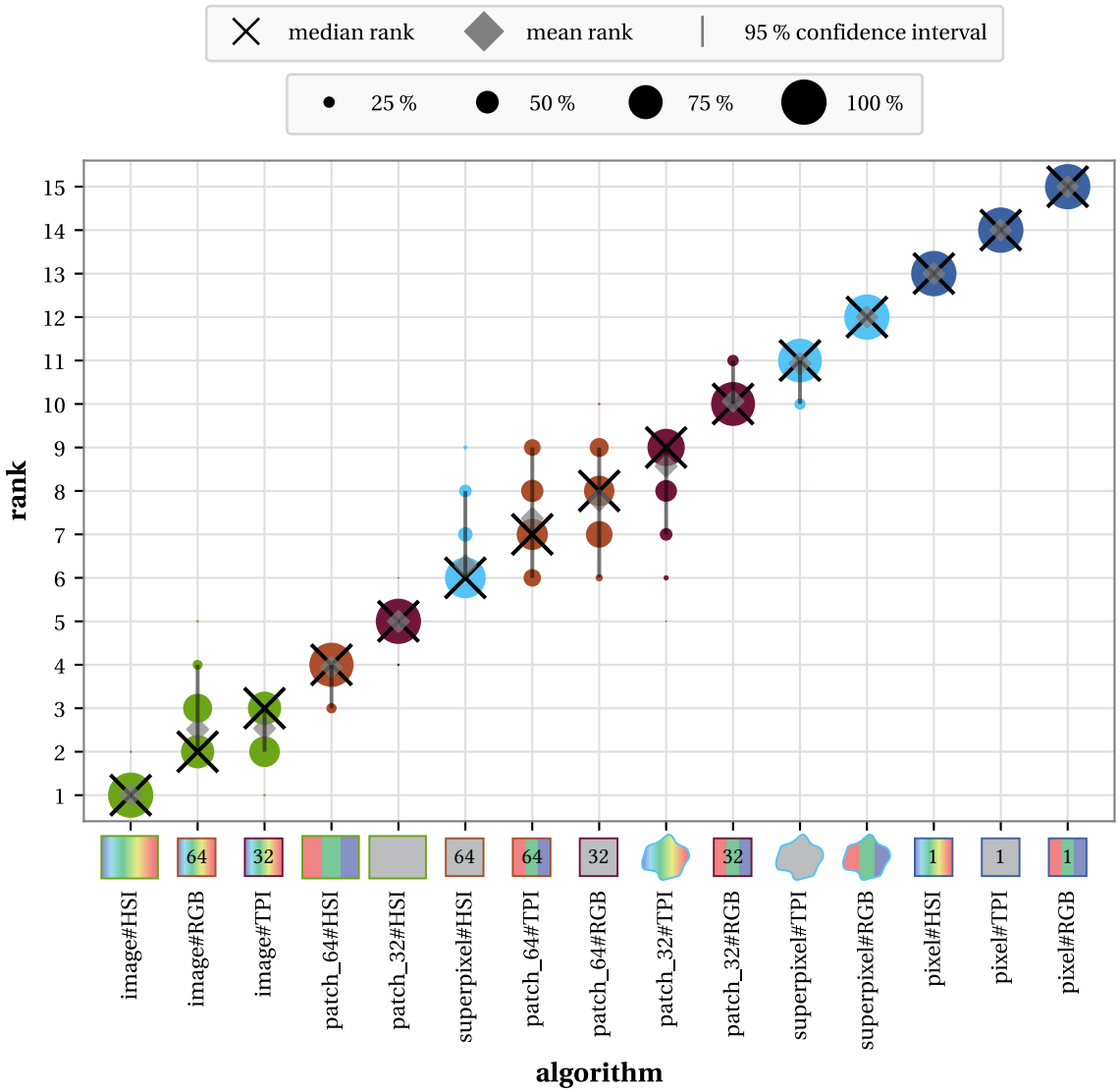


Figure B.11: Ranking stability of our segmentation algorithms with respect to sampling variability using the average surface distance (ASD). Following the concept from [364], bootstrap sampling was performed to assess the ranking stability of our segmentation algorithms across different spatial granularities (pixel, superpixel, patch_32, patch_64 and image) and modalities (RGB, tissue parameter images (TPI), and hyperspectral imaging (HSI)) and modalities (RGB, tissue parameter images (TPI), and hyperspectral imaging (HSI)). For each blob at position (a, rank r), its area is proportional to the frequency of algorithm a achieving rank r across 1000 bootstrap samples. Each sample comprises 5 subject-level ASD values. For each method, black crosses indicate the median rank, gray diamonds show the mean rank, and gray lines represent the 95 % quantile of the bootstrap results. Ranking stability figures for the Dice similarity coefficient (DSC) and normalized surface Dice (NSD) are available in Figure 5.6 and Figure B.10, respectively. Figure adapted from [308, 311].

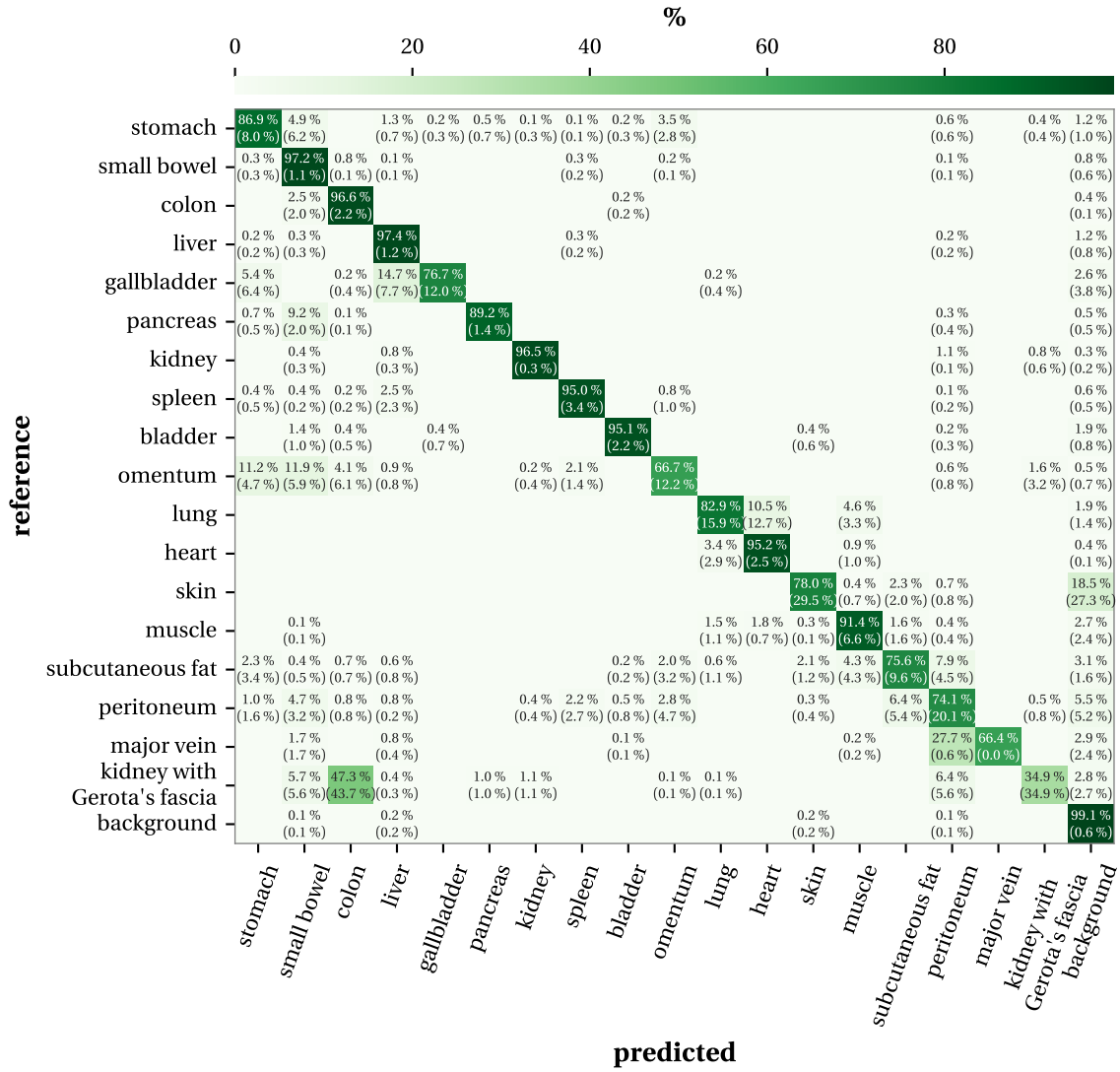


Figure B.12: Confusion matrix for image-based segmentation using tissue parameter images data. Each entry (i, j) denotes the average proportion of pixels from the reference class i that are classified as class j, with values below 0.1 % omitted for clarity. Confusion matrices were row-normalized using pixel data from all images of a single subject, and the subject-specific matrices were averaged across subjects to produce the final confusion matrix. The standard deviation across subjects is indicated in brackets. Diagonal entries correspond to recall (sensitivity). Figures for the hyperspectral imaging and RGB modalities are provided in Figure 5.10 and Figure B.13, respectively. Figure adapted from [308, 311].

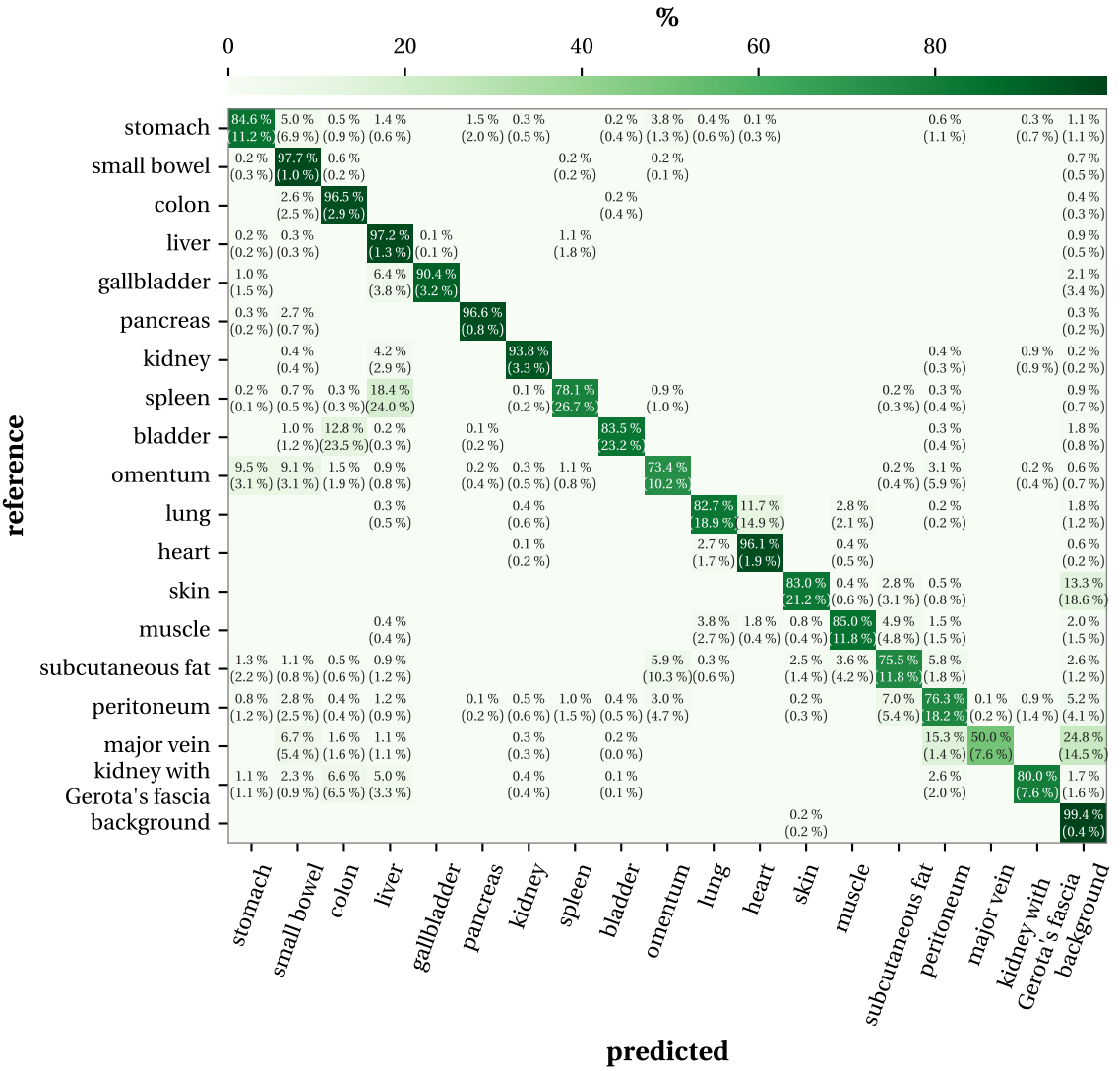


Figure B.13: Confusion matrix for image-based segmentation using RGB data. Each entry (i, j) denotes the average proportion of pixels from the reference class i that are classified as class j, with values below 0.1 % omitted for clarity. Confusion matrices were row-normalized using pixel data from all images of a single subject, and the subject-specific matrices were averaged across subjects to produce the final confusion matrix. The standard deviation across subjects is indicated in brackets. Diagonal entries correspond to recall (sensitivity). Figures for the hyperspectral imaging and tissue parameter images modalities are provided in Figure 5.10 and Figure B.12, respectively. Figure adapted from [308, 311].

B.3 Robust Surgical Scene Segmentation Under Geometric Domain Shifts

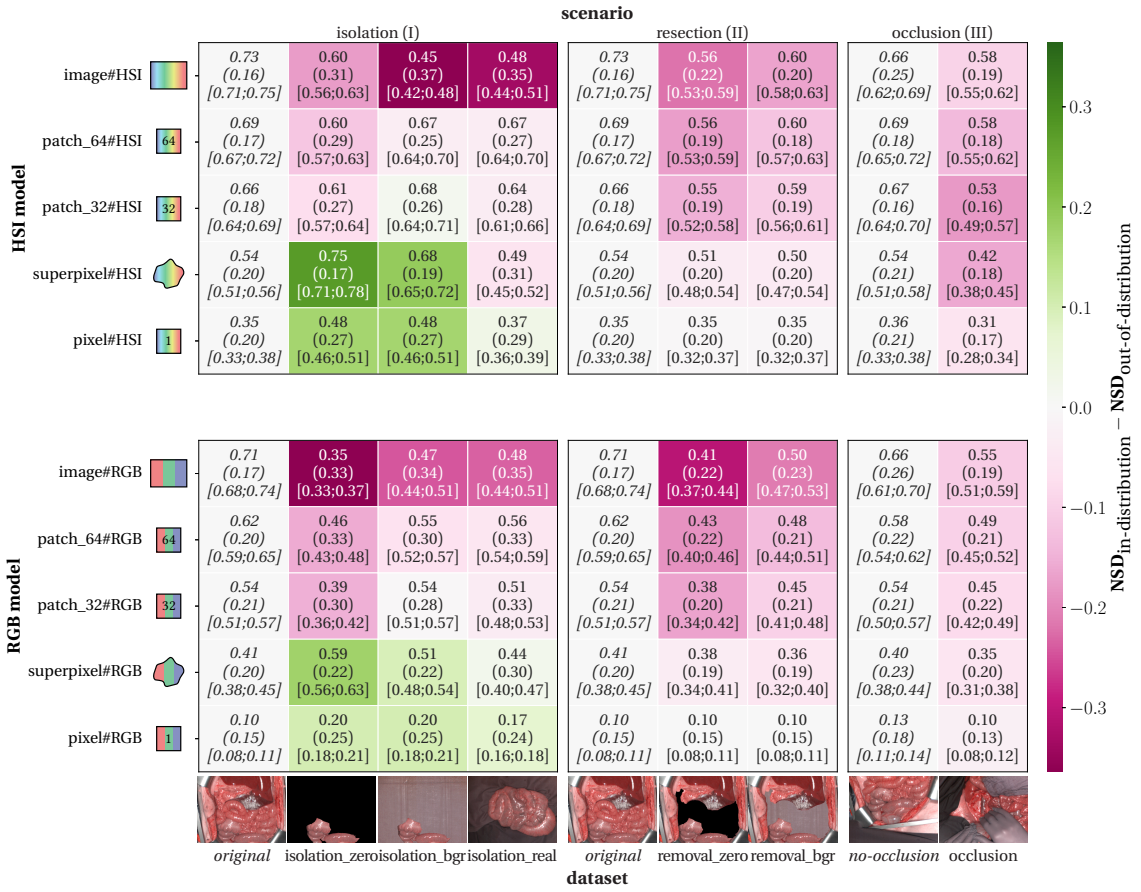


Figure B.14: Role of the input modality and spatial granularity in segmentation performance degradation under geometric domain shifts as measured by the normalized surface Dice (NSD). The segmentation performance is reported for 3 clinical scenarios: organs in isolation (I), organ resections (II), and situs occlusions (III). Columns represent the corresponding in-distribution datasets (highlighted in italic) and out-of-distribution (OOD) datasets. Rows indicate different models, each combining one of two modalities (RGB or hyperspectral imaging (HSI)) with one of 5 spatial granularities: pixel, superpixel, patches of size 32×32 (patch_32) or 64×64 (patch_64), and image. The numbers represent the average DSC across classes, with standard deviations denoted in brackets. The color-coding reflects the difference in DSC relative to the corresponding in-distribution DSC for the same model. Results for the Dice similarity coefficient (DSC) are shown in Figure 6.2. Figure adapted from [309].

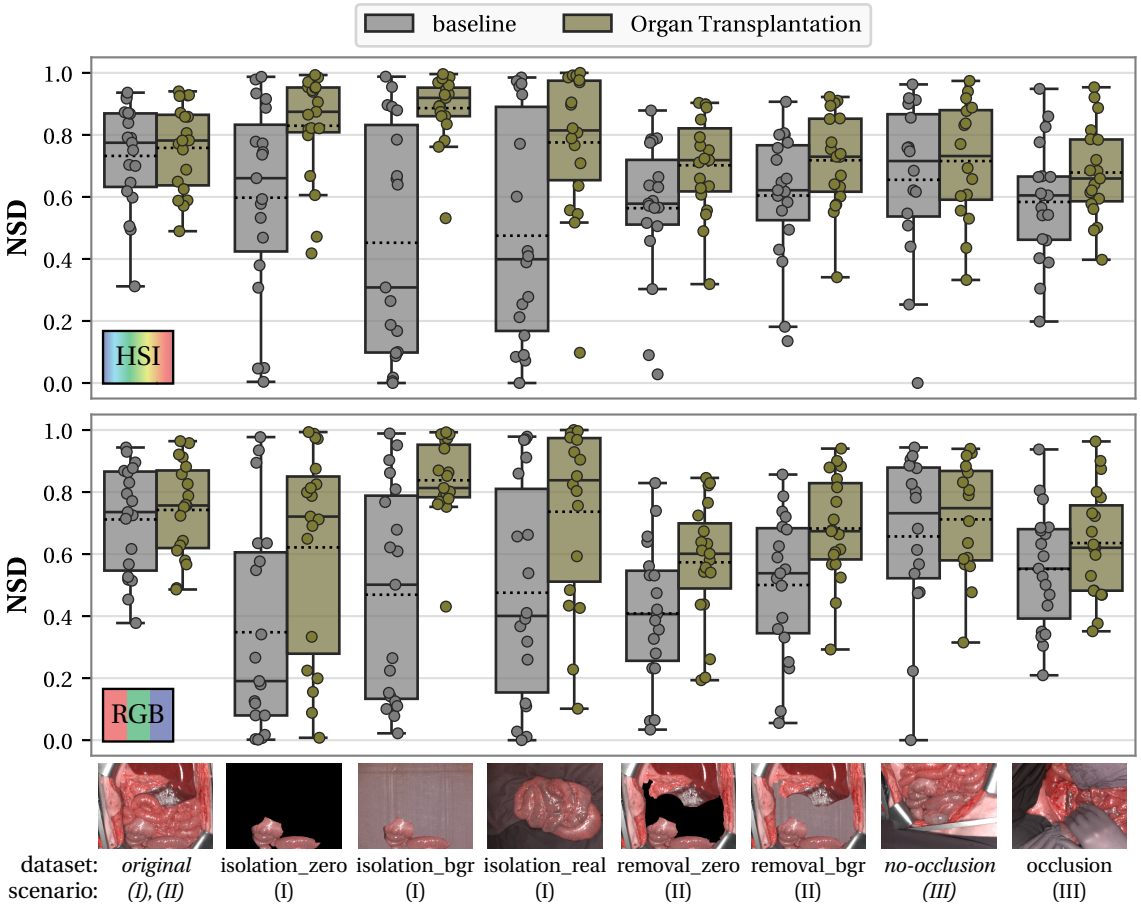


Figure B.15: Performance comparison of the baseline model and the Organ Transplantation model under geometric domain shifts using the normalized surface Dice (NSD). Distributions of class-wise NSD scores are shown for the baseline model and the Organ Transplantation model across the 3 clinical scenarios (I) organs in isolation, (II) organ resections, and (III) situs occlusions, with in-distribution datasets highlighted in italic. The boxplots illustrate the quartiles of the distribution across classes, with whiskers showing the range excluding outliers. The median is shown as a solid line, the mean as a dotted line, and the markers correspond to individual classes. Results for the Dice similarity coefficient (DSC) are shown in Figure 6.4. Figure adapted from [314, 309, 311].

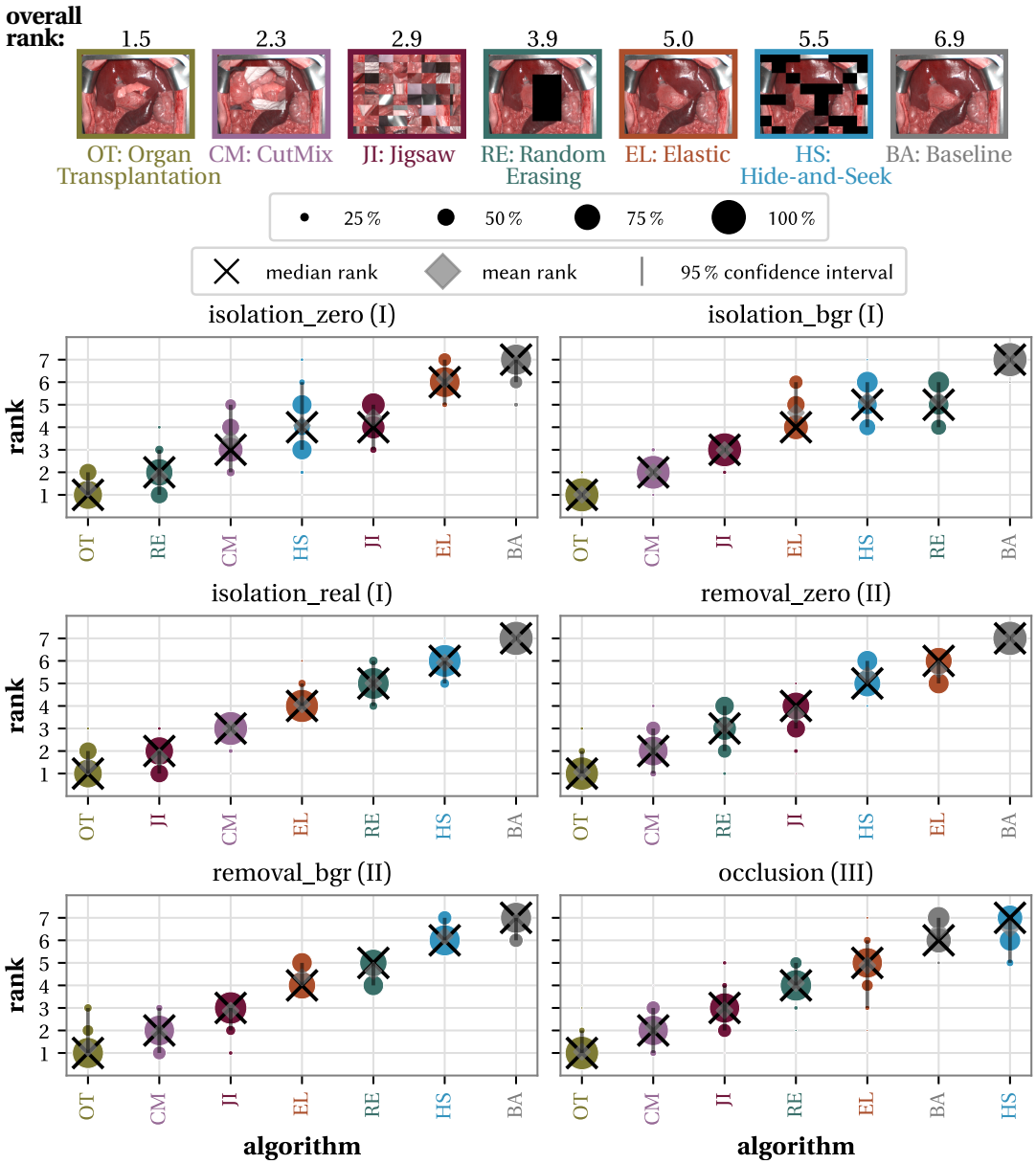


Figure B.16: Uncertainty-aware normalized surface Dice (NSD)-based ranking of different data augmentation methods for addressing geometric domain shifts. Following the concept from [364], bootstrap sampling was performed to assess the ranking stability with respect to sampling variability of our image#HSI models utilizing the data augmentation techniques Organ Transplantation (OT), CutMix (CM), Jigsaw (JI), Random Erasing (RE), Elastic transformations (EL), Hide-and-Seek (HS) and Baseline geometric transformations (BA). For each blob at position (a, r), its area is proportional to the frequency of algorithm a achieving rank r across 1000 bootstrap samples. For each method, black crosses indicate the median rank, gray diamonds show the mean rank, and gray lines represent the 95 % quantile of the bootstrap results. Ranking stability results for the Dice similarity coefficient (DSC) are shown in Figure 6.4. Figure adapted from [314, 309, 311].

B.4 AI-Driven Skin Spectral Imaging for Rapid Sepsis Diagnosis and Mortality Prediction in Critically Ill Patients

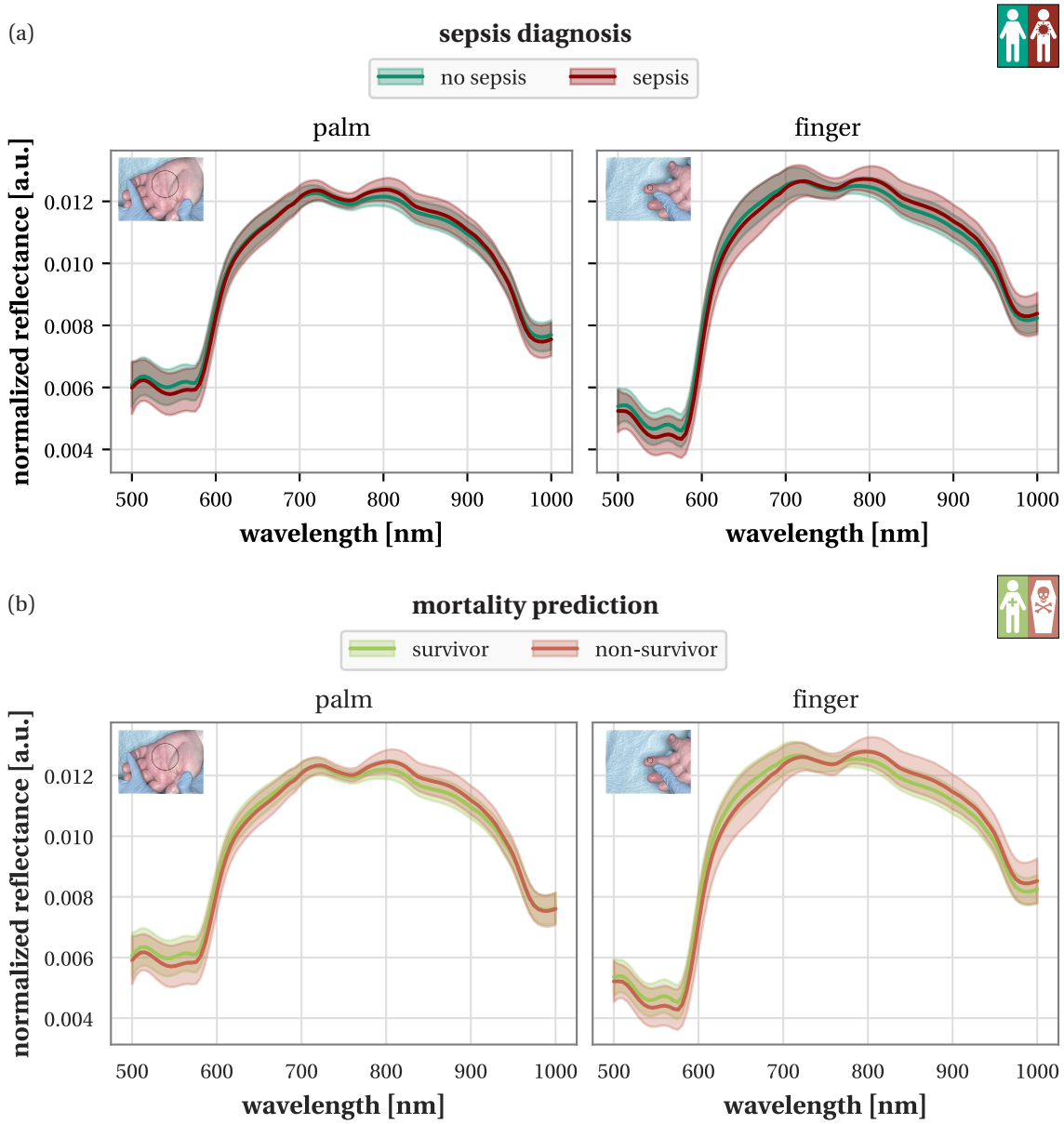


Figure B.17: Characteristic spectra for (a) non-septic and septic patients and (b) survivors and non-survivors. The plots show ℓ^1-normalized spectra for the measurement sites palm (left) and finger (right). Median spectra were first computed for each annotated image region and then averaged across patients (solid lines), with shaded areas indicating one standard deviation. Figure adapted from [306].

Table B.1: Statistical analysis of significant differences in functional tissue parameters for non-septic and septic patients, and survivors and non-survivors. Stratified by measurement site (first column), two-sided Welch's t-tests [363] were conducted to assess significant differences in functional tissue parameter indices (third column) by sepsis and survival status (second column). The table summarizes the p-values, degrees of freedom (DOF), t-statistic and 95 % confidence interval (CI). p-values are colored according to whether they are below or above the Bonferroni-corrected significance level of 0.0125. Table adapted from [306].

site	target	functional parameter	p-value	DOF	t-statistic	95 % CI
palm	sepsis	oxygen saturation	$7.1 \cdot 10^{-4}$	208	-3.44	$[-0.06; -0.02]$
palm	sepsis	perfusion index	$1.1 \cdot 10^{-1}$	205	-1.63	$[-0.03; 0.00]$
palm	sepsis	hemoglobin index	$6.2 \cdot 10^{-5}$	198	4.09	$[0.03; 0.08]$
palm	sepsis	water index	$4.5 \cdot 10^{-10}$	222	6.53	$[0.03; 0.06]$
palm	survival	oxygen saturation	$6.8 \cdot 10^{-4}$	79	-3.54	$[-0.09; -0.02]$
palm	survival	perfusion index	$2.5 \cdot 10^{-3}$	82	-3.12	$[-0.06; -0.01]$
palm	survival	hemoglobin index	$6.0 \cdot 10^{-4}$	81	3.57	$[0.03; 0.09]$
palm	survival	water index	$7.0 \cdot 10^{-5}$	93	4.16	$[0.02; 0.05]$
finger	sepsis	oxygen saturation	$1.4 \cdot 10^{-4}$	176	-3.89	$[-0.07; -0.02]$
finger	sepsis	perfusion index	$1.5 \cdot 10^{-3}$	196	-3.22	$[-0.06; -0.02]$
finger	sepsis	hemoglobin index	$4.4 \cdot 10^{-7}$	205	5.22	$[0.05; 0.10]$
finger	sepsis	water index	$1.2 \cdot 10^{-1}$	194	1.56	$[-0.00; 0.02]$
finger	survival	oxygen saturation	$3.7 \cdot 10^{-4}$	75	-3.73	$[-0.10; -0.03]$
finger	survival	perfusion index	$5.4 \cdot 10^{-4}$	79	-3.61	$[-0.09; -0.03]$
finger	survival	hemoglobin index	$4.6 \cdot 10^{-4}$	81	3.65	$[0.03; 0.11]$
finger	survival	water index	$5.6 \cdot 10^{-1}$	84	-0.59	$[-0.02; 0.01]$

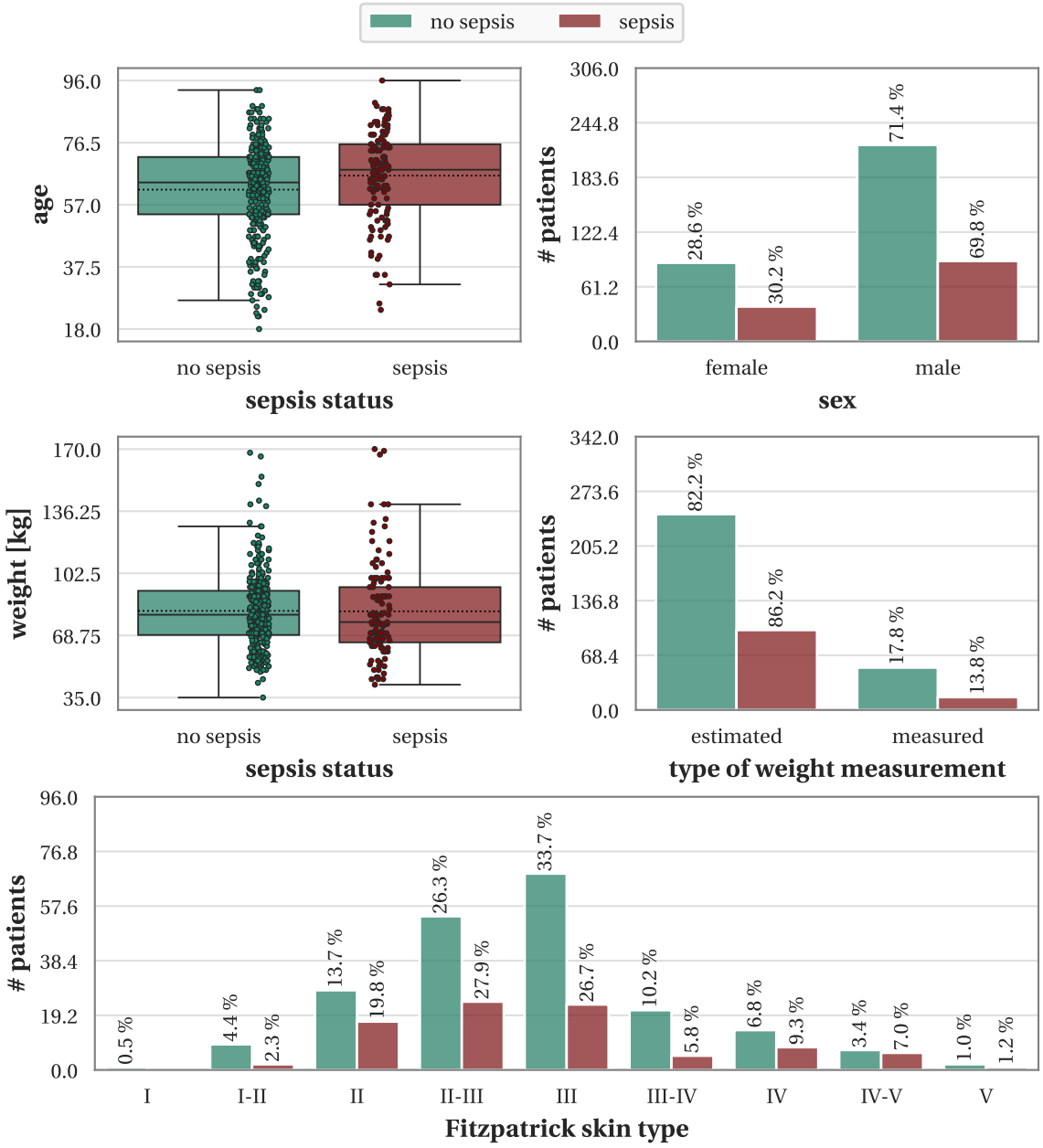


Figure B.18: Distribution of demographic parameters among non-septic and septic patients. Shown are the distributions of age, sex, weight, type of weight measurement (measured or estimated) and Fitzpatrick skin type. For continuous parameters, boxplots indicate the quartiles, with whiskers extending to the range excluding outliers. Solid and dashed lines mark the median and mean, respectively. Each dot corresponds to one patient. For categorical parameters, bar plots display the number of patients (# patients) per category, with percentages given relative to all patients of the same sepsis status.

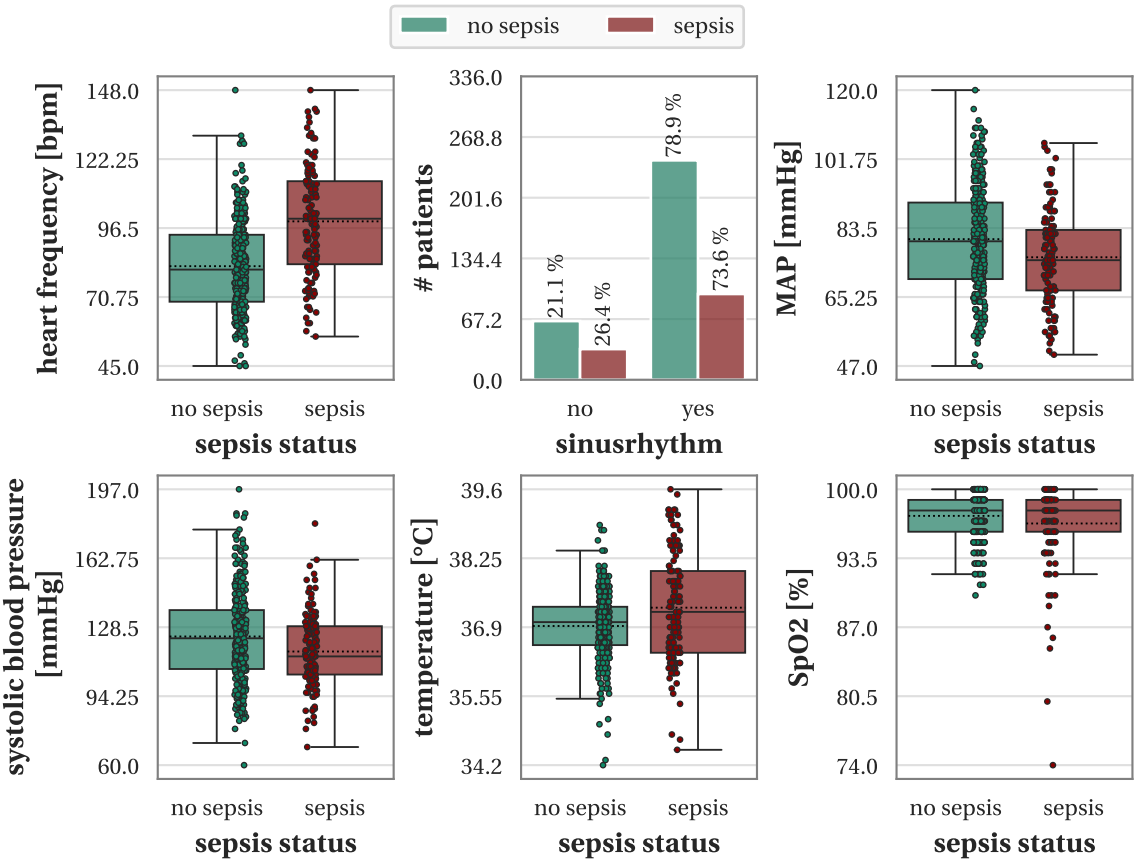


Figure B.19: Distribution of vital parameters among non-septic and septic patients. Shown are the distributions of heart frequency, presence of a sinusrhythm, mean arterial pressure (MAP), systolic blood pressure, temperature and pulse oxymetrical oxygen saturation (SpO_2). For continuous parameters, boxplots indicate the quartiles, with whiskers extending to the range excluding outliers. Solid and dashed lines mark the median and mean, respectively. Each dot corresponds to one patient. For categorical parameters, bar plots display the number of patients (# patients) per category, with percentages given relative to all patients of the same sepsis status.

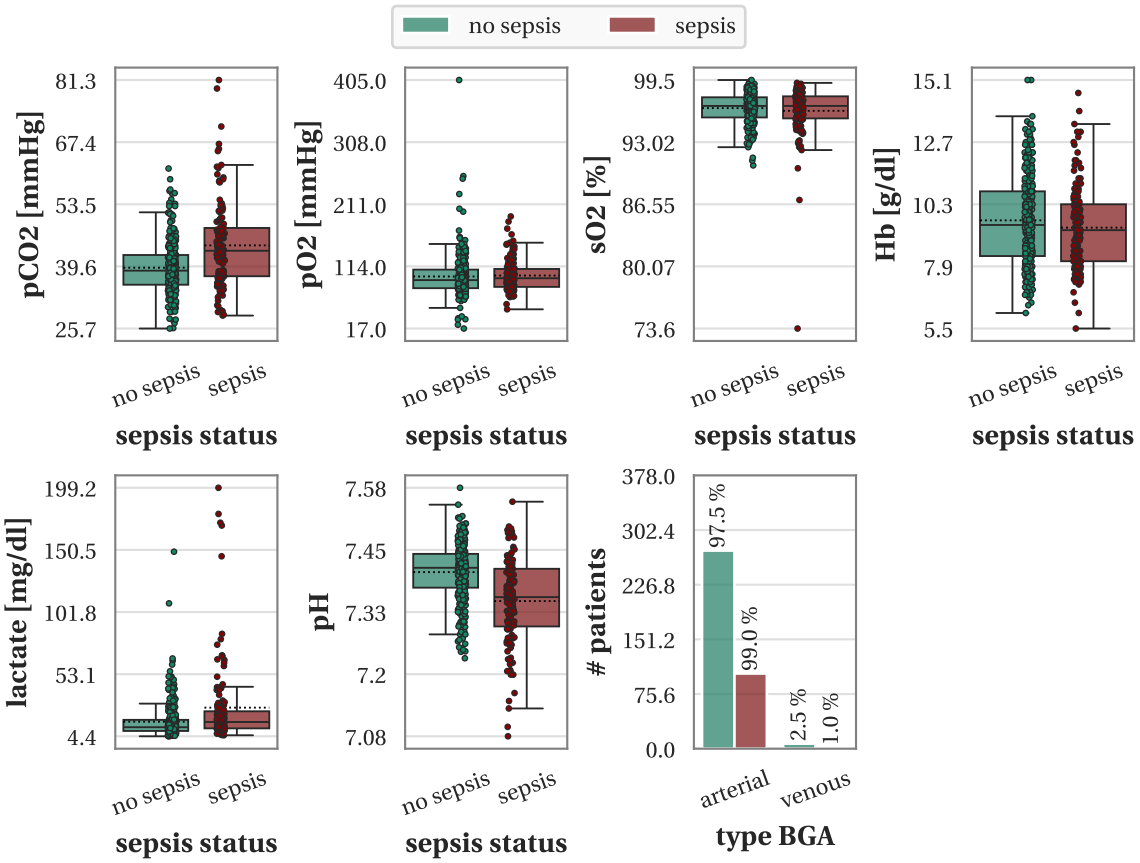


Figure B.20: Distribution of blood gas analysis (BGA) measurements among non-septic and septic patients. Shown are the distributions of carbon dioxide partial pressure (pCO2), oxygen partial pressure (pO2), oxygen saturation (sO2), hemoglobin (Hb), lactate, pH value and type of BGA (arterial or venous). For continuous parameters, boxplots indicate the quartiles, with whiskers extending to the range excluding outliers. Solid and dashed lines mark the median and mean, respectively. Each dot corresponds to one patient. For categorical parameters, bar plots display the number of patients (# patients) per category, with percentages given relative to all patients of the same sepsis status.

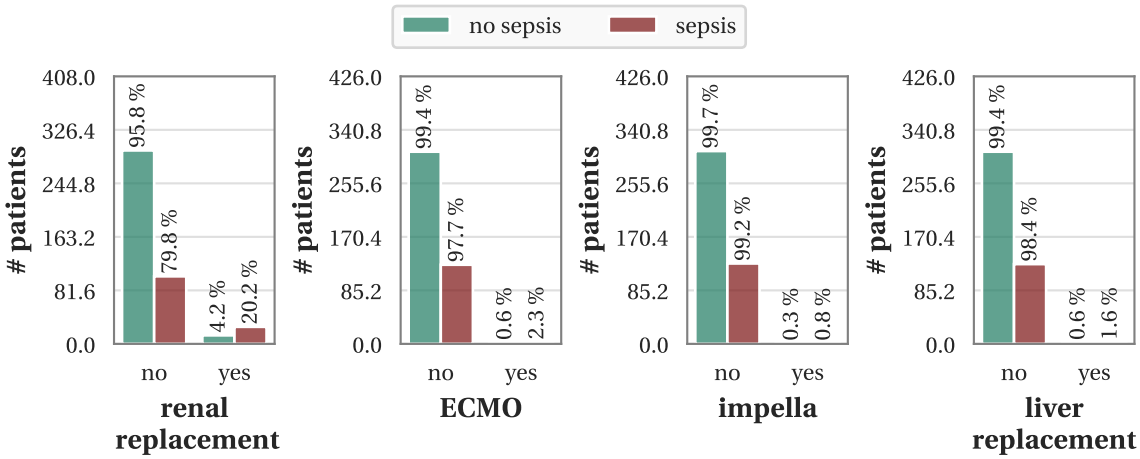


Figure B.21: Distribution of organ replacement therapies among non-septic and septic patients. Shown are the distributions of renal replacement, extracorporeal membrane oxygenation (ECMO), impella and liver replacement therapies. Bar plots display the number of patients (# patients) per category, with percentages given relative to all patients of the same sepsis status.

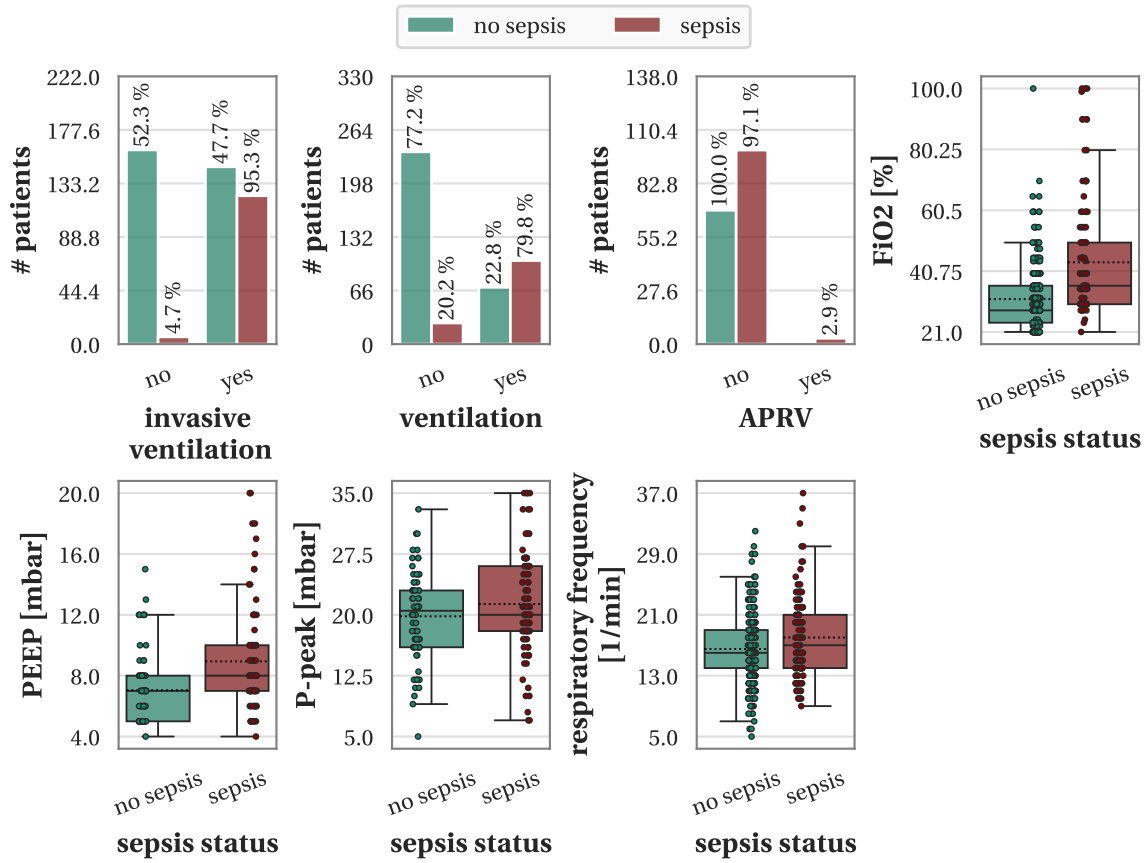


Figure B.22: Distribution of ventilation parameters among non-septic and septic patients. Shown are the distributions of invasive ventilation, ventilation, airway pressure release ventilation (APRV), fraction of inspired oxygen (FiO2), positive endexpiratory pressure (PEEP), peak inspiratory pressure (P-peak) and respiratory frequency. For continuous parameters, boxplots indicate the quartiles, with whiskers extending to the range excluding outliers. Solid and dashed lines mark the median and mean, respectively. Each dot corresponds to one patient. For categorical parameters, bar plots display the number of patients (# patients) per category, with percentages given relative to all patients of the same sepsis status.

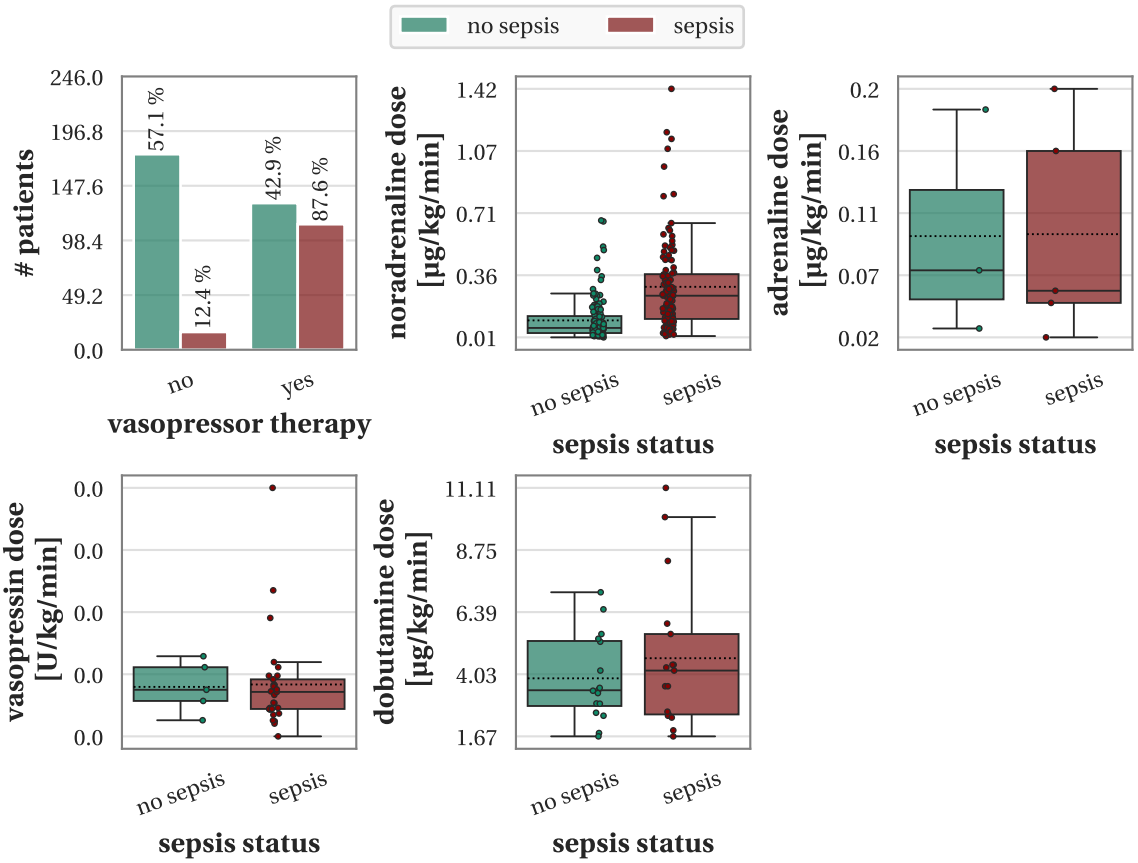


Figure B.23: Distribution of vasopressor/inotrope dosing among non-septic and septic patients. Shown are the distributions of whether vasopressors/inotropes were administered, and if so, the dosing of noradrenaline, adrenaline, vasopressin and dobutamine. For continuous parameters, boxplots indicate the quartiles, with whiskers extending to the range excluding outliers. Solid and dashed lines mark the median and mean, respectively. Each dot corresponds to one patient. For categorical parameters, bar plots display the number of patients (# patients) per category, with percentages given relative to all patients of the same sepsis status.

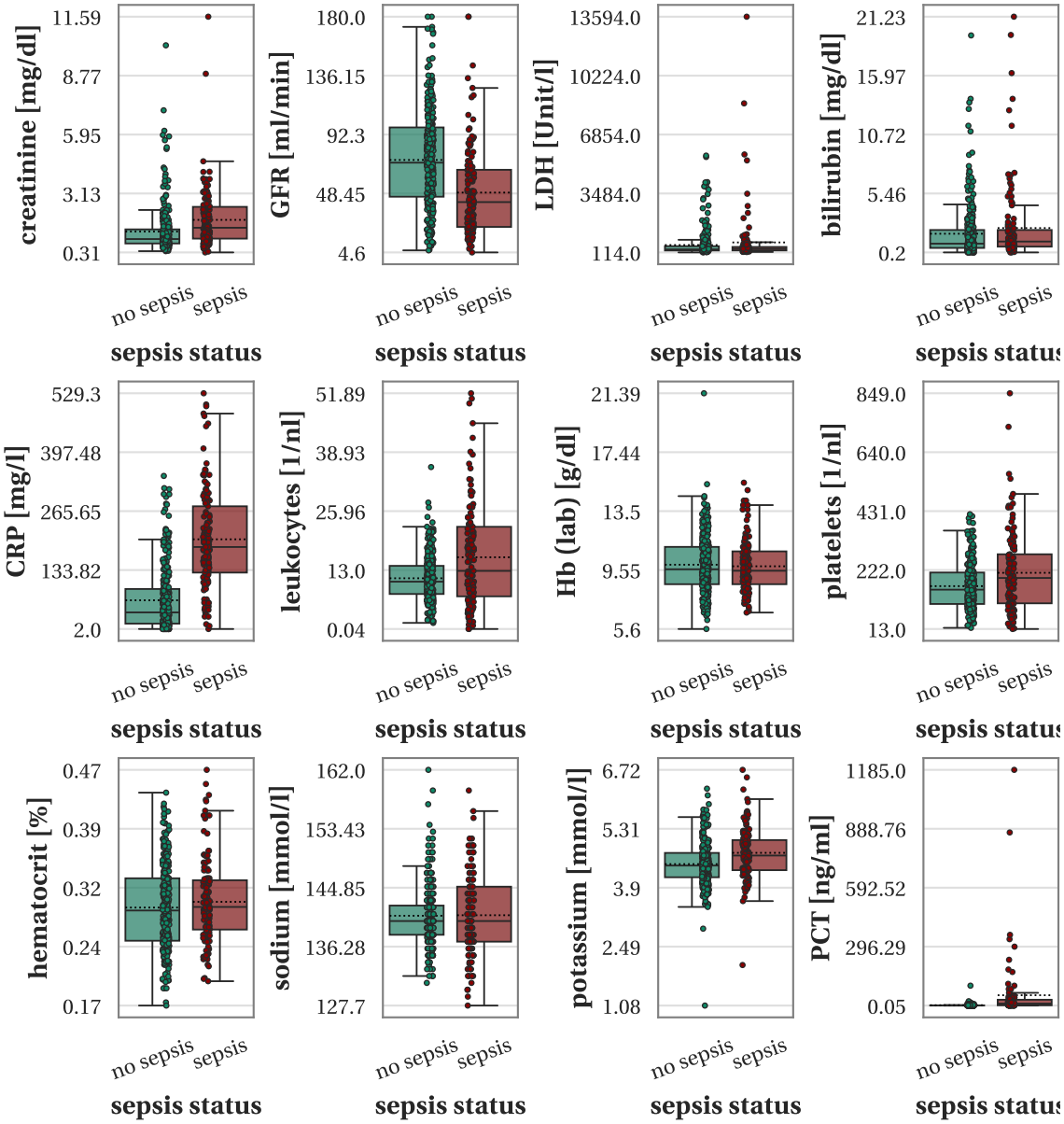


Figure B.24: Distribution of laboratory parameters among non-septic and septic patients. Shown are the distributions of creatinine, glomerular filtration rate (GFR), lactate dehydrogenase (LDH), bilirubin, C-reactive protein (CRP), leukocytes, hemoglobin (Hb), platelets, hematocrit, sodium, potassium and procalcitonin (PCT). Boxplots indicate the quartiles, with whiskers extending to the range excluding outliers. Solid and dashed lines mark the median and mean, respectively. Each dot corresponds to one patient.

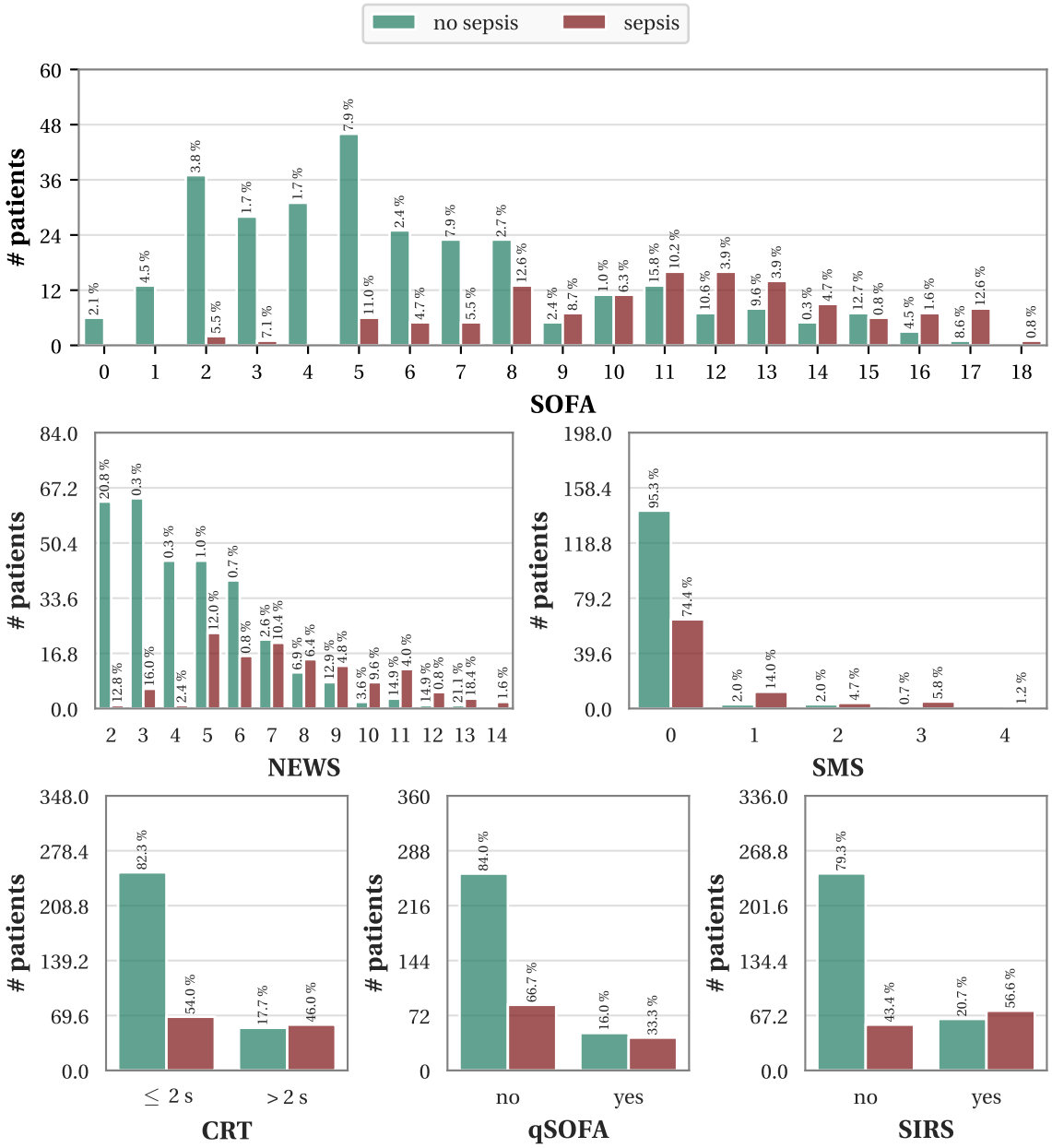


Figure B.25: Distribution of established diagnostic and prognostic clinical scores among non-septic and septic patients. Shown are the distributions of the Sequential Organ Failure Assessment (SOFA) score, national early warning score (NEWS), skin mottling score (SMS), capillary refill time (CRT), quick Sequential Organ Failure Assessment (qSOFA) score and Systemic Inflammatory Response Syndrome (SIRS) criteria. Bar plots display the number of patients (# patients) per category, with percentages given relative to all patients of the same sepsis status.

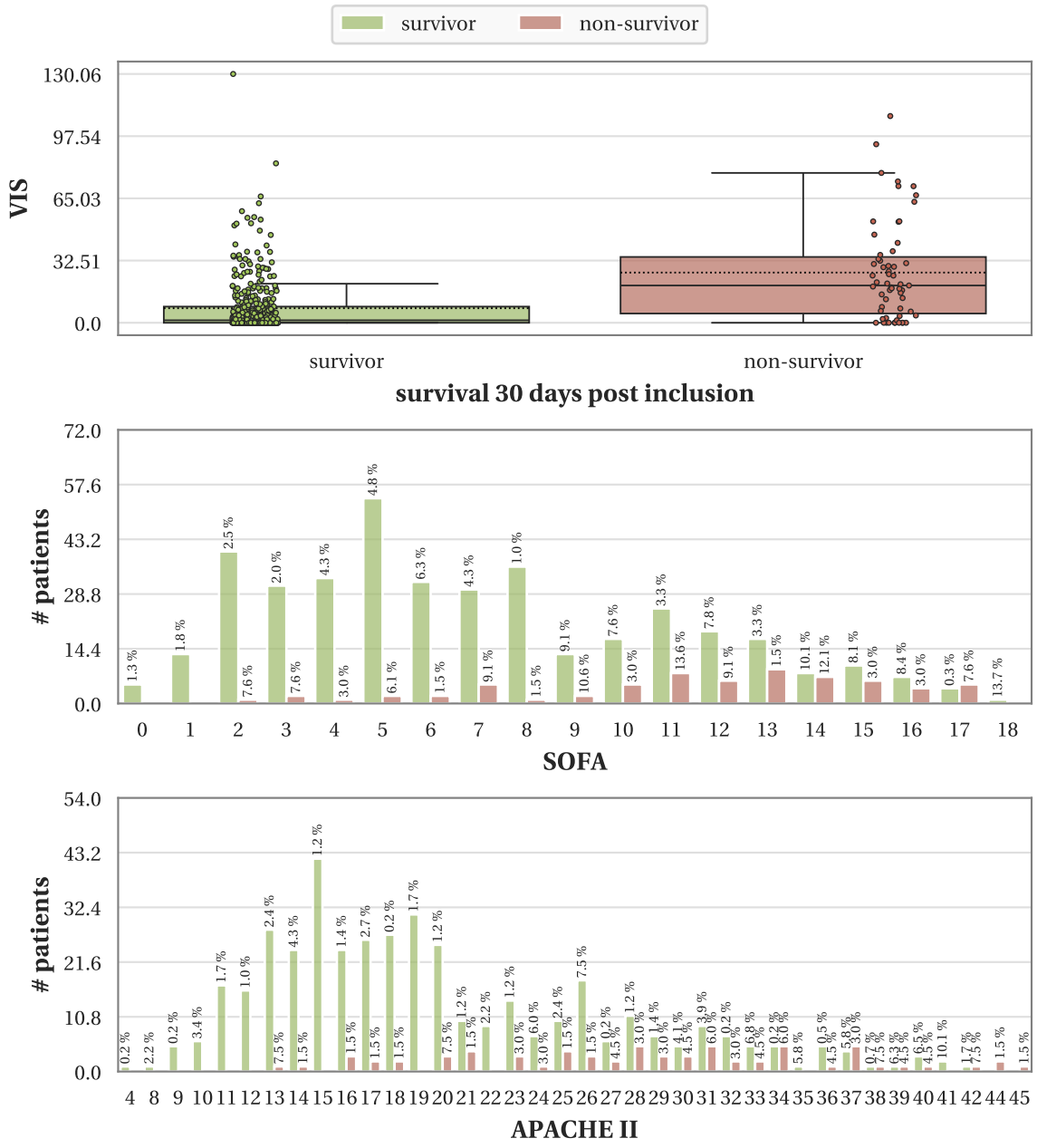


Figure B.26: Distribution of established diagnostic and prognostic clinical scores among survivors and non-survivors. Shown are the distributions of the vasoactive inotropic score (VIS), Sequential Organ Failure Assessment (SOFA) score and Acute Physiology and Chronic Health Evaluation (APACHE) II score. For continuous parameters, boxplots indicate the quartiles, with whiskers extending to the range excluding outliers. Solid and dashed lines mark the median and mean, respectively. Each dot corresponds to one patient. For categorical parameters, bar plots display the number of patients (# patients) per category, with percentages given relative to all patients of the same survival status.

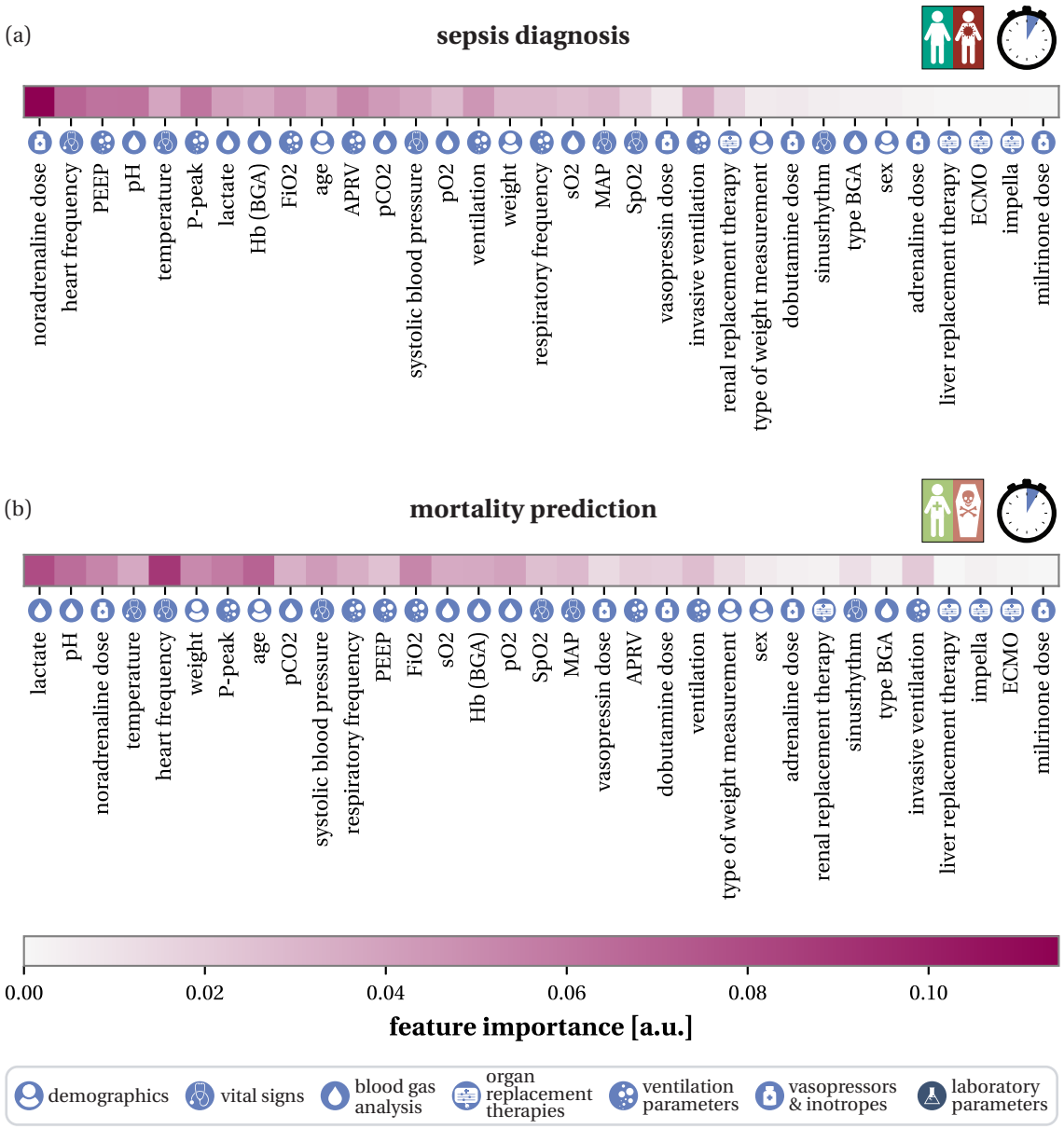


Figure B.27: Importance of clinical features available within 1 h of intensive care unit (ICU) admission for (a) sepsis diagnosis and (b) mortality prediction. Colors indicate the feature importance, measured as the reduction in Gini importance when a clinical feature is used for data splitting within a decision tree node of the clinical data model. Features are ordered from most to least important (left to right) based on recursive feature elimination [130] using the clinical data model, starting from the full set of clinical features available within 1 h of ICU admission. Figure adapted from [306].

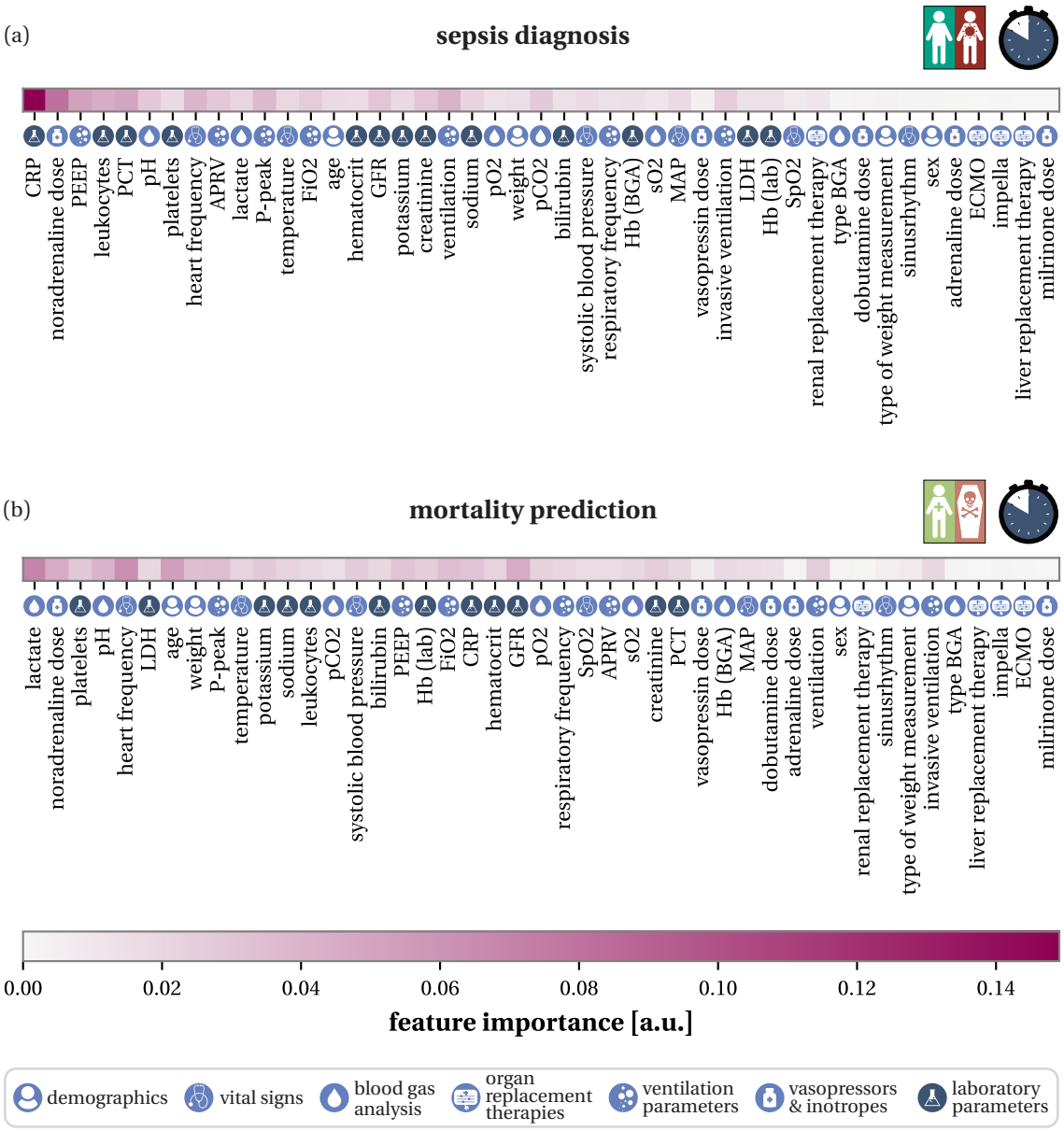


Figure B.28: Importance of clinical features available within 10 h of intensive care unit (ICU) admission for (a) sepsis diagnosis and (b) mortality prediction. Colors indicate the feature importance, measured as the reduction in Gini importance when a clinical feature is used for data splitting within a decision tree node of the clinical data model. Features are ordered from most to least important (left to right) based on recursive feature elimination [130] using the clinical data model, starting from the full set of clinical features available within 10 h of ICU admission. Figure adapted from [306].

List of Acronyms

HbO_2 oxyhemoglobin

SpO_2 pulse oxymetrical oxygen saturation

StO_2 tissue oxygen saturation

AI artificial intelligence

APACHE Acute Physiology and Chronic Health Evaluation

APRV airway pressure release ventilation

ASD average surface distance

AUROC area under the receiver operating characteristic curve

BGA blood gas analysis

CAMI Computer Assisted Medical Interventions

CCC computational color constancy

CE cross-entropy

CI confidence interval

CMOS Complementary Metal-Oxide-Semiconductor

CNN convolutional neural network

CPU central processing unit

CRP C-reactive protein

CRT capillary refill time

dHb deoxyhemoglobin

DKFZ German Cancer Research Center

DL deep learning

DOF degrees of freedom

DSC Dice similarity coefficient

ECMO extracorporeal membrane oxygenation

EHR electronic health record

ELU exponential linear unit

FiO2 fraction of inspired oxygen

GFR glomerular filtration rate

GPU graphics processing unit

Hb hemoglobin

HIDSS4Health Helmholtz Information & Data Science School for Health

ICU intensive care unit

IEEE Institute of Electrical and Electronics Engineers

IMSY Intelligent Medical Systems

IPCAI International Conference on Information Processing in Computer-Assisted Interventions

LDH lactate dehydrogenase

LeakyReLU leaky rectified linear unit

LED light-emitting diode

LMICs low- and middle-income countries

MAP mean arterial pressure

MICCAI International Conference on Medical Image Computing and Computer Assisted Intervention

ML machine learning

MONAI Medical Open Network for AI

MSI multispectral imaging

NEWS national early warning score

NIR near-infrared

NPI tissue perfusion index

NSD normalized surface Dice

OOD out-of-distribution

OR operating room

P-peak peak inspiratory pressure

pCO2 carbon dioxide partial pressure

PCT procalcitonin

PEEP positive endexpiratory pressure

pO2 oxygen partial pressure

qSOFA quick Sequential Organ Failure Assessment

ReLU rectified linear unit

ResNet residual network

RFE recursive feature elimination

ROC receiver operating characteristic

RQ research question

SD standard deviation

SIC sepsis-induced coagulopathy

SIRS Systemic Inflammatory Response Syndrome

SI spectral imaging

SLIC simple linear iterative clustering

SMS skin mottling score

sO2 oxygen saturation

SOFA Sequential Organ Failure Assessment

SWA stochastic weight averaging

THI tissue hemoglobin index

TLI tissue lipid index

TWI tissue water index

VIS vasoactive inotropic score

WHISPERS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing

HSI hyperspectral imaging

TPI tissue parameter images

List of Figures

List of Tables

Bibliography

[1] Accel Abarca and Albert Theuwissen. “A CMOS Image Sensor Dark Current Compensation Using In-Pixel Temperature Sensors”. In: *Sensors (Basel)* 23.22 (Nov. 2023). DOI: 10.3390/s23229109 (cit. on p. 80).

[2] Barsha Abhisheka, Saroj Kumar Biswas, Biswajit Purkayastha, Dolly Das, and Alexandre Escargueil. “Recent trend in medical imaging modalities and their applications in disease diagnosis: a review”. In: *Multimedia Tools and Applications* 83.14 (Apr. 1, 2024), pp. 43035–43070. ISSN: 1573-7721. DOI: 10.1007/s11042-023-17326-1 (cit. on p. 3).

[3] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. “SLIC Superpixels Compared to State-of-the-Art Superpixel Methods”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (Nov. 2012). Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 2274–2282. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2012.120 (cit. on p. 125).

[4] Kasper Adelborg, Julie B Larsen, and Anne-Mette Hvas. “Disseminated intravascular coagulation: epidemiology, biomarkers, and management”. In: *British Journal of Haematology* 192.5 (2021), pp. 803–818. DOI: 10.1111/bjh.17172 (cit. on p. 37).

[5] Mahmoud Afifi and Michael S Brown. “What else can fool deep learning? Addressing color constancy errors on deep neural network performance”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 243–252 (cit. on p. 58).

[6] Abien Fred Agarap. “Deep learning using rectified linear units (relu)”. In: *arXiv preprint arXiv:1803.08375* (2018) (cit. on p. 50).

[7] SuperAnnotate AI. *SuperAnnotate | Empowering Enterprises with Custom LLM/GenAI/CV Models*. URL: https://www.superannotate.com (visited on 11/17/2023) (cit. on p. 119).

[8] H Ait-Oufella, S Lemoinne, PY Boelle, A Galbois, JL Baudel, J Lemant, J Joffre, D Margetis, B Guidet, E Maury, et al. “Mottling score predicts survival in septic shock”. In: *Intensive care medicine* 37 (2011), pp. 801–807. DOI: 10.1007/s00134-011-2163-y (cit. on p. 182).

[9] Hamed Akbari, Yukio Kosugi, Kazuyuki Kojima, and Naofumi Tanaka. "Wavelet-Based Compression and Segmentation of Hyperspectral Images in Surgery". In: *Medical Imaging and Augmented Reality*. Ed. by Takeyoshi Dohi, Ichiro Sakuma, and Hongen Liao. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008, pp. 142–149. ISBN: 978-3-540-79982-5. DOI: 10.1007/978-3-540-79982-5_16 (cit. on pp. 115, 117, 118, 155).

[10] Hassan Al Hajj, Mathieu Lamard, Pierre-Henri Conze, Soumali Roychowdhury, Xiaowei Hu, Gabija Maralkait, Odysseas Zisimopoulos, Muneer Ahmad Dedmari, Fenqiang Zhao, Jonas Prellberg, Manish Sahu, Adrian Galdran, Teresa Araújo, Duc My Vo, Chandan Panda, Navdeep Dahiya, Satoshi Kondo, Zhengbing Bian, Arash Vahdat, Jonas Bialopetraviius, Evangello Flouty, Chenhui Qiu, Sabrina Dill, Anirban Mukhopadhyay, Pedro Costa, Guilherme Aresta, Senthil Ramamurthy, Sang-Woong Lee, Aurélio Campilho, Stefan Zachow, Shunren Xia, Sailesh Conjeti, Danail Stoyanov, Jogundas Armaitis, Pheng-Ann Heng, William G. Macready, Béatrice Cochener, and Gwenolé Quellec. "CATARACTS: Challenge on automatic tool annotation for cataRACT surgery". In: *Medical Image Analysis* 52 (Feb. 2019), pp. 24–41. ISSN: 1361-8415. DOI: 10.1016/j.media.2018.11.008 (cit. on p. 114).

[11] Erik Alerstam, William Chun Yip Lo, Tianyi David Han, Jonathan Rose, Stefan Andersson-Engels, and Lothar Lilge. "Next-generation acceleration and code optimization for light transport in turbid media using GPUs". In: *Biomed. Opt. Express* 1.2 (Sept. 2010), pp. 658–675. DOI: 10.1364/BOE.1.000658 (cit. on p. 27).

[12] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, Avinash Kori, Varghese Alex, Ganapathy Krishnamurthi, David Rauber, Robert Mendel, Christoph Palm, Sophia Bano, Guinther Saibro, Chi-Sheng Shih, Hsun-An Chiang, Juntang Zhuang, Junlin Yang, Vladimir Iglovikov, Anton Dobrenkii, Madhu Reddiboina, Anubhav Reddy, Xingtong Liu, Cong Gao, Mathias Unberath, Myeonghyeon Kim, Chanho Kim, Chaewon Kim, Hyejin Kim, Gyeongmin Lee, Ihsan Ullah, Miguel Luna, Sang Hyun Park, Mahdi Azizian, Danail Stoyanov, Lena Maier-Hein, and Stefanie Speidel. "2018 Robotic Scene Segmentation Challenge". In: *arXiv:2001.11190 [cs]* (Aug. 2020). arXiv: 2001.11190. URL: http://arxiv.org/abs/2001.11190 (cit. on pp. 114, 116, 155).

[13] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, Luis Herrera, Wenqi Li, Vladimir Iglovikov, Huoling Luo, Jian Yang, Danail Stoyanov, Lena Maier-Hein, Stefanie Speidel, and Mahdi Azizian. *2017 Robotic Instrument Segmentation Challenge*. 2019. arXiv: 1902.06426 [cs.CV] (cit. on p. 114).

[14] Khaled Alomar, Halil Ibrahim Aysel, and Xiaohao Cai. “Data Augmentation in Classification and Segmentation: A Survey and New Strategies”. In: *Journal of Imaging* 9.2 (Feb. 2023), p. 46. ISSN: 2313-433X. DOI: 10.3390/jimaging9020046 (cit. on p. 154).

[15] Alhanoof Althnian, Duaa AlSaeed, Heyam Al-Baity, Amani Samha, Alanoud Bin Dris, Najla Alzakari, Afnan Abou Elwafa, and Heba Kurdi. “Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain”. In: *Applied Sciences* 11.2 (2021). ISSN: 2076-3417. DOI: 10.3390/app11020796 (cit. on p. 219).

[16] Jacob Amersfoort, Guy Eelen, and Peter Carmeliet. “Immunomodulation by endothelial cellspartnering up with the immune system?” In: *Nature Reviews Immunology* 22.9 (2022), pp. 576–588. DOI: 10.1038/s41577-022-00694-4 (cit. on p. 36).

[17] Vijay Anand, Zilu Zhang, Sameer S Kadri, Michael Klompas, Chanu Rhee, CDC Prevention Epicenters Program, et al. “Epidemiology of quick sequential organ failure assessment criteria in undifferentiated patients and association with suspected infection and sepsis”. In: *Chest* 156.2 (2019), pp. 289–297. DOI: 10.1016/j.chest.2019.03.032 (cit. on p. 39).

[18] Nicola de Angelis, Fausto Catena, Riccardo Memeo, Federico Coccolini, Aleix Martínez-Pérez, Oreste M Romeo, Belinda De Simone, Salomone Di Saverio, Raffaele Brustia, Rami Rhaiem, et al. “2020 WSES guidelines for the detection and management of bile duct injury during cholecystectomy”. In: *World journal of emergency surgery* 16.1 (2021), p. 30. DOI: 10.1186/s13017-021-00369-w (cit. on p. 31).

[19] Jaskirat Arora, Asher A Mendelson, and Alison Fox-Robichaud. “Sepsis: network pathophysiology and implications for early diagnosis”. In: *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 324.5 (2023), R613–R624. DOI: 10.1152/ajpregu.00003.2023 (cit. on pp. 34, 36, 178).

[20] Caerwyn Ash, Michael Dubec, Kelvin Donne, and Tim Bashford. “Effect of wavelength and beam width on penetration in light-tissue interaction using computational methods”. In: *Lasers in Medical Science* 32.8 (Nov. 1, 2017), pp. 1909–1918. ISSN: 1435-604X. DOI: 10.1007/s10103-017-2317-4 (cit. on p. 20).

[21] Leonardo Ayala, Tim J. Adler, Silvia Seidlitz, Sebastian Wirkert, Christina Engels, Alexander Seitel, Jan Sellner, Alexey Aksenov, Matthias Bodenbach, Pia Bader, Sebastian Baron, Anant Vemuri, Manuel Wiesenfarth, Nicholas Schreck, Diana Mindroc, Minu Tizabi, Sebastian Pirmann, Brittaney Everitt, Annette Kopp-Schneider, Dogu Teber, and Lena Maier-Hein. “Spectral imaging enables contrast agent-free real-time ischemia monitoring in laparoscopic surgery”. In: *Science Advances* 9.10 (2023), eadd6778. DOI: 10.1126/sciadv.add6778 (cit. on pp. 4, 6, 22, 24, 27, 29, 216, 217).

[22] Leonardo Ayala, Fabian Isensee, Sebastian J Wirkert, Anant S Vemuri, Klaus H Maier-Hein, Baowei Fei, and Lena Maier-Hein. "Band selection for oxygenation estimation with multispectral/hyperspectral imaging". In: *Biomed Opt Express* 13.3 (Feb. 2022), pp. 1224–1242. DOI: 10.1364/BOE.441214 (cit. on p. 218).

[23] Leonardo Ayala, Diana Mindroc-Filimon, Maike Rees, Marco Hübner, Jan Sellner, Silvia Seidlitz, Minu Tizabi, Sebastian Wirkert, Alexander Seitel, and Lena Maier-Hein. "The SPECTRAL Perfusion Arm Clamping dAtaset (SPECTRAL-PACA) for video-rate functional imaging of the skin". In: *Scientific Data* 11.1 (May 25, 2024), p. 536. ISSN: 2052-4463. DOI: 10.1038/s41597-024-03307-y (cit. on p. 220).

[24] Leonardo Ayala, Silvia Seidlitz, Anant Vemuri, Sebastian J Wirkert, Thomas Kirchner, Tim J Adler, Christina Engels, Dogu Teber, and Lena Maier-Hein. "Light source calibration for multispectral imaging in surgery". In: *Int J Comput Assist Radiol Surg* 15.7 (June 2020), pp. 1117–1125. DOI: 10.1007/s11548-020-02195-y (cit. on pp. 25, 57, 62, 68, 69, 72–74, 213, 223).

[25] Leonardo Ayala, Sebastian J. Wirkert, Anant Vemuri, Tim Adler, Silvia Seidlitz, Sebastian Pirmann, Christina Engels, Dogu Teber, and Lena Maier-Hein. "Video-rate multispectral imaging in laparoscopic surgery: First-in-human application". In: (2021). DOI: 10.48550/arXiv.2105.13901 (cit. on pp. 21, 217).

[26] Leonardo Antonio Ayala Menjivar. "Translational Functional Imaging in Surgery Enabled by Deep Learning". PhD thesis. 2023. DOI: 10.11588/heidok.00033281 (cit. on pp. 57, 68, 69, 72–74).

[27] Nicolás Ayobi, Santiago Rodríguez, Alejandra Pérez, Isabela Hernández, Nicolás Aparicio, Eugénie Dessevres, Sebastián Peña, Jessica Santander, Juan Ignacio Caicedo, Nicolás Fernández, et al. "Pixel-Wise Recognition for Holistic Surgical Scene Understanding". In: *arXiv preprint arXiv:2401.11174* (2024). DOI: 10.48550/arXiv.2401.11174 (cit. on p. 114).

[28] Marcus A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin S. Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Bethany Percha, Thomas M. Snyder, and Joel T. Dudley. "Deep learning predicts hip fracture using confounding patient and healthcare variables". In: *npj Digital Medicine* 2.1 (Apr. 2019), pp. 1–10. ISSN: 2398-6352. DOI: 10.1038/s41746-019-0105-1. (Visited on 05/19/2021) (cit. on pp. 80, 109).

[29] Wesley B Baker, Ashwin B Parthasarathy, David R Busch, Rickson C Mesquita, Joel H Greenberg, and AG Yodh. "Modified Beer-Lambert law for blood flow". In: *Biomedical optics express* 5.11 (2014), pp. 4053–4075. DOI: 10.1364/BOE.5.004053 (cit. on p. 26).

[30] Elisa Bannone, Toby Collins, Alessandro Esposito, Lorenzo Cinelli, Matteo De Pastena, Patrick Pessaux, Emanuele Felli, Elena Andreotti, Nariaki Okamoto, Manuel Barberio, et al. “Surgical optomics: hyperspectral imaging and deep learning towards precision intraoperative automatic tissue recognitionresults from the EX-MACHYNA trial”. In: *Surgical endoscopy* (2024), pp. 1–15. DOI: 10.1007/s00464-024-10880-1 (cit. on pp. 115, 117, 155, 215, 220).

[31] Wolfgang Banzhaf, Penousal Machado, and Mengjie Zhang. *Handbook of Evolutionary Machine Learning*. Springer, 2023. DOI: 10.1007/978-981-99-3814-8 (cit. on p. 46).

[32] Emily S Bartlett, Andrew Lim, Sean Kivlehan, Lia I Losonczy, Srinivas Murthy, Richard Lowsby, Alfred Papali, Madiha Raees, Bhavna Seth, Natalie Cobb, et al. “Critical care delivery across health care systems in low-income and low-middle-income country settings: A systematic review”. In: *Journal of Global Health* 13 (2023). DOI: 10.7189/jogh.13.04141 (cit. on p. 205).

[33] Philip Baum, Johannes Diers, Sven Lichthardt, Carolin Kastner, Nicolas Schlegel, Christoph-Thomas Germer, and Armin Wiegering. “Mortality and Complications Following Visceral Surgery: A Nationwide Analysis Based on the Diagnostic Categories Used in German Hospital Invoicing Data”. In: *Deutsches Arzteblatt international* 116.44 (Nov. 2019), pp. 739–746. ISSN: 1866-0452. DOI: 10.3238/arztebl.2019.0739. URL: https://europepmc.org/articles/PMC6912125 (cit. on p. 28).

[34] Alexander Baumann, Leonardo Ayala, Alexander Studier-Fischer, Jan Sellner, Berkin Özdemir, Karl-Friedrich Kowalewski, Slobodan Ilic, Silvia Seidlitz, and Lena Maier-Hein. “Deep intra-operative illumination calibration of hyperspectral cameras”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Cham: Springer Nature Switzerland, 2024. DOI: 10.1007/978-3-031-72089-5_12 (cit. on pp. 58, 76, 80, 172).

[35] Jan Behmann, Kelvin Acebron, Dzhaner Emin, Simon Bennertz, Shizue Matsubara, Stefan Thomas, David Bohnenkamp, Matheus T. Kuska, Jouni Jussila, Harri Salo, Anne-Katrin Mahlein, and Uwe Rascher. “Specim IQ: Evaluation of a New, Miniaturized Handheld Hyperspectral Camera and Its Application for Plant Phenotyping and Disease Detection”. In: *Sensors* 18.2 (2018). ISSN: 1424-8220. DOI: 10.3390/s18020441. URL: https://www.mdpi.com/1424-8220/18/2/441 (cit. on p. 218).

[36] Eyal Bercovich and Marcia C Javitt. “Medical Imaging: From Roentgen to the Digital Revolution, and Beyond”. In: *Rambam Maimonides Med J* 9.4 (Oct. 2018). DOI: 10.5041/RMMJ.10355 (cit. on p. 3).

[37] Olivier Bertrand and Catherine Tallon-Baudry. "Oscillatory gamma activity in humans: a possible role for object representation". In: *International Journal of Psychophysiology* 38.3 (2000), pp. 211–223. DOI: 10.1016/S0167-8760(00)00166-5 (cit. on p. 58).

[38] Binod Bhattarai, Ronast Subedi, Rebati Raman Gaire, Eduard Vazquez, and Danail Stoyanov. "Histogram of Oriented Gradients meet deep learning: A novel multi-task deep network for 2D surgical image semantic segmentation". In: *Medical Image Analysis* 85 (2023), p. 102747. ISSN: 1361-8415. DOI: 10.1016/j.media.2023.102747 (cit. on pp. 116, 155).

[39] Simone Bianco, Claudio Cusano, and Raimondo Schettini. "Color constancy using CNNs". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2015, pp. 81–89 (cit. on p. 58).

[40] Irving J. Bigio and Sergio Fantini. *Quantitative Biomedical Optics: Theory, Methods, and Applications*. Cambridge Texts in Biomedical Engineering. Cambridge University Press, 2016. DOI: 10.1017/CBO9781139029797 (cit. on p. 15).

[41] Elizabeth Bilevicius, Desanka Dragosavac, Sanja Dragosavac, Sebastião Araújo, Antonio LE Falcão, and Renato GG Terzi. "Multiple organ failure in septic patients". In: *Brazilian Journal of Infectious Diseases* 5 (2001), pp. 103–110. DOI: 10.1590/s1413-86702001000300001 (cit. on p. 37).

[42] Frank Bloos. "Diagnosis and therapy of sepsis". In: *Journal of Emergency and Critical Care Medicine* 2.1 (2018). DOI: 10.21037/jeccm.2017.12.08 (cit. on pp. 12, 179).

[43] Sebastian Bodenstedt, Max Allan, Anthony Agustinos, Xiaofei Du, Luis Garcia-Peraza-Herrera, Hannes Kenngott, Thomas Kurmann, Beat Müller-Stich, Sebastien Ourselin, Daniil Pakhomov, et al. "Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery". In: *arXiv preprint arXiv:1805.02475* (2018). DOI: 10.48550/arXiv.1805.02475 (cit. on p. 114).

[44] Sebastian Bodenstedt, Dominik Rivoir, Alexander Jenke, Martin Wagner, Michael Breucha, Beat Müller-Stich, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. "Active learning using deep Bayesian networks for surgical workflow analysis". In: *International Journal of Computer Assisted Radiology and Surgery* 14.6 (June 1, 2019), pp. 1079–1087. ISSN: 1861-6429. DOI: 10.1007/s11548-019-01963-9 (cit. on p. 220).

[45] Rositsa Bogdanova, Pierre Boulanger, and Bin Zheng. "Depth Perception of Surgeons in Minimally Invasive Surgery". In: *Surg Innov* 23.5 (Mar. 2016), pp. 515–524. DOI: 10.1177/1553350616639141 (cit. on p. 28).

[46] Roger C Bone, Robert A Balk, Frank B Cerra, R Phillip Dellinger, Alan M Fein, William A Knaus, Roland MH Schein, and William J Sibbald. “Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis”. In: *Chest* 101.6 (1992), pp. 1644–1655. DOI: 10.1378/chest.101.6.1644 (cit. on p. 182).

[47] Carlo E Bonferroni. “Il calcolo delle assicurazioni su gruppi di teste”. In: *Studi in onore del professore salvatore ortu carboni* (1935), pp. 13–60 (cit. on p. 192).

[48] Nada N Boustany, Stephen A Boppart, and Vadim Backman. “Microscopic imaging and spectroscopy with scattered light”. In: *Annual review of biomedical engineering* 12.1 (2010), pp. 285–314. DOI: 10.1146/annurev-bioeng-061008-124811 (cit. on p. 17).

[49] Sakshi Bramhe and Swanand S Pathak. “Robotic Surgery: A Narrative Review”. In: *Cureus* 14.9 (Sept. 2022), e29179. DOI: 10.7759/cureus.29179 (cit. on p. 28).

[50] Leo Breiman. “Bagging predictors”. In: *Machine learning* 24 (1996), pp. 123–140. DOI: 10.1007/BF00058655 (cit. on p. 48).

[51] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324 (cit. on p. 44).

[52] Gershon Buchsbaum. “A spatial processor model for object colour perception”. In: *Journal of the Franklin institute* 310.1 (1980), pp. 1–26. DOI: 10.1016/0016-0032(80)90058-7 (cit. on p. 59).

[53] Wilhelm Burger and Mark J. Burge. *Digital Image Processing: An Algorithmic Introduction*. Springer International Publishing, 2022. ISBN: 978-3-031-05744-1. DOI: 10.1007/978-3-031-05744-1 (cit. on p. 49).

[54] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. “Albumentations: Fast and Flexible Image Augmentations”. In: *Information* 11.2 (Feb. 2020), p. 125. DOI: 10.3390/info11020125. URL: https://www.mdpi.com/2078-2489/11/2/125 (cit. on p. 127).

[55] Min Cao, Guozheng Wang, and Jianfeng Xie. “Immune dysregulation in sepsis: experiences, lessons and perspectives”. In: *Cell death discovery* 9.1 (2023), p. 465. DOI: 10.1038/s41420-023-01766-7 (cit. on p. 34).

[56] M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murray, Andriy Myronenko, Can Zhao, Dong Yang, Vishwesh Nath, Yufan He, Ziyue Xu, Ali Hatamizadeh, Wentao Zhu, Yun Liu, Mingxin Zheng, Yucheng Tang, Isaac Yang, Michael Zephyr, Behrooz Hashemian, Sachidanand Alle, Mohammad Zalbagi Darestani, Charlie Budd, Marc Modat, Tom Vercauteren, Guotai Wang, Yiwen Li, Yipeng Hu, Yunguan Fu, Benjamin Gorman, Hans Johnson, Brad Genereaux, Barbaros S. Erdal, Vikash Gupta, Andres

Diaz-Pinto, Andre Dourson, Lena Maier-Hein, Paul F. Jaeger, Michael Baumgartner, Jayashree Kalpathy-Cramer, Mona Flores, Justin Kirby, Lee A.D. Cooper, Holger R. Roth, Daguang Xu, David Bericat, Ralf Floca, S. Kevin Zhou, Haris Shuaib, et al. "MONAI: An open-source framework for deep learning in healthcare". In: (Nov. 2022). DOI: 10.48550/arXiv.2211.02701 (cit. on pp. 144, 215).

[57] Claudio Carini and Attila A. Seyhan. "Tribulations and future opportunities for artificial intelligence in precision medicine". In: *Journal of Translational Medicine* 22.1 (Apr. 30, 2024), p. 411. ISSN: 1479-5876. DOI: 10.1186/s12967-024-05067-0 (cit. on p. 218).

[58] Jane E Carré, Jean-Christophe Orban, Lorenza Re, Karen Felsmann, Wiebke Iffert, Michael Bauer, Hagir B Suliman, Claude A Piantadosi, Terry M Mayhew, Patrick Breen, et al. "Survival in critical illness is associated with early activation of mitochondrial biogenesis". In: *American journal of respiratory and critical care medicine* 182.6 (2010), pp. 745–751. DOI: 10.1164/rccm.201003-0326OC (cit. on p. 37).

[59] Fernando Cervantes-Sanchez, Marianne Maktabi, Hannes Köhler, Robert Sucher, Nada Rayes, Juan Gabriel Avina-Cervantes, Ivan Cruz-Aceves, and Claire Chalopin. "Automatic tissue segmentation of hyperspectral images in liver and head neck surgeries using machine learning". In: *Artificial Intelligence Surgery* 1 (Aug. 2021), pp. 22–37. DOI: 10.20517/ais.2021.05. URL: https://aisjournal.net/article/view/4291 (cit. on pp. 115, 117, 118, 138, 155).

[60] Claire Chalopin, Felix Nickel, Annekatrin Pfahl, Hannes Köhler, Marianne Maktabi, René Thieme, Robert Sucher, Boris Jansen-Winkeln, Alexander Studier-Fischer, Silvia Seidlitz, Lena Maier-Hein, Thomas Neumuth, Andreas Melzer, Beat Peter Müller-Stich, and Ines Gockel. "Artificial intelligence and hyperspectral imaging for image-guided assistance in minimally invasive surgery". In: *Chirurgie (Heidelberg, Germany)* 93.10 (Oct. 2022), pp. 940–947. ISSN: 2731-6971. DOI: 10.1007/s00104-022-01677-w (cit. on p. 29).

[61] Zitian Chen, Yanwei Fu, Kaiyu Chen, and Yu-Gang Jiang. "Image Block Augmentation for One-Shot Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 3379–3386. ISSN: 2374-3468, 2159-5399 (cit. on pp. 154, 157).

[62] Zhenxing Cheng, Simon T Abrams, Julien Toh, Susan Siyu Wang, Zhi Wang, Qian Yu, Weiping Yu, Cheng-Hock Toh, and Guozheng Wang. "The critical roles and mechanisms of immune cell death in sepsis". In: *Frontiers in immunology* 11 (2020), p. 1918. DOI: 10.3389/fimmu.2020.01918 (cit. on p. 42).

[63] Niki Christou, Alexia Roux-David, David N Naumann, Stephane Bouvier, Thibaud Rivaille, Sophiane Derbal, Abdelkader Taibi, Anne Fabre, Fabien Fredon, Sylvaine Durand-Fontanier, et al. "Bile duct injury during cholecystectomy:

necessity to learn how to do and interpret intraoperative cholangiography". In: *Frontiers in medicine* 8 (2021), p. 637987. DOI: 10.3389/fmed.2021.637987 (cit. on p. 31).

[64] William Chu, Avinash Chennamsetty, Robert Toroussian, and Clayton Lau. "Anaphylactic shock after intravenous administration of indocyanine green during robotic partial nephrectomy". In: *Urology case reports* 12 (2017), pp. 37–38. DOI: 10.1016/j.eucr.2017.02.006 (cit. on p. 29).

[65] Neil T. Clancy, Geoffrey Jones, Lena Maier-Hein, Daniel S. Elson, and Danail Stoyanov. "Surgical spectral imaging". In: *Medical Image Analysis* 63 (2020), p. 101699. ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101699. URL: https://www.sciencedirect.com/science/article/pii/S1361841520300645 (cit. on pp. 3, 20, 21, 216).

[66] Ela Claridge and Dena Hidovi-Rowe. "Model based inversion for deriving maps of histological parameters characteristic of cancer from ex-vivo multispectral images of the colon". In: *IEEE Trans Med Imaging* 33.4 (Nov. 2013), pp. 822–835. DOI: 10.1109/TMI.2013.2290697 (cit. on p. 25).

[67] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)". In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2016. URL: http://arxiv.org/abs/1511.07289 (cit. on pp. 50, 51, 125, 188).

[68] Jonathan Cohen, Jean-Louis Vincent, Neill KJ Adhikari, Flavia R Machado, Derek C Angus, Thierry Calandra, Katia Jaton, Stefano Giulieri, Julie Delaloye, Steven Opal, et al. "Sepsis: a roadmap for future research". In: *The Lancet infectious diseases* 15.5 (2015), pp. 581–614. DOI: 10.1016/S1473-3099(15)70112-X (cit. on pp. 37, 38).

[69] Toby Collins, Adrien Bartoli, Nicolas Bourdel, and Michel Canis. "Segmenting the Uterus in Monocular Laparoscopic Images without Manual Input". In: Cham: Springer, 2015. DOI: 10.1007/978-3-319-24574-4_22 (cit. on pp. 114–116, 155).

[70] Toby Collins, Marianne Maktabi, Manuel Barberio, Valentin Bencteux, Boris Jansen-Winkeln, Claire Chalopin, Jacques Marescaux, Alexandre Hostettler, Michele Diana, and Ines Gockel. "Automatic recognition of colon and esophagogastric cancer with machine learning and hyperspectral imaging". In: *Diagnostics* 11.10 (2021), p. 1810. DOI: 10.3390/diagnostics11101810 (cit. on pp. 115, 117, 118, 138, 155).

[71] Jhonatan Contreras and Thomas Bocklitz. "Explainable artificial intelligence for spectroscopy data: a review". In: *Pflügers Archiv - European Journal of Physiology* (Aug. 1, 2024). ISSN: 1432-2013. DOI: 10.1007/s00424-024-02997-y (cit. on p. 219).

[72] NVIDIA Corporation. *Convolutional Layers User's Guide*. Feb. 1, 2023. URL: https://docs.nvidia.com/deeplearning/performance/dl-performance-convolutional/index.html (visited on 02/08/2024) (cit. on p. 50).

[73] Rong Cui, He Yu, Tingfa Xu, Xiaoxue Xing, Xiaorui Cao, Kang Yan, and Jiexi Chen. "Deep learning in medical hyperspectral images: A review". In: *Sensors* 22.24 (2022), p. 9790. DOI: 10.3390/s22249790 (cit. on pp. 21, 58, 215, 218, 220).

[74] Partha Das, Yang Liu, Sezer Karaoglu, and Theo Gevers. "Generative models for multi-illumination color constancy". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1194–1203 (cit. on p. 58).

[75] Matthew Davies, Mary B. Stuart, Matthew J. Hobbs, Andrew J. S. McGonigle, and Jon R. Willmott. "Image Correction and In Situ Spectral Calibration for Low-Cost, Smartphone Hyperspectral Imaging". In: *Remote Sensing* 14.5 (2022). ISSN: 2072-4292. DOI: 10.3390/rs14051152 (cit. on pp. 80, 218).

[76] Daniel De Backer, Katia Donadello, Yasser Sakr, Gustavo Ospina-Tascon, Diamantino Salgado, Sabino Scolletta, and Jean-Louis Vincent. "Microcirculatory alterations in patients with severe sepsis: impact of time of assessment and relationship with outcome". In: *Critical care medicine* 41.3 (2013), pp. 791–799. DOI: 10.1097/CCM.0b013e3182742e8b (cit. on p. 178).

[77] David T Delpy, Mark Cope, Pieter van der Zee, Simon Arridge, Susan Wray, and JS Wyatt. "Estimation of optical pathlength through tissue from direct time of flight measurement". In: *Physics in Medicine & Biology* 33.12 (1988), p. 1433. DOI: 10.1088/0031-9155/33/12/008 (cit. on p. 26).

[78] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919. June 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848 (cit. on pp. 126, 219).

[79] Lee R. Dice. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3 (1945), pp. 297–302. ISSN: 00129658, 19399170. URL: http://www.jstor.org/stable/1932409 (cit. on p. 129).

[80] Maximilian Dietrich, S. Marx, M. von der Forst, T. Bruckner, F. C. F. Schmitt, M. O. Fiedler, F. Nickel, A. Studier-Fischer, B. P. Müller-Stich, T. Hackert, T. Brenner, M. A. Weigand, F. Uhle, and K. Schmidt. "Bedside hyperspectral imaging indicates a microcirculatory sepsis pattern - an observational study". In: *Microvascular Research* 136.104164 (July 2021). ISSN: 0026-2862. DOI: 10.1016/j.mvr.2021.104164. (Visited on 05/01/2021) (cit. on pp. 108, 179, 184, 198).

[81] Maximilian Dietrich, Sebastian Marx, Thomas Bruckner, Felix Nickel, Beat Peter Müller-Stich, Thilo Hackert, Markus A. Weigand, Florian Uhle, Thorsten Brenner, and Karsten Schmidt. "Bedside hyperspectral imaging for the evaluation of microcirculatory alterations in perioperative intensive care medicine: a study protocol for an observational clinical pilot study (HySpI-ICU)". In: *BMJ Open* 10.9 (Sept. 2020). ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2019-035742. (Visited on 12/10/2020) (cit. on pp. 80, 184, 198).

[82] Maximilian Dietrich, Sebastian Marx, Maik von der Forst, Thomas Bruckner, Felix C F Schmitt, Mascha O. Fiedler, Felix Nickel, Alexander Studier-Fischer, Beat P. Mueller-Stich, Tilo Hackert, Thorsten Brenner, Markus A. Weigand, Florian Uhle, and Karsten Schmidt. "Hyperspectral Imaging for Perioperative Monitoring of Microcirculatory Tissue Oxygenation and Tissue Water Content in Pancreatic Surgery An Observational Clinical Pilot Study". In: *mansucript under review in Perioperative Medicine (PERI-D-20-00132)* (Dec. 2020) (cit. on p. 184).

[83] Maximilian Dietrich, Berkin Özdemir, Daniel Gruneberg, Clara Petersen, Alexander Studier-Fischer, Maik von der Forst, Felix C. F. Schmitt, Mascha O. Fiedler, Felix Nickel, Beat Peter Müller-Stich, Thorsten Brenner, Markus A. Weigand, Florian Uhle, and Karsten Schmidt. "Hyperspectral Imaging for the Evaluation of Microcirculatory Tissue Oxygenation and Perfusion Quality in Haemorrhagic Shock: A Porcine Study". In: *Biomedicines* 9.12 (2021). ISSN: 2227-9059. DOI: 10.3390/biomedicines9121829 (cit. on pp. 6, 80, 206).

[84] Maximilian Dietrich, Silvia Seidlitz, Katharina Hölzl, Ayca von Garrel, Jan Sellner, Stephan Katzenschlager, Tobias Hölle, Dania Fischer, Maik von der Forst, Felix C. F. Schmitt, Alexander Studier-Fischer, Markus A. Weigand, and Lena Maier-Hein. *AI-Based Spectral Imaging Biomarkers for Recognizing Sepsis and Predicting Mortality in Intensive Care: A Prospective Study of 483 Critically Ill Patients*. Oral presentation at the 5th Conference on Clinical Translation of Medical Image Computing and Computer Assisted Intervention (CLINICCAI), Daejeon, South Korea. Sept. 25, 2025 (cit. on p. 214).

[85] Maximilian Dietrich, Silvia Seidlitz, Nicholas Schreck, Manuel Wiesenfarth, Patrick Godau, Minu Tizabi, Jan Sellner, Sebastian Marx, Samuel Knödler, Michael M. Allers, Leonardo Ayala, Karsten Schmidt, Thorsten Brenner, Alexander Studier-Fischer, Felix Nickel, Beat P. Müller-Stich, Annette Kopp-Schneider, Markus A. Weigand, and Lena Maier-Hein. "Machine learning-based analysis of hyperspectral images for automated sepsis diagnosis". In: *arXiv preprint arXiv:2106.08445* (2021). DOI: 10.48550/arXiv.2106.08445 (cit. on pp. 12, 177, 179, 184, 198).

[86] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. *Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection*. Aug. 2017 (cit. on pp. 154, 157).

[87] Marc Ebner. "Color Constancy". In: *Computer Vision: A Reference Guide*. Ed. by Katsushi Ikeuchi. Cham: Springer International Publishing, 2021, pp. 168–175. ISBN: 978-3-030-63416-2. DOI: 10.1007/978-3-030-63416-2_454 (cit. on pp. 58, 59).

[88] Christoph Engel, Frank M Brunkhorst, Hans-Georg Bone, Reinhard Brunkhorst, Herwig Gerlach, Stefan Grond, Matthias Gruendling, Guenter Huhle, Ulrich Jaschinski, Stefan John, et al. "Epidemiology of sepsis in Germany: results from a national prospective multicenter study". In: *Intensive care medicine* 33 (2007), pp. 606–618. DOI: 10.1007/s00134-006-0517-7 (cit. on p. 34).

[89] Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M Coopersmith, Craig French, Flávia R Machado, Lauralyn Mcintyre, Marlies Ostermann, Hallie C Prescott, et al. "Executive summary: surviving sepsis campaign: international guidelines for the management of sepsis and septic shock 2021". In: *Critical care medicine* 49.11 (2021), pp. 1974–1982. DOI: 10.1097/CCM.0000000000005357 (cit. on pp. 41, 42).

[90] Tom Evans. "Diagnosis and management of sepsis". In: *Clinical Medicine* 18.2 (2018), p. 146. DOI: 10.7861/clinmedicine.18-2-146 (cit. on pp. 34, 36).

[91] Maria Ewerlöf, Marcus Larsson, and E. Göran Salerud. "Spatial and temporal skin blood volume and saturation estimation using a multispectral snapshot imaging camera". In: *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XV*. Ed. by Daniel L. Farkas, Dan V. Nicolau, and Robert C. Leif. SPIE, Feb. 2017. DOI: 10.1117/12.2251928 (cit. on p. 25).

[92] Himar Fabelo, Martin Halicek, Samuel Ortega, Adam Szolna, Jesus Morera, Roberto Sarmiento, Gustavo M. Callico, and Baowei Fei. "Surgical Aid Visualization System for Glioblastoma Tumor Identification based on Deep Learning and In-Vivo Hyperspectral Images of Human Patients". In: *Proceedings of SPIE–the International Society for Optical Engineering* 10951 (Feb. 2019), p. 1095110. ISSN: 0277-786X. DOI: 10.1117/12.2512569 (cit. on pp. 115, 117, 118, 155).

[93] Himar Fabelo, Samuel Ortega, Silvester Kabwama, Gustavo M. Callico, Diederik Bulters, Adam Szolna, Juan F. Pineiro, and Roberto Sarmiento. "HELICoiD project: a new use of hyperspectral imaging for brain cancer detection in real-time during neurosurgical operations". In: *Hyperspectral Imaging Sensors: Innovative Applications and Sensor Standards 2016*. Vol. 9860. SPIE, May 2016, p. 986002. DOI: 10.1117/12.2223075 (cit. on pp. 117, 118).

[94] Himar Fabelo, Samuel Ortega, Daniele Ravi, B. Ravi Kiran, Coralia Sosa, Diederik Bulters, Gustavo M. Callicó, Harry Bulstrode, Adam Szolna, Juan F. Piñeiro, Silvester Kabwama, Daniel Madroñal, Raquel Lazcano, Aruma J-OShanahan, Sara Bisshopp, María Hernández, Abelardo Báez, Guang-Zhong Yang, Bogdan Stanciulescu, Rubén Salvador, Eduardo Juárez, and Roberto Sarmiento. "Spatio-spectral classification of hyperspectral images for brain cancer detection during

surgical operations". In: *PLOS ONE* 13.3 (Mar. 2018), e0193721. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0193721. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0193721 (cit. on pp. 117, 155).

[95] Baowei Fei. "Chapter 3.6 - Hyperspectral imaging in medical applications". In: *Data Handling in Science and Technology*. Ed. by José Manuel Amigo. Vol. 32. Hyperspectral Imaging. Elsevier, Jan. 2020, pp. 523–565. DOI: 10.1016/B978-0-444-63977-6.00021-3. URL: https://www.sciencedirect.com/science/article/pii/B9780444639776000213 (cit. on pp. 25, 115, 148).

[96] Axelle Felicio-Briegel, Matthäus Linek, Ronald Sroka, Adrian Rühm, Christian Freymüller, Magdalena Stocker, Philipp Baumeister, Christoph Reichel, and Veronika Volgger. "Hyperspectral imaging for monitoring of free flaps of the oral cavity: A feasibility study". In: *Lasers Surg Med* 56.2 (Jan. 2024), pp. 165–174. DOI: 10.1002/lsm.23756 (cit. on p. 80).

[97] Ricard Ferrer, Ignacio Martin-Loeches, Gary Phillips, Tiffany M Osborn, Sean Townsend, R Phillip Dellinger, Antonio Artigas, Christa Schorr, and Mitchell M Levy. "Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program". In: *Critical care medicine* 42.8 (2014), pp. 1749–1755. DOI: 10.1097/CCM.0000000000000330 (cit. on pp. 10, 39).

[98] Paul Fieguth. *An introduction to pattern recognition and machine learning*. Springer, 2022. DOI: 10.1007/978-3-030-95995-1 (cit. on pp. 42, 43, 46, 48).

[99] Graham D Finlayson and Gerald Schaefer. "Solving for colour constancy using a constrained dichromatic reflection model". In: *International Journal of Computer Vision* 42 (2001), pp. 127–144. DOI: 10.1023/A:1011120214885 (cit. on p. 60).

[100] Graham D Finlayson and Elisabetta Trezzi. "Shades of gray and colour constancy". In: *Color and Imaging Conference*. Vol. 12. Society of Imaging Science and Technology. 2004, pp. 37–41. DOI: 10.2352/CIC.2004.12.1.art00008 (cit. on pp. 59, 60).

[101] Louise Finlayson, Isla RM Barnard, Lewis McMillan, Sally H Ibbotson, C Tom A Brown, Ewan Eadie, and Kenneth Wood. "Depth penetration of light into skin as a function of wavelength from 200 to 1000 nm". In: *Photochemistry and Photobiology* 98.4 (2022), pp. 974–981. DOI: 10.1111/php.13550 (cit. on p. 20).

[102] Carolin Fleischmann, Daniel O Thomas–Rueddel, Michael Hartmann, Christiane S Hartog, Tobias Welte, Steffen Heublein, Ulf Dennler, and Konrad Reinhart. "Hospital incidence and mortality rates of sepsis: an analysis of hospital episode (DRG) statistics in Germany from 2007 to 2013". In: *Deutsches Ärzteblatt International* 113.10 (2016), p. 159. DOI: 10.3238/arztebl.2016.0159 (cit. on p. 32).

[103] Carolin Fleischmann-Struzek and Kristina Rudd. "Challenges of assessing the burden of sepsis". In: *Medizinische Klinik-Intensivmedizin und Notfallmedizin* 118.Suppl 2 (2023), pp. 68–74. DOI: 10.1007/s00063-023-01088-7 (cit. on p. 32).

[104] Lucas M Fleuren, Thomas LT Klausch, Charlotte L Zwager, Linda J Schoonmade, Tingjie Guo, Luca F Roggeveen, Eleonora L Swart, Armand RJ Girbes, Patrick Thoral, Ari Ercole, et al. "Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy". In: *Intensive care medicine* 46 (2020), pp. 383–400. DOI: 10.1007/s00134-019-05872-y (cit. on p. 178).

[105] Gerald Friedland. *Information-Driven Machine Learning: Data Science as an Engineering Discipline*. Springer, 2024. DOI: 10.1007/978-3-031-39477-5 (cit. on p. 53).

[106] Yunguan Fu, Maria R. Robu, Bongjin Koo, Crispin Schneider, Stijn van Laarhoven, Danail Stoyanov, Brian Davidson, Matthew J. Clarkson, and Yipeng Hu. "More Unlabelled Data or Label More Data? A Study on Semi-supervised Laparoscopic Image Segmentation". In: *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Ed. by Qian Wang, Fausto Milletari, Hien V. Nguyen, Shadi Albarqouni, M. Jorge Cardoso, Nicola Rieke, Ziyue Xu, Konstantinos Kamnitsas, Vishal Patel, Badri Roysam, Steve Jiang, Kevin Zhou, Khoa Luu, and Ngan Le. Cham: Springer International Publishing, 2019, pp. 173–180. ISBN: 978-3-030-33391-1. DOI: 10.1007/978-3-030-33391-1_20. URL: http://arxiv.org/abs/1908.08035 (cit. on pp. 114, 116, 155).

[107] Michael G Gaies, James G Gurney, Alberta H Yen, Michelle L Napoli, Robert J Gajarski, Richard G Ohye, John R Charpie, and Jennifer C Hirsch. "Vasoactive–inotropic score as a predictor of morbidity and mortality in infants after cardiopulmonary bypass". In: *Pediatric critical care medicine* 11.2 (2010), pp. 234–238. DOI: 10.1097/PCC.0b013e3181b806fc (cit. on p. 182).

[108] Giorgio Gandaglia, Peter Schatteman, Geert De Naeyer, Frederiek D'Hondt, and Alexandre Mottrie. "Novel technologies in urologic surgery: a rapidly changing scenario". In: *Current urology reports* 17 (2016), pp. 1–8. DOI: 10.1007/s11934-016-0577-3 (cit. on p. 29).

[109] Luis C Garcia Peraza Herrera, Conor Horgan, Sebastien Ourselin, Michael Ebner, and Tom Vercauteren. "Hyperspectral image segmentation: a preliminary study on the Oral and Dental Spectral Image Database (ODSI-DB)". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 11.4 (2023), pp. 1290–1298. DOI: 10.1080/21681163.2022.2160377 (cit. on pp. 115, 117, 155).

[110] Azat Garifullin, Peeter Kööbi, Pasi Ylitepsa, Kati Ådjers, Markku Hauta-Kasari, Hannu Uusitalo, and Lasse Lensu. “Hyperspectral Image Segmentation of Retinal Vasculature, Optic Disc and Macula”. In: *2018 Digital Image Computing: Techniques and Applications (DICTA)*. Dec. 2018, pp. 1–5. DOI: 10.1109/DICTA.2018.8615761 (cit. on pp. 115, 117, 155).

[111] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. “A survey of uncertainty in deep neural networks”. In: *Artificial Intelligence Review* 56.1 (Oct. 1, 2023), pp. 1513–1589. ISSN: 1573-7462. DOI: 10.1007/s10462-023-10562-9 (cit. on p. 219).

[112] Larsa Gawria, Ahmed Jaber, Richard Peter Gerardus Ten Broek, Gianmaria Bernasconi, Rachel Rosenthal, Harry Van Goor, and Salome Dell-Kuster. “Appraisal of Intraoperative Adverse Events to Improve Postoperative Care”. In: *J Clin Med* 12.7 (Mar. 2023) (cit. on p. 28).

[113] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. “Shortcut learning in deep neural networks”. In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673. DOI: 10.1038/s42256-020-00257-z (cit. on pp. 12, 109).

[114] Evalyn I George, COL Timothy C Brand, Jacques Marescaux, et al. “Origins of robotic surgery: from skepticism to standard of care”. In: *JSLS: Journal of the Society of Laparoendoscopic Surgeons* 22.4 (2018). DOI: 10.4293/JSLS.2018.00039 (cit. on p. 28).

[115] Negin Ghamsarian, Yosuf El-Shabrawi, Sahar Nasirihaghighi, Doris Putzgruber-Adamitsch, Martin Zinkernagel, Sebastian Wolf, Klaus Schoeffmann, and Raphael Sznitman. “Cataract-1K: Cataract Surgery Dataset for Scene Segmentation, Phase Recognition, and Irregularity Detection”. In: *arXiv preprint arXiv:2312.06295* (2023). DOI: 10.48550/arXiv.2312.06295 (cit. on pp. 116, 155).

[116] Isaias Ghebrehiwet, Nazar Zaki, Rafat Damseh, and Mohd Saberi Mohamad. “Revolutionizing personalized medicine with generative AI: a systematic review”. In: *Artificial Intelligence Review* 57.5 (Apr. 2024). ISSN: 1573-7462. DOI: 10.1007/s10462-024-10768-5 (cit. on pp. 10, 43).

[117] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. “Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, June 2021, pp. 2917–2927. ISBN: 978-1-6654-4509-2 (cit. on p. 157).

[118] Evangelos J Giamarellos-Bourboulis, Anna C Aschenbrenner, Michael Bauer, Christoph Bock, Thierry Calandra, Irit Gat-Viks, Evdoxia Kyriazopoulou, Mihaela Lupse, Guillaume Monneret, Peter Pickkers, et al. "The pathophysiology of sepsis and precision-medicine-based immunotherapy". In: *Nature immunology* 25.1 (2024), pp. 19–28. DOI: 10.1038/s41590-023-01660-5 (cit. on p. 34).

[119] Eli Gibson, Maria R. Robu, Stephen Thompson, P. Eddie Edwards, Crispin Schneider, Kurinchi Gurusamy, Brian Davidson, David J. Hawkes, Dean C. Barratt, and Matthew J. Clarkson. "Deep residual networks for automatic segmentation of laparoscopic videos of the liver". In: ed. by Robert J. Webster and Baowei Fei. Orlando, Florida, United States, Mar. 2017, p. 101351M. DOI: 10.1117/12.2255975 (cit. on pp. 114–116, 155).

[120] Diaspective Vision GmbH. *Tivita Tissue FAQs*. URL: https://diaspective-vision.com/wp-content/uploads/2021/02/0101001-MD-012-b_TIVITA-Tissue-FAQ_EN.pdf (visited on 02/13/2025) (cit. on pp. 81, 86, 88).

[121] Alexander FH Goetz, Gregg Vane, Jerry E Solomon, and Barrett N Rock. "Imaging spectrometry for earth remote sensing". In: *science* 228.4704 (1985), pp. 1147–1153. DOI: 10.1126/science.228.4704.1147 (cit. on p. 3).

[122] Julia Gong, F. Christopher Holsinger, Julia E. Noel, Sohei Mitani, Jeff Jopling, Nikita Bedi, Yoon Woo Koh, Lisa A. Orloff, Claudio R. Cernea, and Serena Yeung. "Using deep learning to identify the recurrent laryngeal nerve during thyroidectomy". In: *Scientific Reports* 11.1 (July 2021), p. 14306. ISSN: 2045-2322. DOI: 10.1038/s41598-021-93202-y. URL: https://www.nature.com/articles/s41598-021-93202-y (cit. on pp. 114, 116, 155).

[123] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: http://www.deeplearningbook.org (cit. on pp. 42, 43, 49, 50, 52).

[124] Eberhard Grambow, Michael Dau, Amadeus Holmer, Vicky Lipp, Bernhard Frerich, Ernst Klar, Brigitte Vollmar, and Peer Wolfgang Kämmerer. "Hyperspectral imaging for monitoring of perfusion failure upon microvascular anastomosis in the rat hind limb". In: *Microvasc Res* 116 (Nov. 2017), pp. 64–70. DOI: 10.1016/j.mvr.2017.10.005 (cit. on p. 108).

[125] Maria Grammatikopoulou, Evangello Flouty, Abdolrahim Kadkhodamohammadi, Gwenolé Quellec, Andre Chow, Jean Nehme, Imanol Luengo, and Danail Stoyanov. "CaDIS: Cataract dataset for surgical RGB-image segmentation". In: *Medical Image Analysis* 71 (July 2021), p. 102053. ISSN: 1361-8415. DOI: 10.1016/j.media.2021.102053 (cit. on pp. 8, 116, 155).

[126] Michael D. Grossberg and Shree K. Nayar. "What is the space of camera response functions?" In: *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.* 2 (2003), pp. II–602. DOI: 10.1109/CVPR.2003.1211522 (cit. on p. 26).

[127] SepNet Critical Care Trials Group. “Incidence of severe sepsis and septic shock in German intensive care units: the prospective, multicentre INSEP study”. In: *Intensive care medicine* 42 (2016), pp. 1980–1989. DOI: 10.1007/s00134-016-4504-3 (cit. on p. 34).

[128] Amit Gupta, Tanuj Singla, Jaine John Chennatt, Lena Elizabath David, Shaik Sameer Ahmed, and Deepak Rajput. “Artificial intelligence: A new tool in surgeon’s hand”. In: *J. Educ. Health Promot.* 11 (Mar. 2022), p. 93. DOI: 10.4103/jehp.jehp_625_21 (cit. on p. 31).

[129] Khurshid A Guru, Ehsan T Esfahani, Syed J Raza, Rohit Bhat, Katy Wang, Yana Hammond, Gregory Wilding, James O Peabody, and Ashirwad J Chowriappa. “Cognitive skills assessment during robot-assisted surgery: separating the wheat from the chaff”. In: *BJU international* 115.1 (2015), pp. 166–174. DOI: 10.1111/bju.12657 (cit. on p. 28).

[130] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. “Gene selection for cancer classification using support vector machines”. In: *Machine learning* 46 (2002), pp. 389–422. DOI: 10.1023/A:1012487302797 (cit. on pp. 192, 199–201, 266, 267).

[131] Bastiaan W Haak and W Joost Wiersinga. “The role of the gut microbiota in sepsis”. In: *The lancet Gastroenterology & hepatology* 2.2 (2017), pp. 135–143. DOI: 10.1016/S2468-1253(16)30119-4 (cit. on pp. 38, 42).

[132] Mikael Häggström. “Medical gallery of Mikael Häggström 2014”. In: *WikiJournal of Medicine* 1.2 (2014). ISSN: 2002-4436. DOI: 10.15347/wjm/2014.008 (cit. on p. 30).

[133] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90 (cit. on pp. 53, 187, 189).

[134] G.E. Healey and R. Kondepudy. “Radiometric CCD camera calibration and noise estimation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16.3 (1994), pp. 267–276. DOI: 10.1109/34.276126 (cit. on p. 26).

[135] Tobias Heimann, Bram van Ginneken, Martin A. Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, GyÖrgy Bekes, Fernando Bello, Gerd Binnig, Horst Bischof, Alexander Bornik, Peter M. M. Cashman, Ying Chi, AndrÉs Cordova, Benoit M. Dawant, MÁrta Fidrich, Jacob D. Furst, Daisuke Furukawa, Lars Grenacher, Joachim Hornegger, Dagmar KainmÜller, Richard I. Kitney, Hidefumi Kobatake, Hans Lamecker, Thomas Lange, Jeongjin Lee, Brian Lennon, Rui Li, Senhu Li, Hans-Peter Meinzer, GÁbor Nemeth, Daniela S. Raicu, Anne-Mareike Rau, Eva M. van Rikxoort, MikaËl Rousson, LÁszlÓ Rusko, Kinda A. Saddi, GÜnter Schmidt,

Dieter Seghers, Akinobu Shimizu, Pieter Slagmolen, Erich Sorantin, Grzegorz Soza, Ruchaneewan Susomboon, Jonathan M. Waite, Andreas Wimmer, and Ivo Wolf. "Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets". In: *IEEE Transactions on Medical Imaging* 28.8 (Aug. 2009). Conference Name: IEEE Transactions on Medical Imaging, pp. 1251–1265. ISSN: 1558-254X. DOI: 10.1109/TMI.2009.2013851 (cit. on p. 130).

[136] Daniel J Henning, Jeremy R Carey, Kimie Oedorf, Danielle E Day, Colby S Redfield, Colin J Huguenel, Jonathan C Roberts, Leon D Sanchez, Richard E Wolfe, and Nathan I Shapiro. "The absence of fever is associated with higher mortality and decreased antibiotic and IV fluid administration in emergency department patients with suspected septic shock". In: *Critical care medicine* 45.6 (2017), e575–e582. DOI: 10.1097/CCM.0000000000002311 (cit. on pp. 10, 39, 178).

[137] Suzana Herculano-Houzel. "The human brain in numbers: a linearly scaled-up primate brain". In: *Frontiers in human neuroscience* 3 (2009), p. 857. DOI: 10.3389/neuro.09.031.2009 (cit. on p. 52).

[138] EMC Hillman. "Optical brain imaging in vivo: techniques and applications from animal to man". In: *J. Biomed. Opt.* 12 (2007), pp. 1–28 (cit. on p. 26).

[139] Victor WM van Hinsbergh. "Endotheliumrole in regulation of coagulation and inflammation". In: *Seminars in immunopathology*. Vol. 34. Springer. 2012, pp. 93–106. DOI: 10.1007/s00281-011-0285-5 (cit. on p. 36).

[140] Tim Holland-Letz and Annette Kopp-Schneider. "Drawing statistical conclusions from experiments with multiple quantitative measurements per subject". In: *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 152 (Nov. 2020), pp. 30–33. ISSN: 1879-0887. DOI: 10.1016/j.radonc.2020.08.009 (cit. on p. 130).

[141] Amadeus Holmer, Jörg Marotz, Philip Wahl, Michael Dau, and Peer W. Kämmerer. "Hyperspectral imaging in perfusion and wound diagnostics – methods and algorithms for the determination of tissue parameters". In: *Biomedical Engineering / Biomedizinische Technik* 63.5 (2018), pp. 547–556. ISSN: 0013-5585. DOI: 10.1515/bmt-2017-0155. URL: https://www.degruyter.com/view/j/bmte.2018.63.issue-5/bmt-2017-0155/bmt-2017-0155.xml (cit. on pp. 4, 21, 25, 27, 86, 119, 124, 187, 188).

[142] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. *What do we need to build explainable AI systems for the medical domain?* 2017. DOI: 10.48550/arXiv.1712.09923. arXiv: 1712.09923 [cs.AI] (cit. on p. 219).

[143] Richard S Hotchkiss, Craig M Coopersmith, Jonathan E McDunn, and Thomas A Ferguson. "The sepsis seesaw: tilting toward immunosuppression". In: *Nature medicine* 15.5 (2009), pp. 496–497. DOI: 10.1038/nm0509-496 (cit. on pp. 34, 41).

[144] Richard S Hotchkiss, Lyle L Moldawer, Steven M Opal, Konrad Reinhart, Isaiah R Turnbull, and Jean-Louis Vincent. “Sepsis and septic shock”. In: *Nature reviews Disease primers* 2.1 (2016), pp. 1–21. DOI: 10.1038/nrdp.2016.45 (cit. on p. 41).

[145] Richard S Hotchkiss, Guillaume Monneret, and Didier Payen. “Sepsis-induced immunosuppression: from cellular dysfunctions to immunotherapy”. In: *Nature Reviews Immunology* 13.12 (2013), pp. 862–874. DOI: 10.1038/nri3552 (cit. on pp. 37, 42).

[146] Richard S Hotchkiss, Paul E Swanson, Bradley D Freeman, Kevin W Tinsley, J Perren Cobb, George M Matuschak, Timothy G Buchman, and Irene E Karl. “Apoptotic cell death in patients with sepsis, shock, and multiple organ dysfunction”. In: *Critical care medicine* 27.7 (1999), pp. 1230–1251 (cit. on p. 38).

[147] Yuanming Hu, Baoyuan Wang, and Stephen Lin. “Fc4: Fully convolutional color constancy with confidence-weighted pooling”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4085–4094 (cit. on p. 58).

[148] Longqian Huang, Ruichen Luo, Xu Liu, and Xiang Hao. “Spectral imaging with deep learning”. In: *Light: Science & Applications* 11.1 (Mar. 2022), p. 61 (cit. on pp. 42, 218).

[149] Min Huang, Shaoli Cai, and Jingqian Su. “The pathogenesis of sepsis and potential therapeutic targets”. In: *International journal of molecular sciences* 20.21 (2019), p. 5376. DOI: 10.3390/ijms20215376 (cit. on p. 37).

[150] Sophie C Huijskens, Irma W E M van Dijk, Jorrit Visser, Brian V Balgobind, D te Lindert, Coen R N Rasch, Tanja Alderliesten, and Arjan Bel. “Abdominal organ position variation in children during image-guided radiotherapy”. In: *Radiation Oncology* 13.1 (Sept. 2018), p. 173 (cit. on p. 29).

[151] Yoshie Imai, Yu Kato, Hideki Kadoi, Takahiko Horiuchi, and Shoji Tominaga. “Estimation of Multiple Illuminants Based on Specular Highlight Detection”. In: *Computational Color Imaging*. Ed. by Raimondo Schettini, Shoji Tominaga, and Alain Trémeau. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 85–98. DOI: 10.1007/978-3-642-20404-3_7 (cit. on p. 60).

[152] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *arXiv:1502.03167 [cs]* (Mar. 2015). arXiv: 1502.03167. URL: http://arxiv.org/abs/1502.03167 (cit. on p. 128).

[153] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature Methods* 18.2 (Feb. 1, 2021), pp. 203–211. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01008-z (cit. on p. 126).

[154] Khandaker Reajul Islam, Johayra Prithula, Jaya Kumar, Toh Leong Tan, Mamun Bin Ibne Reaz, Md Shaheenur Islam Sumon, and Muhammad EH Chowdhury. "Machine learning-based early prediction of sepsis using electronic health records: A systematic review". In: *Journal of clinical medicine* 12.17 (2023), p. 5658. DOI: 10.3390/jcm12175658 (cit. on p. 178).

[155] Shinya Iwase, Taka-aki Nakada, Tadanaga Shimada, Takehiko Oami, Takashi Shimazui, Nozomi Takahashi, Jun Yamabe, Yasuo Yamao, and Eiryo Kawakami. "Prediction algorithm for ICU mortality and length of stay using machine learning". In: *Scientific reports* 12.1 (2022), p. 12912. DOI: 10.1038/s41598-022-17091-5 (cit. on p. 12).

[156] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. "Averaging Weights Leads to Wider Optima and Better Generalization". In: *Conference on Uncertainty in Artificial Intelligence*. 2018. URL: https://api.semanticscholar.org/CorpusID:3833416 (cit. on pp. 128, 189).

[157] Steven L Jacques. "Optical properties of biological tissues: a review". In: *Physics in Medicine & Biology* 58.11 (May 2013), R37. DOI: 10.1088/0031-9155/58/11/R37 (cit. on pp. 17, 18).

[158] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer New York, 2013. DOI: 10.1007/978-1-4614-7138-7 (cit. on pp. 44, 46, 50).

[159] Boris Jansen-Winkeln, N Holfert, H Köhler, Y Moulla, JP Takoh, SM Rabe, M Mehdorn, M Barberio, C Chalopin, T Neumuth, et al. "Determination of the transection margin during colorectal resection with hyperspectral imaging (HSI)". In: *International journal of colorectal disease* 34 (2019), pp. 731–739. DOI: 10.1007/s00384-019-03250-0 (cit. on p. 29).

[160] Francis Arthur Jenkins and Harvey Elliott White. *Fundamentals of optics: By francis a. jenkins and harvey e. white*. McGraw-Hill, 1976 (cit. on p. 17).

[161] Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven A. Hicks, Vajira Thambawita, Enrique Garcia-Ceja, Michael A. Riegler, Thomas de Lange, Peter T. Schmidt, Håvard D. Johansen, Dag Johansen, and Pål Halvorsen. "Kvasir-Instrument: Diagnostic and Therapeutic Tool Segmentation Dataset in Gastrointestinal Endoscopy". In: *MultiMedia Modeling*. Ed. by Jakub Loko, Tomá Skopal, Klaus Schoeffmann, Vasileios Mezaris, Xirong Li, Stefanos Vrochidis, and Ioannis Patras. Cham: Springer International Publishing, 2021, pp. 218–229. ISBN: 978-3-030-67835-7. DOI: 10.1007/978-3-030-67835-7_19 (cit. on p. 114).

[162] Yueming Jin, Yang Yu, Cheng Chen, Zixu Zhao, Pheng-Ann Heng, and Danail Stoyanov. “Exploring Intra- and Inter-Video Relation for Surgical Semantic Scene Segmentation”. In: *IEEE Transactions on Medical Imaging* 41.11 (2022), pp. 2991–3002. DOI: 10.1109/TMI.2022.3177077 (cit. on pp. 116, 155).

[163] Daniel Hernández Juárez, Sarah Parisot, Benjamin Busam, Ales Leonardis, Gregory G Slabaugh, and Steven McDonagh. “A Multi-Hypothesis Approach to Color Constancy.” In: *CVPR*. 2020, pp. 2267–2277 (cit. on p. 58).

[164] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin ídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (July 2021), pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2 (cit. on pp. 10, 43).

[165] Abdolrahim Kadkhodamohammadi, Imanol Luengo, Santiago Barbarisi, Hinde Taleb, Evangello Flouty, and Danail Stoyanov. “Feature Aggregation Decoder for Segmenting Laparoscopic Scenes”. In: *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*. Ed. by Luping Zhou, Duygu Sarikaya, Seyed Mostafa Kia, Stefanie Speidel, Anand Malpani, Daniel Hashimoto, Mohamad Habes, Tommy Löfstedt, Kerstin Ritter, and Hongzhi Wang. Vol. 11796. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 3–11. ISBN: 978-3-030-32695-1. DOI: 10.1007/978-3-030-32695-1_1. URL: http://link.springer.com/10.1007/978-3-030-32695-1_1 (cit. on pp. 116, 155).

[166] Kanav Kahol, Mario J Leyba, Mary Deka, Vikram Deka, Stephanie Mayes, Marshall Smith, John J Ferrara, and Sethuraman Panchanathan. “Effect of fatigue on psychomotor and cognitive skills”. In: *The American Journal of Surgery* 195.2 (2008), pp. 195–204. DOI: 10.1016/j.amjsurg.2007.10.004 (cit. on p. 28).

[167] Ibrahem Kandel and Mauro Castelli. “The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset”. In: *ICT Express* 6.4 (2020), pp. 312–315. ISSN: 2405-9595. DOI: 10.1016/j.icte.2020.04.010. URL: https://www.sciencedirect.com/science/article/pii/S2405959519303455 (cit. on p. 128).

[168] Eiji Kaneko, Hirofumi Aoki, and Masato Tsukada. “Daylight Spectrum Estimation from Hyper-and Multispectral Image without Area Extraction of Uniform Materials”. In: *2015 11th International Conference on Signal-Image Technology &*

Internet-Based Systems (SITIS). 2015, pp. 53–60. DOI: 10.1109/SITIS.2015.67 (cit. on p. 60).

[169] Mithun Kumar Kar, Malaya Kumar Nath, and Debanga Raj Neog. "A Review on Progress in Semantic Image Segmentation and Its Application to Medical Images". In: *SN Computer Science* 2.5 (July 2021), p. 397. ISSN: 2661-8907. DOI: 10.1007/s42979-021-00784-5. (Visited on 02/20/2023) (cit. on p. 154).

[170] Jaka Katranik, Franjo Pernu, and Botjan Likar. "Radiometric calibration and noise estimation of acousto-optic tunable filter hyperspectral imaging systems". In: *Appl Opt* 52.15 (May 2013), pp. 3526–3537. DOI: 10.1364/AO.52.003526 (cit. on p. 26).

[171] George M Katz, Angel Mozo, and John P Reuben. "Filament interaction in intact muscle fibers monitored by light scattering." In: *Proceedings of the National Academy of Sciences* 76.9 (1979), pp. 4421–4424. DOI: 10.1073/pnas.76.9.4421 (cit. on p. 26).

[172] Sigita Kazune, Anastasija Caica, Karina Volceka, Olegs Suba, Uldis Rubins, and Andris Grabovskis. "Relationship of mottling score, skin microcirculatory perfusion indices and biomarkers of endothelial dysfunction in patients with septic shock: an observational study". In: *Critical Care* 23 (2019), pp. 1–9. DOI: 10.1186/s13054-019-2589-0 (cit. on p. 179).

[173] Henrik Kehlet, Troels S Jensen, and Clifford J Woolf. "Persistent postsurgical pain: risk factors and prevention". In: *The lancet* 367.9522 (2006), pp. 1618–1625. DOI: 10.1016/S0140-6736(06)68700-X (cit. on p. 31).

[174] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. "Key challenges for delivering clinical impact with artificial intelligence". In: *BMC Medicine* 17.1 (Oct. 29, 2019), p. 195. ISSN: 1741-7015. DOI: 10.1186/s12916-019-1426-2 (cit. on p. 219).

[175] Haris Ahmad Khan, Jean-Baptiste Thomas, Jon Yngve Hardeberg, and Olivier Laligant. "Illuminant estimation in multispectral imaging". In: *JOSA A* 34.7 (2017), pp. 1085–1098. DOI: 10.1364/JOSAA.34.001085 (cit. on pp. 57, 59).

[176] Uzair Khan, Paheding Sidike, Colin Elkin, and Vijay Devabhaktuni. "Trends in deep learning for medical hyperspectral image analysis". In: *IEEE Access* 9 (2021). arXiv: 2011.13974, pp. 79534–79548. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3068392 (cit. on p. 189).

[177] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6980 (cit. on p. 128).

[178] Daichi Kitaguchi, Toru Fujino, Nobuyoshi Takeshita, Hiro Hasegawa, Kensaku Mori, and Masaaki Ito. “Limited generalizability of single deep neural network for surgical instrument segmentation in different surgical environments”. In: *Scientific Reports* 12.1 (July 22, 2022), p. 12575. ISSN: 2045-2322. DOI: 10.1038/s41598-022-16923-8 (cit. on p. 154).

[179] Peter MC Klein Klouwenberg, Olaf L Cremer, Lonneke A van Vught, David SY Ong, Jos F Frencken, Marcus J Schultz, Marc J Bonten, and Tom van der Poll. “Likelihood of infection in patients with presumed sepsis at the time of intensive care unit admission: a cohort study”. In: *Critical care* 19 (2015), pp. 1–8. DOI: 10.1186/s13054-015-1035-1 (cit. on p. 42).

[180] William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. “APACHE II: a severity of disease classification system”. In: *Critical care medicine* 13.10 (1985), pp. 818–829 (cit. on p. 182).

[181] Hannes Köhler, Boris Jansen-Winkeln, Marianne Maktabi, Manuel Barberio, Jonathan Takoh, Nico Holfert, Yusef Moulla, Stefan Niebisch, Michele Diana, Thomas Neumuth, et al. “Evaluation of hyperspectral imaging (HSI) for the measurement of ischemic conditioning effects of the gastric conduit during esophagectomy”. In: *Surgical endoscopy* 33 (2019), pp. 3775–3782. DOI: 10.1007/s00464-019-06675-4 (cit. on p. 29).

[182] Judith Kohnke, Kevin Pattberg, Felix Nensa, Henning Kuhlmann, Thorsten Brenner, Karsten Schmidt, René Hosch, and Florian Espeter. “A proof of concept for microcirculation monitoring using machine learning based hyperspectral imaging in critically ill patients: a monocentric observational study”. In: *Critical Care* 28.1 (2024), p. 230. DOI: 10.1186/s13054-024-05023-w (cit. on p. 179).

[183] Esther N D Kok, Roeland Eppenga, Koert F D Kuhlmann, Harald C Groen, Ruben van Veen, Jolanda M van Dieren, Thomas R de Wijkerslooth, Monique van Leerdam, Doenja M J Lambregts, Wouter J Heerink, Nikie J Hoetjes, Oleksandra Ivashchenko, Geerard L Beets, Arend G J Aalbers, Jasper Nijkamp, and Theo J M Ruers. “Accurate surgical navigation with real-time tumor tracking in cancer surgery”. In: *npj Precision Oncology* 4.1 (Apr. 2020), p. 8. DOI: 10.1038/s41698-020-0115-0 (cit. on p. 31).

[184] Fiona R Kolbinger, Sebastian Bodenstedt, Matthias Carstens, Stefan Leger, Stefanie Krell, Franziska M Rinner, Thomas P Nielen, Johanna Kirchberg, Johannes Fritzmann, Jürgen Weitz, et al. “Artificial Intelligence for context-aware surgical guidance in complex robot-assisted oncological procedures: An exploratory feasibility study”. In: *European Journal of Surgical Oncology* (2023), p. 106996. DOI: 10.1016/j.ejso.2023.106996 (cit. on p. 31).

[185] Fiona R. Kolbinger, Franziska M. Rinner, Alexander C. Jenke, Matthias Carstens, Stefanie Krell, Stefan Leger, Marius Distler, Jürgen Weitz, Stefanie Speidel, and Sebastian Bodenstedt. "Anatomy segmentation in laparoscopic surgery: comparison of machine learning and human expertise – an experimental study". In: *International Journal of Surgery* 109.10 (2023). URL: https://journals.lww.com/international-journal-of-surgery/fulltext/2023/10000/anatomy_segmentation_in_laparoscopic_surgery_.10.aspx (cit. on pp. 114, 116, 155).

[186] Matthieu Komorowski, Ashleigh Green, Kate C. Tatham, Christopher Seymour, and David Antcliffe. "Sepsis biomarkers and diagnostic tools with a focus on machine learning". In: *EBioMedicine* 86 (2022). DOI: 10.1016/j.ebiom.2022.104394 (cit. on p. 178).

[187] Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. "Second opinion needed: communicating uncertainty in medical machine learning". In: *npj Digital Medicine* 4.1 (Jan. 5, 2021), p. 4. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00367-3 (cit. on p. 219).

[188] Axel Kulcke, Amadeus Holmer, Philip Wahl, Frank Siemers, Thomas Wild, and Georg Daeschlein. "A compact hyperspectral camera for measurement of perfusion parameters in medicine". In: *Biomedical Engineering / Biomedizinische Technik* 63.5 (Oct. 2018), pp. 519–527. ISSN: 1862-278X, 0013-5585. DOI: 10.1515/bmt-2017-0145. URL: http://www.degruyter.com/view/j/bmte.2018.63.issue-5/bmt-2017-0145/bmt-2017-0145.xml (cit. on pp. 4, 21, 25).

[189] Solomon Kullback and Richard A. Leibler. "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1 (1951). Publisher: Institute of Mathematical Statistics, pp. 79–86. ISSN: 0003-4851. URL: https://www.jstor.org/stable/2236703 (cit. on p. 126).

[190] Ajith AK Kumar. "Mortality prediction in the icu: The daunting task of predicting the unpredictable". In: *Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine* 26.1 (2022), p. 13. DOI: 10.5005/jp-journals-10071-24063 (cit. on p. 12).

[191] Yuta Kumazu, Nao Kobayashi, Naoki Kitamura, Elleuch Rayan, Paul Neculoiu, Toshihiro Misumi, Yudai Hojo, Tatsuro Nakamura, Tsutomu Kumamoto, Yasunori Kurahashi, et al. "Automated segmentation by deep learning of loose connective tissue fibers to define safe dissection planes in robot-assisted gastrectomy". In: *Scientific Reports* 11.1 (2021), p. 21198. DOI: 10.1038/s41598-021-00557-3 (cit. on pp. 28, 30).

[192] Firas Laakom, Jenni Raitoharju, Alexandros Iosifidis, Jarno Nikkanen, and Moncef Gabbouj. "Color constancy convolutional autoencoder". In: *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2019, pp. 1085–1090. DOI: 10.1109/SSCI44817.2019.9002684 (cit. on p. 58).

[193] M Lcis, S Kazune, Z Marcinkevics, U Rubins, and A Grabovskis. “Hybrid optical prototype for sepsis bedside diagnostics”. In: *Novel Optical Systems, Methods, and Applications XXII*. Vol. 11105. SPIE. 2019, pp. 123–129. DOI: 10.1117/12.2529230 (cit. on p. 179).

[194] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. “Quantifying the Carbon Emissions of Machine Learning”. In: (Oct. 2019). URL: https://arxiv.org/abs/1910.09700 (cit. on p. 147).

[195] Edwin H Land. “The retinex theory of color vision”. In: *Scientific American* 237.6 (1977), pp. 108–129 (cit. on p. 59).

[196] Max-Heinrich Laves, Jens Bicker, Lüder A. Kahrs, and Tobias Ortmaier. “A dataset of laryngeal endoscopic images with comparative study on convolution neural network-based semantic segmentation”. In: *International Journal of Computer Assisted Radiology and Surgery* 14.3 (Mar. 2019), pp. 483–492. ISSN: 1861-6429. DOI: 10.1007/s11548-018-01910-0 (cit. on pp. 116, 155).

[197] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *Nature* 521.7553 (May 1, 2015), pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539 (cit. on pp. 43, 49).

[198] Ann-Kathrin Lederer, Sophia Chikhladze, Eva Kohnert, Roman Huber, and Alexander Müller. “Current Insights: The Impact of Gut Microbiota on Postoperative Complications in Visceral SurgeryA Narrative Review”. In: *Diagnostics* 11.11 (2021). ISSN: 2075-4418. DOI: 10.3390/diagnostics11112099. URL: https://www.mdpi.com/2075-4418/11/11/2099 (cit. on p. 28).

[199] Aleksandra Leligdowicz, Lauren F Chun, Alejandra Jauregui, Kathryn Vessel, Kathleen D Liu, Carolyn S Calfee, and Michael A Matthay. “Human pulmonary endothelial cell permeability after exposure to LPS-stimulated leukocyte supernatants derived from patients with early sepsis”. In: *American Journal of Physiology-Lung Cellular and Molecular Physiology* 315.5 (2018), pp. L638–L644. DOI: 10.1152/ajplung.00286.2018 (cit. on p. 36).

[200] Aleksandra Leligdowicz, Peter M Dodek, Monica Norena, Hubert Wong, Aseem Kumar, Anand Kumar, and Co-operative Antimicrobial Therapy of Septic Shock Database Research Group. “Association between source of infection and hospital mortality in patients who have septic shock”. In: *American journal of respiratory and critical care medicine* 189.10 (2014), pp. 1204–1213. DOI: 10.1164/rccm.201310-1875OC (cit. on p. 34).

[201] Christophe Lelubre and Jean-Louis Vincent. “Mechanisms and treatment of organ failure in sepsis”. In: *Nature Reviews Nephrology* 14.7 (2018), pp. 417–427. DOI: 10.1038/s41581-018-0005-7 (cit. on pp. 37, 38).

[202] Raquel Leon, Himar Fabelo, Samuel Ortega, Ines A. Cruz-Guerrero, Daniel Ulises Campos-Delgado, Adam Szolna, Juan F. Piñeiro, Carlos Espino, Aruma J. O'Shanahan, Maria Hernandez, David Carrera, Sara Bisshopp, Coralia Sosa, Francisco J. Balea-Fernandez, Jesus Morera, Bernardino Clavo, and Gustavo M. Callico. "Hyperspectral imaging benchmark based on machine learning for intraoperative brain tumour detection". In: *npj Precision Oncology* 7.1 (Nov. 14, 2023), p. 119. ISSN: 2397-768X. DOI: 10.1038/s41698-023-00475-9 (cit. on pp. 3, 31, 115, 117, 155).

[203] Raquel Leon, Beatriz Martinez-Vega, Himar Fabelo, Samuel Ortega, Veronica Melian, Irene Castaño, Gregorio Carretero, Pablo Almeida, Aday Garcia, Eduardo Quevedo, Javier A. Hernandez, Bernardino Clavo, and Gustavo M. Callico. "Non-Invasive Skin Cancer Diagnosis Using Hyperspectral Imaging for In-Situ Clinical Support". In: *Journal of Clinical Medicine* 9.6 (2020). ISSN: 2077-0383. DOI: 10.3390/jcm9061662 (cit. on pp. 3, 31).

[204] Marcel Levi, Marcus Schultz, and Tom van der Poll. "Sepsis and thrombosis". In: *Seminars in thrombosis and hemostasis*. Vol. 39. 05. Thieme Medical Publishers. 2013, pp. 559–566. DOI: 10.1055/s-0033-1343894 (cit. on p. 37).

[205] Qingli Li, Xiaofu He, Yiting Wang, Hongying Liu, Dongrong Xu, and Fangmin Guo. "Review of spectral imaging technology in biomedical engineering: achievements and challenges". In: *Journal of Biomedical Optics* 18.10 (2013), p. 100901. DOI: 10.1117/1.JBO.18.10.100901 (cit. on pp. 3, 217).

[206] Xin Li, Lei Zhang, Jingsi Yang, and Fei Teng. "Role of Artificial Intelligence in Medical Image Analysis: A Review of Current Trends and Future Directions". In: *Journal of Medical and Biological Engineering* 44.2 (Apr. 2024), pp. 231–243. ISSN: 2199-4757. DOI: 10.1007/s40846-024-00863-x (cit. on pp. 10, 43).

[207] Jeremiah K H Lim, Qiao-Xin Li, Tim Ryan, Phillip Bedggood, Andrew Metha, Algis J Vingrys, Bang V Bui, and Christine T O Nguyen. "Retinal hyperspectral imaging in the 5xFAD mouse model of Alzheimer's disease". In: *Scientific Reports* 11.1 (Mar. 2021), p. 6387. DOI: 10.1038/s41598-021-85554-2 (cit. on p. 3).

[208] Stephen Lin. "Illumination Estimation". In: *Computer Vision: A Reference Guide*. Ed. by Katsushi Ikeuchi. Cham: Springer International Publishing, 2021, pp. 599–604. ISBN: 978-3-030-63416-2. DOI: 10.1007/978-3-030-63416-2_516 (cit. on p. 58).

[209] Min Liu, Yubin Han, Jiazheng Wang, Can Wang, Yaonan Wang, and Erik Meijering. "LSKANet: Long Strip Kernel Attention Network for Robotic Surgical Scene Segmentation". In: *IEEE Transactions on Medical Imaging* 43.4 (2024), pp. 1308–1322. DOI: 10.1109/TMI.2023.3335406 (cit. on pp. 116, 155).

[210] Vincent Liu, Gabriel J Escobar, John D Greene, Jay Soule, Alan Whippy, Derek C Angus, and Theodore J Iwashyna. "Hospital deaths in patients with sepsis from 2 independent cohorts". In: *Jama* 312.1 (2014), pp. 90–92. DOI: 10.1001/jama.2014.5804 (cit. on p. 32).

[211] Tyler J Loftus, Benjamin Shickel, Matthew M Ruppert, Jeremy A Balch, Tezcan Ozrazgat-Baslanti, Patrick J Tighe, Philip A Efron, William R Hogan, Parisa Rashidi, Gilbert R Upchurch Jr, and Azra Bihorac. "Uncertainty-aware deep learning in healthcare: A scoping review". In: *PLOS Digit Health* 1.8 (Aug. 2022). DOI: 10.1371/journal.pdig.0000085 (cit. on p. 219).

[212] Bert K Lopansri, Russell R Miller III, John P Burke, Mitchell Levy, Steven Opal, Richard E Rothman, Franco R DAlessio, Venkataramana K Sidhaye, Robert Balk, Jared A Greenberg, et al. "Physician agreement on the diagnosis of sepsis in the intensive care unit: estimation of concordance and analysis of underlying factors in a multicenter cohort". In: *Journal of intensive care* 7 (2019), pp. 1–17. DOI: 10.1186/s40560-019-0368-2 (cit. on pp. 12, 179).

[213] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: *arXiv preprint arXiv:1711.05101* (2019). DOI: 10.48550/arXiv.1711.05101 (cit. on pp. 188, 189).

[214] Mayar Lotfy, Anna Alperovich, Tommaso Giannantonio, Bjorn Barz, Xiaohan Zhang, Felix Holm, Nassir Navab, Felix Boehm, Carolin Schwamborn, Thomas K. Hoffmann, and Patrick J. Schuler. *Robust Tumor Segmentation with Hyperspectral Imaging and Graph Neural Networks*. 2023. arXiv: 2311.11782 [eess.IV] (cit. on pp. 115, 117, 155).

[215] Guolan Lu and Baowei Fei. "Medical hyperspectral imaging: a review". In: *J Biomed Opt* 19.1 (Jan. 2014), p. 10901. DOI: 10.1117/1.JBO.19.1.010901 (cit. on p. 25).

[216] Yu-Wen Luo, Hai-Yong Chen, Zhen Li, Wei-Peng Liu, Ke Wang, Li Zhang, Pan Fu, Wen-Qian Yue, and Gui-Bin Bian. "Fast instruments and tissues segmentation of micro-neurosurgical scene using high correlative non-local network". In: *Computers in Biology and Medicine* 153 (2023), p. 106531. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2022.106531. URL: https://www.sciencedirect.com/science/article/pii/S0010482522012392 (cit. on pp. 116, 155).

[217] Nicholas J Lynch, Colin L Willis, Christopher C Nolan, Silke Roscher, Maxine J Fowler, Eberhard Weihe, David E Ray, and Wilhelm J Schwaeble. "Microglial activation and increased synthesis of complement component C1q precedes blood–brain barrier dysfunction in rats". In: *Molecular immunology* 40.10 (2004), pp. 709–716. DOI: 10.1016/j.molimm.2003.08.009 (cit. on p. 36).

[218] Ling Ma, Kelden Pruitt, and Baowei Fei. "A hyperspectral surgical microscope with super-resolution reconstruction for intraoperative image guidance". In: *Proceedings of SPIE–the International Society for Optical Engineering*. Vol. 12930. NIH Public Access. 2024. DOI: 10.1117/12.3008789 (cit. on pp. 217, 218).

[219] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. "Rectifier nonlinearities improve neural network acoustic models". In: *Proc. icml*. Vol. 30. 1. Atlanta, GA. 2013, p. 3 (cit. on pp. 50, 51).

[220] Sabrina Madad Zadeh, Tom Francois, Lilian Calvet, Pauline Chauvet, Michel Canis, Adrien Bartoli, and Nicolas Bourdel. "SurgAI: deep learning for computerized laparoscopic image understanding in gynaecology". In: *Surgical Endoscopy* 34.12 (Dec. 2020), pp. 5377–5383. ISSN: 1432-2218. DOI: 10.1007/s00464-019-07330-8 (cit. on pp. 114, 116, 155).

[221] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, Hirenkumar Nakawala, Adrian Park, Carla Pugh, Danail Stoyanov, Swaroop S. Vedula, Kevin Cleary, Gabor Fichtinger, Germain Forestier, Bernard Gibaud, Teodor Grantcharov, Makoto Hashizume, Doreen Heckmann-Nötzel, Hannes G. Kenngott, Ron Kikinis, Lars Mündermann, Nassir Navab, Sinan Onogur, Tobias RoSS, Raphael Sznitman, Russell H. Taylor, Minu D. Tizabi, Martin Wagner, Gregory D. Hager, Thomas Neumuth, Nicolas Padoy, Justin Collins, Ines Gockel, Jan Goedeke, Daniel A. Hashimoto, Luc Joyeux, Kyle Lam, Daniel R. Leff, Amin Madani, Hani J. Marcus, Ozanan Meireles, Alexander Seitel, Dogu Teber, Frank Ückert, Beat P. Müller-Stich, Pierre Jannin, et al. "Surgical data science – from concepts toward clinical translation". In: *Medical Image Analysis* 76 (2022), p. 102306. ISSN: 1361-8415. DOI: 10.1016/j.media.2021.102306. URL: https://www.sciencedirect.com/science/article/pii/S1361841521003510 (cit. on pp. 8, 215, 216).

[222] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, A. Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew B. Blaschko, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Robert Haase, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, et al. "Metrics reloaded: recommendations for image analysis validation". In: *Nature Methods* 21.2 (Feb. 1, 2024),

pp. 195–212. ISSN: 1548-7105. DOI: 10.1038/s41592-023-02151-z (cit. on pp. 43, 129, 130, 144, 160, 191).

[223] Lena Maier-Hein, Swaroop S. Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, Makoto Hashizume, Darko Katic, Hannes Kenngott, Michael Kranzfelder, Anand Malpani, Keno März, Thomas Neumuth, Nicolas Padoy, Carla Pugh, Nicolai Schoch, Danail Stoyanov, Russell Taylor, Martin Wagner, Gregory D. Hager, and Pierre Jannin. "Surgical data science for next-generation interventions". In: *Nature Biomedical Engineering* 1.9 (Sept. 1, 2017), pp. 691–696. ISSN: 2157-846X. DOI: 10.1038/s41551-017-0132-7 (cit. on p. 8).

[224] Lena Maier-Hein, Martin Wagner, Tobias Ross, Annika Reinke, Sebastian Bodenstedt, Peter M. Full, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Anna Kisilenko, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Minu D. Tizabi, Matthias Eisenmann, Tim J. Adler, Janek Gröhl, Melanie Schellenberg, Silvia Seidlitz, T. Y. Emmy Lai, Bünyamin Pekdemir, Veith Roethlingshoefer, Fabian Both, Sebastian Bittel, Marc Mengler, Lars Mündermann, Martin Apitz, Annette Kopp-Schneider, Stefanie Speidel, Felix Nickel, Pascal Probst, Hannes G. Kenngott, and Beat P. Müller-Stich. "Heidelberg colorectal data set for surgical data science in the sensor operating room". In: *Scientific Data* 8.1 (Apr. 12, 2021), p. 101. ISSN: 2052-4463. DOI: 10.1038/s41597-021-00882-2 (cit. on p. 114).

[225] Lena Maier-Hein, Sebastian Josef Wirkert, Anant Suraj Vemuri, Leonardo Antonio Ayala Menjivar, Silvia Seidlitz, Thomas Kirchner, and Tim Adler. "Method and system for augmented imaging in open treatment using multispectral information". EP3829416B1. June 9, 2021 (cit. on pp. 57, 213, 223).

[226] Lena Maier-Hein, Sebastian Josef Wirkert, Anant Suraj Vemuri, Leonardo Antonio Ayala Menjivar, Silvia Seidlitz, Thomas Kirchner, and Tim Adler. "Method and system for augmented imaging using multispectral information". US20220012874A1. Jan. 13, 2022 (cit. on pp. 57, 213, 223).

[227] Marianne Maktabi, Benjamin Huber, Toni Pfeiffer, and Torsten Schulz. "Detection of flap malperfusion after microsurgical tissue reconstruction using hyperspectral imaging and machine learning". In: *Scientific Reports* 15.1 (May 2025), p. 15637. DOI: 10.1038/s41598-025-98874-4 (cit. on p. 80).

[228] Harshita Mangotra, Sahima Srivastava, Garima Jaiswal, Ritu Rani, and Arun Sharma. "Hyperspectral imaging for early diagnosis of diseases: A review". In: *Expert Systems* 40.8 (2023), e13311. DOI: 10.1111/exsy.13311 (cit. on pp. 3, 29, 31).

[229] Francesca Manni, Roger Fonollà, Fons van der Sommen, Svetlana Zinger, Caifeng Shan, Esther Kho, Susan Brouwer de Koning, Theo Ruers, and Peter H.N. de With. "Hyperspectral imaging for colon cancer classification in surgical specimens: towards optical biopsy during image-guided surgery". In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2020, pp. 1169–1173. DOI: 10.1109/EMBC44109.2020.9176543 (cit. on pp. 3, 31).

[230] A. Mansouri, F.S. Marzani, and P. Gouton. "Development of a Protocol for CCD Calibration: Application to a Multispectral Imaging System". In: *International Journal of Robotics and Automation* 20.2 (2005). ISSN: 1925-7090. DOI: 10.2316/journal.206.2005.2.206-2784 (cit. on p. 26).

[231] Salman Maqbool, Aqsa Riaz, Hasan Sajid, and Osman Hasan. *m2caiSeg: Semantic Segmentation of Laparoscopic Images using Convolutional Neural Networks*. 2020. arXiv: 2008.10134 [cs.CV] (cit. on pp. 114, 116, 155).

[232] Howard Mark and Jerome Workman Jr. "Bias and slope correction". In: *Spectroscopy* 32.2 (Feb. 2017), pp. 24–30 (cit. on p. 81).

[233] Adriana Marques, Carla Torre, Rui Pinto, Bruno Sepodes, and João Rocha. "Treatment advances in sepsis and septic shock: modulating pro-and anti-inflammatory mechanisms". In: *Journal of Clinical Medicine* 12.8 (2023), p. 2892. DOI: 10.3390/jcm12082892 (cit. on pp. 38, 41, 42).

[234] Tyler R McClintock, Marc A Bjurlin, James S Wysock, Michael S Borofsky, Tracy P Marien, Chinonyerem Okoro, and Michael D Stifelman. "Can selective arterial clamping with fluorescence imaging preserve kidney function during robotic partial nephrectomy?" In: *Urology* 84.2 (2014), pp. 327–334. DOI: 10.1016/j.urology.2014.02.044 (cit. on p. 29).

[235] Shaohui Mei, Yunhao Geng, Junhui Hou, and Qian Du. "Learning hyperspectral images from RGB images via a coarse-to-fine CNN". In: *Science China Information Sciences* 65 (2022), pp. 1–14. DOI: 10.1007/s11432-020-3102-9 (cit. on p. 218).

[236] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *2016 Fourth International Conference on 3D Vision (3DV)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2016, pp. 565–571. DOI: 10.1109/3DV.2016.79 (cit. on p. 126).

[237] Rohit Mittal and Craig M Coopersmith. "Redefining the gut as the motor of critical illness". In: *Trends in molecular medicine* 20.4 (2014), pp. 214–223. DOI: 10.1016/j.molmed.2013.08.004 (cit. on p. 38).

[238] Gaby N. Moawad, Savannah Smith, and Jordan Klebanoff. “The US Perspective of Benefit of Minimally Invasive Surgery: Why Is This Important Now?” In: *Robotic Surgery*. Ed. by Farid Gharagozloo, Vipul R. Patel, Pier Cristoforo Giulianotti, Robert Poston, Rainer Gruessner, and Mark Meyer. Cham: Springer International Publishing, 2021, pp. 1217–1221. ISBN: 978-3-030-53594-0. DOI: 10.1007/978-3-030-53594-0_112 (cit. on p. 28).

[239] Sara Moccia, Sebastian J. Wirkert, Hannes Kenngott, Anant S. Vemuri, Martin Apitz, Benjamin Mayer, Elena De Momi, Leonardo S. Mattos, and Lena Maier-Hein. “Uncertainty-Aware Organ Classification for Surgical Data Science Applications in Laparoscopy”. In: *IEEE Transactions on Biomedical Engineering* 65.11 (Nov. 2018), pp. 2649–2659. ISSN: 0018-9294, 1558-2531. DOI: 10.1109/TBME.2018.2813015. URL: https://ieeexplore.ieee.org/document/8310960 (cit. on pp. 58, 115, 117, 155).

[240] Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, eds. *Neural Networks: Tricks of the Trade: Second Edition*. Vol. 7700. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8. URL: http://link.springer.com/10.1007/978-3-642-35289-8 (cit. on pp. 128, 147).

[241] Michael Moor, Nicolas Bennett, Drago Pleko, Max Horn, Bastian Rieck, Nicolai Meinshausen, Peter Bühlmann, and Karsten Borgwardt. “Predicting sepsis using deep learning across international sites: a retrospective development and validation study”. In: *EClinicalMedicine* 62 (2023). DOI: 10.1016/j.eclinm.2023.102124 (cit. on pp. 10, 178).

[242] Zachary Mostel, Abraham Perl, Matthew Marck, Syed F Mehdi, Barbara Lowell, Sagar Bathija, Ramchandani Santosh, Valentin A Pavlov, Sangeeta S Chavan, and Jesse Roth. “Post-sepsis syndrome–an evolving entity that afflicts survivors of sepsis”. In: *Molecular Medicine* 26 (2020), pp. 1–14. DOI: 10.1186/s10020-019-0132-z (cit. on pp. 32, 41).

[243] Judith R Mourant, Murat Canpolat, C Brocker, O Esponda-Ramos, Tamara M Johnson, A Matanock, K Stetter, and James P Freyer. “Light scattering from cells: the contribution of the nucleus and the effects of proliferative status”. In: *Journal of biomedical optics* 5.2 (2000), pp. 131–137. DOI: 10.1117/1.429979 (cit. on p. 17).

[244] Hala Muaddi, Melanie El Hafid, Woo Jin Choi, Erin Lillie, Charles de Mestral, Avery Nathens, Therese A Stukel, and Paul J Karanicolas. “Clinical Outcomes of Robotic Surgery Compared to Conventional Surgical Approaches (Laparoscopic or Open): A Systematic Overview of Reviews”. In: *Annals of Surgery* 273.3 (2021). DOI: 10.1097/SLA.0000000000003915 (cit. on p. 28).

[245] Christopher JL Murray, Kevin Shunji Ikuta, Fablina Sharara, Lucien Swetschinski, Gisela Robles Aguilar, Authia Gray, Chieh Han, Catherine Bisignano, Puja Rao, Eve Wool, et al. "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis". In: *The lancet* 399.10325 (2022), pp. 629–655. DOI: 10.1016/S0140-6736(21)02724-0 (cit. on p. 42).

[246] Christina Nedeva, Joseph Menassa, and Hamsa Puthalakath. "Sepsis: inflammation is a necessary evil". In: *Frontiers in cell and developmental biology* 7 (2019), p. 108. DOI: 10.3389/fcell.2019.00108 (cit. on p. 36).

[247] Dmitri Nepogodiev, Janet Martin, Bruce Biccard, Alex Makupe, Aneel Bhangu, Adesoji Ademuyiwa, Adewale Oluseye Adisa, Maria-Lorena Aguilera, Sohini Chakrabortee, J. Edward Fitzgerald, Dhruva Ghosh, James C. Glasbey, Ewen M. Harrison, J.C. Allen Ingabire, Hosni Salem, Marie Carmela Lapitan, Ismail Lawani, David Lissauer, Laura Magill, Rachel Moore, Daniel C. Osei-Bordom, Thomas D. Pinkney, Ahmad Uzair Qureshi, Antonio Ramos-De la Medina, Sarah Rayne, Sudha Sundar, Stephen Tabiri, Azmina Verjee, Raul Yepez, O. James Garden, Richard Lilford, Peter Brocklehurst, and Dion G. Morton. "Global burden of postoperative death". In: *The Lancet* 393.10170 (Feb. 2, 2019), p. 401. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(18)33139-8 (cit. on p. 28).

[248] F. Nickel, A. Studier-Fischer, B. Özdemir, J. Odenthal, L.R. Müller, S. Knoedler, K.F. Kowalewski, I. Camplisson, M.M. Allers, M. Dietrich, K. Schmidt, G.A. Salg, H.G. Kenngott, A.T. Billeter, I. Gockel, C. Sagiv, O.E. Hadar, J. Gildenblat, L. Ayala, S. Seidlitz, L. Maier-Hein, and B.P. Müller-Stich. "Optimization of anastomotic technique and gastric conduit perfusion with hyperspectral imaging and machine learning in an experimental model for minimally invasive esophagectomy". In: *European Journal of Surgical Oncology* 51.1 (2025), p. 106908. ISSN: 0748-7983. DOI: 10.1016/j.ejso.2023.04.007 (cit. on pp. 4, 6, 29).

[249] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015. URL: http://neuralnetworksanddeeplearning.com (cit. on pp. 43, 50, 53).

[250] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernardino Romera-Paredes, Christopher Kelly, Alan Karthikesalingam, Carlton Chu, Dawn Carnell, Cheng Boon, Derek D'Souza, Syed Ali Moinuddin, Bethany Garie, Yasmin McQuinlan, Sarah Ireland, Kiarna Hampton, Krystle Fuller, Hugh Montgomery, Geraint Rees, Mustafa Suleyman, Trevor Back, Cían Hughes, Joseph R. Ledsam, and Olaf Ronneberger. "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy". In: *arXiv:1809.04430 [physics, stat]* (Jan. 2021). arXiv: 1809.04430. URL: http://arxiv.org/abs/1809.04430 (cit. on pp. 130, 145).

[251] Aksone Nouvong, Byron Hoogwerf, Emile Mohler, Brian Davis, Azita Tajaddini, and Elizabeth Medenilla. “Evaluation of diabetic foot ulcer healing with hyperspectral imaging of oxyhemoglobin and deoxyhemoglobin”. In: *Diabetes care* 32.11 (2009), pp. 2056–2061. DOI: 10.2337/dc08-2246 (cit. on p. 4).

[252] Augustus Odena, Vincent Dumoulin, and Chris Olah. “Deconvolution and Checkerboard Artifacts”. In: *Distill* (2016). DOI: 10.23915/distill.00003. URL: http://distill.pub/2016/deconv-checkerboard (cit. on p. 53).

[253] T. Y. Ohulchanskyy, A. M. Pliss, and P. N. Prasad. “Biophotonics: Harnessing Light for Biology and Medicine”. In: *Biophotonics: Spectroscopy, Imaging, Sensing, and Manipulation*. Ed. by Baldassare Di Bartolo and John Collins. Dordrecht: Springer Netherlands, 2011, pp. 3–17. ISBN: 978-90-481-9977-8. DOI: 10.1007/978-90-481-9977-8_1 (cit. on p. 15).

[254] Ryan K Orosco, Viridiana J Tapia, Joseph A Califano, Bryan Clary, Ezra EW Cohen, Christopher Kane, Scott M Lippman, Karen Messer, Alfredo Molinolo, James D Murphy, et al. “Positive surgical margins in the 10 most common solid cancers”. In: *Scientific reports* 8.1 (2018), p. 5686. DOI: 10.1038/s41598-018-23403-5 (cit. on p. 31).

[255] Samuel Ortega, Martin Halicek, Himar Fabelo, Raul Guerra, Carlos Lopez, Marylene Lejaune, Fred Godtliebsen, Gustavo M Callico, and Baowei Fei. “Hyperspectral imaging and deep learning for the detection of breast cancer cells in digitized histological images”. In: *Proc SPIE Int Soc Opt Eng* 11320 (Mar. 2020). DOI: 10.1117/12.2548609 (cit. on pp. 3, 31).

[256] Ilze Oshina and Janis Spigulis. “Beer–Lambert law for optical tissue diagnostics: current state of the art and the main limitations”. In: *Journal of biomedical optics* 26.10 (2021), pp. 100901–100901. DOI: 10.1117/1.JBO.26.10.100901 (cit. on p. 26).

[257] Leif Østergaard, A Granfeldt, N Secher, A Tietze, NK Iversen, Morten Skovgaard Jensen, Kristian Kjær Andersen, K Nagenthiraja, P Gutiérrez-Lizardi, K Mouridsen, et al. “Microcirculatory dysfunction and tissue oxygenation in critical illness”. In: *Acta Anaesthesiologica Scandinavica* 59.10 (2015), pp. 1246–1259. DOI: 10.1111/aas.12581 (cit. on p. 179).

[258] Ester Pachyn, Maximilian Aumiller, Christian Freymüller, Matthäus Linek, Veronika Volgger, Alexander Buchner, Adrian Rühm, and Ronald Sroka. “Investigation on the influence of the skin tone on hyperspectral imaging for free flap surgery”. In: *Scientific Reports* 14.1 (June 17, 2024), p. 13979. ISSN: 2045-2322. DOI: 10.1038/s41598-024-64549-9 (cit. on p. 80).

[259] Pan Pan, Longxiang Su, Dawei Liu, and Xiaoting Wang. "Microcirculation-guided protection strategy in hemodynamic therapy". In: *Clinical Hemorheology and Microcirculation* 75.2 (2020), pp. 243–253. DOI: 10.3233/CH-190784 (cit. on p. 182).

[260] Carly J Paoli, Mark A Reynolds, Meenal Sinha, Matthew Gitlin, and Elliott Crouser. "Epidemiology and costs of sepsis in the United Statesan analysis based on timing of diagnosis and severity level". In: *Critical care medicine* 46.12 (2018), pp. 1889–1897. DOI: 10.1097/CCM.0000000000003342 (cit. on p. 32).

[261] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 190).

[262] Antonio Pesce, Stefano Palmucci, Gaetano La Greca, and Stefano Puleo. "Iatrogenic bile duct injury: impact and management challenges". In: *Clinical and experimental gastroenterology* (2019), pp. 121–128. DOI: 10.2147/CEG.S169492 (cit. on p. 31).

[263] Hung Viet Pham, Shangshu Qian, Jiannan Wang, Thibaud Lutellier, Jonathan Rosenthal, Lin Tan, Yaoliang Yu, and Nachiappan Nagappan. "Problems and Opportunities in Training Deep Learning Software Systems: An Analysis of Variance". In: *2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. ISSN: 2643-1572. Sept. 2020, pp. 771–783 (cit. on pp. 128, 190).

[264] Leonardo Piccolo, Kristal Bornillo, Sara Micheli, Marco Sorgato, Mauro Ricotta, Elisa Cimetta, and Giovanni Lucchetta. "A penetration efficiency model for the optimization of solid conical microneedles' geometry". In: *Journal of Micromechanics and Microengineering* 34.2 (2024), p. 025009. DOI: 10.1088/1361-6439/ad1e36 (cit. on p. 16).

[265] Charalampos Pierrakos, Dimitrios Velissaris, Max Bisdorff, John C Marshall, and Jean-Louis Vincent. "Biomarkers of sepsis: time for a reappraisal". In: *Critical Care* 24 (2020), pp. 1–15. DOI: 10.1186/s13054-020-02993-5 (cit. on pp. 12, 178).

[266] Mukund Poddar, Jayson S. Marwaha, William Yuan, Santiago Romero-Brufau, and Gabriel A. Brat. "An operational guide to translational clinical machine learning in academic medical centers". In: *npj Digital Medicine* 7.1 (May 17, 2024), p. 129. ISSN: 2398-6352. DOI: 10.1038/s41746-024-01094-9 (cit. on pp. 218, 220).

[267] Rachel Pool, Hernando Gomez, and John A Kellum. "Mechanisms of organ dysfunction in sepsis". In: *Critical care clinics* 34.1 (2018), pp. 63–80. DOI: 10.1016/j.ccc.2017.08.003 (cit. on pp. 36, 38).

[268] Assaf Potruch, Asaf Schwartz, and Yaron Ilan. “The role of bacterial translocation in sepsis: a new target for therapy”. In: *Therapeutic Advances in Gastroenterology* 15 (2022), p. 17562848221094214. DOI: 10.1177/17562848221094214 (cit. on p. 38).

[269] Scott Prahl and Steve Jacques. *Assorted spectra*. https://omlc.org/spectra/. [Online; accessed 2024-09-17] (cit. on p. 18).

[270] Saurabh Prasad and Jocelyn Chanussot, eds. *Hyperspectral Image Analysis: Advances in Machine Learning and Signal Processing*. Advances in Computer Vision and Pattern Recognition. Cham: Springer International Publishing, 2020. ISBN: 978-3-030-38617-7. DOI: 10.1007/978-3-030-38617-7. URL: http://link.springer.com/10.1007/978-3-030-38617-7 (cit. on p. 20).

[271] Hallie C Prescott and Derek C Angus. “Enhancing recovery from sepsis: a review”. In: *Jama* 319.1 (2018), pp. 62–75. DOI: 10.1001/jama.2017.17687 (cit. on p. 32).

[272] Sami Puustinen, Hana Vrzáková, Joni Hyttinen, Tuomas Rauramaa, Pauli Fält, Markku Hauta-Kasari, Roman Bednarik, Timo Koivisto, Susanna Rantala, Mikael von und zu Fraunberg, Juha E. Jääskeläinen, and Antti-Pekka Elomaa. “Hyperspectral Imaging in Brain Tumor SurgeryEvidence of Machine Learning-Based Performance”. In: *World Neurosurgery* 175 (2023), e614–e635. ISSN: 1878-8750. DOI: 10.1016/j.wneu.2023.03.149. URL: https://www.sciencedirect.com/science/article/pii/S1878875023004734 (cit. on pp. 217, 218).

[273] Ahmad Bin Qasim, Alessandro Motta, Alexander Studier-Fischer, Jan Sellner, Leonardo Ayala, Marco Hübner, Marc Bressan, Berkin Özdemir, Karl Friedrich Kowalewski, Felix Nickel, Silvia Seidlitz, and Lena Maier-Hein. “Test-time augmentation with synthetic data addresses distribution shifts in spectral imaging”. In: *International Journal of Computer Assisted Radiology and Surgery* (Mar. 14, 2024). ISSN: 1861-6429. DOI: 10.1007/s11548-024-03085-3 (cit. on p. 172).

[274] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. “Improving language understanding by generative pre-training”. In: () (cit. on pp. 10, 43).

[275] Lisa Raia and Lara Zafrani. “Endothelial activation and microcirculatory disorders in sepsis”. In: *Frontiers in Medicine* 9 (2022), p. 907992. DOI: 10.3389/fmed.2022.907992 (cit. on p. 178).

[276] Laura Elena Raileanu and Kilian Stoffel. “Theoretical Comparison between the Gini Index and Information Gain Criteria”. In: *Annals of Mathematics and Artificial Intelligence* 41.1 (May 2004), pp. 77–93. DOI: 10.1023/B:AMAI.0000018580.96245.c6 (cit. on p. 46).

[277] Anna-Maria Raita-Hakola, Leevi Annala, Vivian Lindholm, Roberts Trops, Antti Näsilä, Heikki Saari, Annamari Ranki, and Ilkka Pölönen. "FPI Based Hyperspectral Imager for the Complex SurfacesCalibration, Illumination and Applications". In: *Sensors* 22.9 (2022), p. 3420. DOI: 10.3390/s22093420 (cit. on p. 218).

[278] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. *Zero-Shot Text-to-Image Generation*. 2021. arXiv: 2102.12092 [cs.CV]. URL: https://arxiv.org/abs/2102.12092 (cit. on pp. 10, 43).

[279] Daniele Ravì, Himar Fabelo, Gustavo Marrero Callic, and Guang-Zhong Yang. "Manifold Embedding and Semantic Segmentation for Intraoperative Guidance With Hyperspectral Brain Imaging". In: *IEEE Transactions on Medical Imaging* 36.9 (Sept. 2017). Conference Name: IEEE Transactions on Medical Imaging, pp. 1845–1857. ISSN: 1558-254X. DOI: 10.1109/TMI.2017.2695523 (cit. on pp. 115, 117, 118, 155).

[280] London RCoP. "National Early Warning Score (NEWS): standardising the assessment of acute-illness severity in the NHS". In: *Report of working party. London: Royal College of Physicians* (2012) (cit. on p. 182).

[281] Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Kai Han, Andrea Vedaldi, and Andrew Zisserman. "Semi-supervised learning with scarce annotations". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, pp. 762–763 (cit. on p. 220).

[282] Ashfaq Ur Rehman, Mingyu Li, Binjian Wu, Yasir Ali, Salman Rasheed, Sana Shaheen, Xinyi Liu, Ray Luo, and Jian Zhang. "Role of Artificial Intelligence in Revolutionizing Drug Discovery". In: *Fundamental Research* (2024). ISSN: 2667-3258. DOI: 10.1016/j.fmre.2024.04.021 (cit. on pp. 10, 43).

[283] Annika Reinke, Minu D. Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A. Emre Kavur, Tim Rädsch, Carole H. Sudre, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Florian Buettner, M. Jorge Cardoso, Veronika Cheplygina, Jianxu Chen, Evangelia Christodoulou, Beth A. Cimini, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Ben Glocker, Patrick Godau, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Fabian Isensee, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Jens Kleesiek, Florian Kofler, Thijs Kooi, Annette Kopp-Schneider, Michal Kozubek, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, et al. "Understanding metric-related pitfalls in image analysis validation". In: *Nature Methods* 21.2 (Feb. 1, 2024), pp. 182–194. ISSN: 1548-7105. DOI: 10.1038/s41592-023-02150-0 (cit. on pp. 129, 130, 144, 160).

[284] Annika Reinke, Minu D. Tizabi, Carole H. Sudre, Matthias Eisenmann, Tim Rädsch, Michael Baumgartner, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Peter Bankhead, Arriel Benis, Matthew Blaschko, Florian Büttner, M. Jorge Cardoso, Jianxu Chen, Veronika Cheplygina, Evangelia Christodoulou, Beth Cimini, Gary S. Collins, Sandy Engelhardt, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Ben Glocker, Patrick Godau, Robert Haase, Fred Hamprecht, Daniel A. Hashimoto, Doreen Heckmann-Nötzel, Peter Hirsch, Michael M. Hoffman, Merel Huisman, Fabian Isensee, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, A. Emre Kavur, Hannes Kenngott, Jens Kleesiek, Andreas Kleppe, Sven Kohler, Florian Kofler, Annette Kopp-Schneider, Thijs Kooi, Michal Kozubek, et al. *Common Limitations of Image Processing Metrics: A Picture Story*. 2023. arXiv: 2104.05642 [eess.IV] (cit. on pp. 129, 144, 160).

[285] Mauricio Reyes, Miguel A Gonzalez Ballester, Zhixi Li, Nina Kozic, See Chin, Ronald M Summers, and Marius George Linguraru. "ANATOMICAL VARIABILITY OF ORGANS VIA PRINCIPAL FACTOR ANALYSIS FROM THE CONSTRUCTION OF AN ABDOMINAL PROBABILISTIC ATLAS". In: *Proc IEEE Int Symp Biomed Imaging* 2009 (2009), pp. 682–685 (cit. on p. 29).

[286] Andrew Rhodes, Laura E Evans, Waleed Alhazzani, Mitchell M Levy, Massimo Antonelli, Ricard Ferrer, Anand Kumar, Jonathan E Sevransky, Charles L Sprung, Mark E Nunnally, et al. "Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016". In: *Intensive care medicine* 43 (2017), pp. 304–377. DOI: 10.1007/s00134-017-4683-6 (cit. on p. 41).

[287] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. "Kornia: an Open Source Differentiable Computer Vision Library for PyTorch". In: *Winter Conference on Applications of Computer Vision*. 2020. URL: https://arxiv.org/pdf/1910.02190.pdf (cit. on pp. 127, 173, 215).

[288] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, Alessandro Ruggiero, Anna Korhonen, Emily Jefferson, Emmanuel Ako, Georg Langs, Ghassem Gozaliasl, Guang Yang, Helmut Prosch, Jacobus Preller, Jan Stanczuk, Jing Tang, Johannes Hofmanninger, Judith Babar, Lorena Escudero Sánchez, Muhunthan Thillai, Paula Martin Gonzalez, Philip Teare, Xiaoxiang Zhu, Mishal Patel, Conor Cafolla, Hojjat Azadbakht, Joseph Jacob, Josh Lowe, Kang Zhang, Kyle Bradley, Marcel Wassin, Markus Holzer, Kangyu Ji, Maria Delgado Ortet, Tao Ai, Nicholas Walton, Pietro Lio, Samuel Stranks, Tolou Shadbahr, Weizhe Lin, Yunfei Zha, Zhangming Niu, et al. "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans". In: *Na-*

ture Machine Intelligence 3.3 (Mar. 1, 2021), pp. 199–217. ISSN: 2522-5839. DOI: 10.1038/s42256-021-00307-0 (cit. on pp. 10, 109).

[289] Sophie Romann, Tristan Wagner, Shadi Katou, Stefan Reuter, Thomas Vogel, Felix Becker, Haluk Morgul, Philipp Houben, Philip Wahl, Andreas Pascher, and Sonia Radunz. "Hyperspectral Imaging for Assessment of Initial Graft Function in Human Kidney Transplantation". In: *Diagnostics* 12.5 (2022). ISSN: 2075-4418. DOI: 10.3390/diagnostics12051194 (cit. on p. 80).

[290] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695 (cit. on pp. 10, 43).

[291] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4 (cit. on pp. 53, 126).

[292] Mélanie Roschewitz, Raghav Mehta, Charles Jones, and Ben Glocker. *Automatic dataset shift identification to support safe deployment of medical imaging AI*. 2025. DOI: 10.48550/arXiv.2411.07940. arXiv: 2411.07940 [cs.AI] (cit. on p. 80).

[293] Norman Rose, Claudia Matthäus-Krämer, Daniel Schwarzkopf, André Scherag, Sebastian Born, Konrad Reinhart, and Carolin Fleischmann-Struzek. "Association between sepsis incidence and regional socioeconomic deprivation and health care capacity in Germany–an ecological study". In: *BMC Public Health* 21 (2021), pp. 1–11. DOI: 10.1186/s12889-021-11629-4 (cit. on p. 34).

[294] Tobias Ross, Annika Reinke, Peter M. Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Pablo Arbeláez, Gui-Bin Bian, Sebastian Bodenstedt, Jon Lindström Bolmgren, Laura Bravo-Sánchez, Hua-Bin Chen, Cristina González, Dong Guo, Pål Halvorsen, Pheng-Ann Heng, Enes Hosgor, Zeng-Guang Hou, Fabian Isensee, Debesh Jha, Tingting Jiang, Yueming Jin, Kadir Kirtac, Sabrina Kletz, Stefan Leger, Zhixuan Li, Klaus H. Maier-Hein, Zhen-Liang Ni, Michael A. Riegler, Klaus Schoeffmann, Ruohua Shi, Stefanie Speidel, Michael Stenzel, Isabell Twick, Gutai Wang, Jiacheng Wang, Liansheng Wang, Lu Wang, Yujie Zhang, Yan-Jie Zhou, Lei Zhu, Manuel Wiesenfarth, Annette Kopp-Schneider, Beat P. Müller-Stich, and Lena Maier-Hein. "Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the ROBUST-MIS 2019 challenge". In: *Medical Image Analysis* 70 (May 2021), p. 101920. ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101920. URL: https:

//www.sciencedirect.com/science/article/pii/S136184152030284X (cit. on p. 8).

[295] Ignacio Rubio, Marcin F Osuchowski, Manu Shankar-Hari, Tomasz Skirecki, Martin Sebastian Winkler, Gunnar Lachmann, Paul La Rosée, Guillaume Monneret, Fabienne Venet, Michael Bauer, et al. "Current gaps in sepsis immunology: new opportunities for translational research". In: *The Lancet infectious diseases* 19.12 (2019), e422–e436. DOI: 10.1016/S1473-3099(19)30567-5 (cit. on p. 38).

[296] Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, et al. "Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study". In: *The Lancet* 395.10219 (2020), pp. 200–211. DOI: 10.1016/S0140-6736(19)32989-7 (cit. on pp. 10, 32, 33, 36, 178, 206).

[297] Tobias Rueckert, Daniel Rueckert, and Christoph Palm. "Methods and datasets for segmentation of minimally invasive surgical instruments in endoscopic images and videos: A review of the state of the art". In: *Computers in Biology and Medicine* 169 (2024), p. 107929. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2024.107929 (cit. on p. 114).

[298] Yasser Sakr, Suzana M Lobo, Rui P Moreno, Herwig Gerlach, V Marco Ranieri, Argyris Michalopoulos, Jean-Louis Vincent, and SOAP Investigators jlvincen@ulb. ac. be. "Patterns and early evolution of organ failure in the intensive care unit and their relation to outcome". In: *Critical care* 16 (2012), pp. 1–9. DOI: 10.1186/cc11868 (cit. on p. 37).

[299] Gabriel Sandblom. "Grand challenges in visceral surgery". In: *Front Surg* 9 (Aug. 2022), p. 1005046. DOI: 10.3389/fsurg.2022.1005046 (cit. on pp. 28, 217).

[300] Angelo Sassaroli and Sergio Fantini. "Comment on the modified Beer–Lambert law for scattering media". In: *Physics in Medicine & Biology* 49.14 (2004), N255. DOI: 10.1088/0031-9155/49/14/N07 (cit. on p. 26).

[301] Christopher M Sauer, Li-Ching Chen, Stephanie L Hyland, Armand Girbes, Paul Elbers, and Leo A Celi. "Leveraging electronic health records for data science: common pitfalls and how to avoid them". In: *The Lancet Digital Health* 4.12 (2022), e893–e898. DOI: 10.1016/S2589-7500(22)00154-6 (cit. on p. 178).

[302] Paul Scheikl, Stefan Laschewski, Anna Kisilenko, Tornike Davitashvili, Benjamin Müller, Manuela Capek, Beat Müller, Martin Wagner, and Franziska Ullrich. "Deep learning for semantic segmentation of organs and tissues in laparoscopic surgery". In: *Current Directions in Biomedical Engineering* 6 (Sept. 2020), p. 20200016. DOI: 10.1515/cdbme-2020-0016 (cit. on pp. 8, 116, 155).

[303] Dominik Scherer, Andreas Müller, and Sven Behnke. "Evaluation of pooling operations in convolutional architectures for object recognition". In: *International conference on artificial neural networks*. Springer. 2010, pp. 92–101. DOI: 10.1007/978-3-642-15825-4_10 (cit. on p. 53).

[304] Thomas Schmoch, Patrick Möhnle, Markus A Weigand, Josef Briegel, Michael Bauer, Frank Bloos, Patrick Meybohm, Didier Keh, Markus Löffler, Gunnar Elke, Thorsten Brenner, Holger Bogatsch, and SepNet–Critical Care Trials Group. "The prevalence of sepsis-induced coagulopathy in patients with sepsis - a secondary analysis of two German multicenter randomized controlled trials". In: *Ann Intensive Care* 13.1 (Jan. 2023), p. 3. DOI: 10.1186/s13613-022-01093-7 (cit. on p. 37).

[305] Marcel André Schneider, Daniel Gero, Matteo Müller, Karoline Horisberger, Andreas Rickenbacher, and Matthias Turina. "Inequalities in access to minimally invasive general surgery: a comprehensive nationwide analysis across 20ăyears". In: *Surgical Endoscopy* 35.11 (Nov. 1, 2021), pp. 6227–6243. ISSN: 1432-2218. DOI: 10.1007/s00464-020-08123-0 (cit. on p. 28).

[306] Silvia Seidlitz, Katharina Hölzl, Ayca von Garrel, Jan Sellner, Stephan Katzenschlager, Tobias Hölle, Dania Fischer, Maik von der Forst, Felix C. F. Schmitt, Alexander Studier-Fischer, Markus A. Weigand, Lena Maier-Hein, and Maximilian Dietrich. "AI-powered skin spectral imaging enables instant sepsis diagnosis and outcome prediction in critically ill patients". In: *Science Advances* 11.29 (2025), eadw1968. DOI: 10.1126/sciadv.adw1968 (cit. on pp. 177, 181, 182, 193, 194, 196, 197, 200, 201, 203, 204, 214, 215, 224, 255, 256, 266, 267).

[307] Silvia Seidlitz, Jan Sellner, Jan Odenthal, Berkin Özdemir, Alexander Studier-Fischer, Samuel Knödler, Leonardo Ayala, Tim J. Adler, Hannes G. Kenngott, Minu Tizabi, Martin Wagner, Felix Nickel, Beat P. Müller-Stich, and Lena Maier-Hein. *Robust deep learning-based semantic organ segmentation in hyperspectral images*. Oral presentation at the 13th International Conference on Information Processing in Computer-Assisted Interventions (IPCAI), Tokyo, Japan. June 7, 2022 (cit. on pp. 113, 213, 224).

[308] Silvia Seidlitz, Jan Sellner, Jan Odenthal, Berkin Özdemir, Alexander Studier-Fischer, Samuel Knödler, Leonardo Ayala, Tim J. Adler, Hannes G. Kenngott, Minu Tizabi, Martin Wagner, Felix Nickel, Beat P. Müller-Stich, and Lena Maier-Hein. "Robust deep learning-based semantic organ segmentation in hyperspectral images". In: *Medical Image Analysis* 80 (Aug. 2022), p. 102488. ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102488 (cit. on pp. 8, 31, 52, 86, 113, 117, 123, 124, 127, 128, 135–137, 139, 140, 142, 143, 146, 147, 150, 155, 188, 213, 215, 224, 247–250).

[309] Silvia Seidlitz, Jan Sellner, Alexander Studier-Fischer, Alessandro Motta, Berkin Özdemir, Beat P. Müller-Stich, Felix Nickel, and Lena Maier-Hein. "Handling

Geometric Domain Shifts in Semantic Segmentation of Surgical RGB and Hyperspectral Images". In: *arXiv preprint arXiv:2408.15373* (2024). DOI: 10.48550/arXiv.2408.15373 (cit. on pp. 117, 153, 158, 163, 165, 167, 168, 170, 224, 251–253).

[310] Silvia Seidlitz, Alexander Studier-Fischer, Maximilian Dietrich, Ayca von Garrel, Katharina Hölzl, Felix Nickel, Markus A. Weigand, and Lena Maier-Hein. *Shedding light on hidden factors: Unvealing biases in medical hyperspectral images*. Oral presentation at the 13th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Athens, Greece. Nov. 2, 2023 (cit. on pp. 79, 213, 223).

[311] Jan Sellner. "Generalizable Surgical Scene Segmentation of Hyperspectral Images". PhD thesis. 2024. DOI: 10.11588/heidok.00035083 (cit. on pp. 16, 23, 30, 49, 51, 52, 113, 122, 124, 127, 135–137, 139, 140, 142, 143, 146, 147, 150, 153, 155, 158, 161, 165, 167, 168, 170, 247–250, 252, 253).

[312] Jan Sellner and Silvia Seidlitz. *Hyperspectral Tissue Classification*. Version v0.0.15. Feb. 5, 2024. DOI: 10.5281/zenodo.6577614. URL: https://github.com/IMSY-DKFZ/htc (cit. on pp. 113, 151, 173, 207, 215, 221).

[313] Jan Sellner, Silvia Seidlitz, and Lena Maier-Hein. *Dealing with I/O bottlenecks in high-throughput model training*. Poster presentation at the PyTorch Conference 2023, San Francisco, United States of America. Oct. 16, 2023. URL: https://e130-hyperspectal-tissue-classification.s3.dkfz.de/figures/PyTorchConference_Poster.pdf (cit. on pp. 171, 172, 215).

[314] Jan Sellner, Silvia Seidlitz, Alexander Studier-Fischer, Alessandro Motta, Berkin Özdemir, Beat Peter Müller-Stich, Felix Nickel, and Lena Maier-Hein. "Semantic Segmentation of Surgical Hyperspectral Images Under Geometric Domain Shifts". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor. Cham: Springer Nature Switzerland, 2023, pp. 618–627. ISBN: 978-3-031-43996-4. DOI: 10.1007/978-3-031-43996-4_59 (cit. on pp. 86, 117, 153, 158, 165, 167, 168, 170, 214, 224, 252, 253).

[315] Jan Sellner, Alexander Studier-Fischer, Ahmad Bin Qasim, Silvia Seidlitz, Nicholas Schreck, Minu Tizabi, Manuel Wiesenfarth, Annette Kopp-Schneider, Samuel Knödler, Caelan Max Haney, Gabriel Salg, Berkin Özdemir, Maximilian Dietrich, Maurice Stephan Michel, Felix Nickel, Karl-Friedrich Kowalewski, and Lena Maier-Hein. "Xeno-learning: knowledge transfer across species in deep learning-based spectral image analysis". In: *Accepted at Nature Biomedical Engineering, arXiv preprint arXiv:2410.19789* (2024). DOI: 10.48550/arXiv.2410.19789 (cit. on pp. 117, 151, 172, 215, 220, 221).

[316] Gregor Sersa, Urban Simoncic, Matija Milanic, et al. "Imaging perfusion changes in oncological clinical applications by hyperspectral imaging: a literature review". In: *Radiology and Oncology* 56.4 (2022), pp. 420–429. DOI: 10.2478/raon-2022-0051 (cit. on p. 29).

[317] Connor Shorten and Taghi M. Khoshgoftaar. "A survey on Image Data Augmentation for Deep Learning". In: *Journal of Big Data* 6.1 (July 6, 2019), p. 60. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0197-0 (cit. on p. 154).

[318] Oleksii Sidorov. "Conditional gans for multi-illuminant color constancy: Revolution or yet another approach?" In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2019 (cit. on p. 58).

[319] Patrice Y. Simard, Dave Steinkraus, and John C. Platt. "Best practices for convolutional neural networks applied to visual document analysis". In: *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings*. 2003, pp. 958–963. DOI: 10.1109/ICDAR.2003.1227801 (cit. on p. 160).

[320] Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)". In: *JAMA* 315.8 (Feb. 2016), pp. 801–810. ISSN: 0098-7484. DOI: 10.1001/jama.2016.0287. (Visited on 05/20/2021) (cit. on pp. 10, 32, 34, 36, 39, 40, 182–184).

[321] Krishna Kumar Singh and Yong Jae Lee. "Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-Supervised Object and Action Localization". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, pp. 3544–3553 (cit. on pp. 154, 157).

[322] Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. "Don't Decay the Learning Rate, Increase the Batch Size". In: Feb. 2018. URL: https://openreview.net/forum?id=B1Yy1BxCZ (cit. on p. 128).

[323] Ace St John, Ilaria Caturegli, Natalia S Kubicki, and Stephen M Kavic. "The Rise of Minimally Invasive Surgery: 16 Year Analysis of the Progressive Replacement of Open Surgery with Laparoscopy". In: *JSLS* 24.4 (Oct. 2020). DOI: 10.4293/JSLS.2020.00076 (cit. on pp. 28, 172).

[324] Nathan P Staff, JaNean Engelstad, Christopher J Klein, Kimberly K Amrami, Robert J Spinner, Peter J Dyck, Mark A Warner, Mary E Warner, and P James B Dyck. "Post-surgical inflammatory neuropathy". In: *Brain* 133.10 (2010), pp. 2866–2880. DOI: 10.1093/brain/awq252 (cit. on p. 31).

[325] Andrew Stockman and Lindsay T. Sharpe. “The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype”. In: *Vision Research* 40.13 (2000), pp. 1711–1737. ISSN: 0042-6989. DOI: 10.1016/S0042-6989(00)00021-3. URL: https://www.sciencedirect.com/science/article/pii/S0042698900000213 (cit. on p. 20).

[326] Alexander Studier-Fischer, Marc Bressan, Ahmad bin Qasim, Berkin Özdemir, Jan Sellner, Silvia Seidlitz, Caelán Haney, Luisa Egen, Maurice Michel, Maximilian Dietrich, Gabriel Alexander Salg, Franck Billmann, Henrik Nienhüser, Thilo Hackert, Beat Müller-Stich, Lena Maier-Hein, Felix Nickel, and Karl-Friedrich Kowalewski. “Spectral characterization of intraoperative renal perfusion using hyperspectral imaging and artificial intelligence”. In: *Scientific Reports* 14.1 (July 27, 2024), p. 17262. ISSN: 2045-2322. DOI: 10.1038/s41598-024-68280-3 (cit. on pp. 4, 6, 29, 80).

[327] Alexander Studier-Fischer, Berkin Özdemir, Maike Rees, Leonardo Ayala, Silvia Seidlitz, Jan Sellner, Karl-Friedrich Kowalewski, Caelán Max Haney, Jan Odentahl, Samuel Knödler, Maximilian Dietrich, Daniel Gruneberg, Thorsten Brenner, Karsten Schmidt, Felix Carl Fabian Schmitt, Markus A. Weigand, Gabriel Alexander Salg, Anna Dupree, Henrik Nienhüser, Arianeb Mehrabi, Thilo Hackert, Beat Müller-Stich, Lena Maier-Hein, and Felix Nickel. “Crystalloid volume versus catecholamines for management of hemorrhagic shock during esophagectomy – assessment of microcirculatory tissue oxygenation of the gastric conduit in a porcine model using hyperspectral imaging – an experimental study”. In: *International Journal of Surgery* (9900). DOI: 10.1097/JS9.0000000000001849 (cit. on pp. 4, 6, 29, 80, 108, 206).

[328] Alexander Studier-Fischer, Silvia Seidlitz, Jan Sellner, Marc Bressan, Berkin Özdemir, Leonardo Ayala, Jan Odenthal, Samuel Knoedler, Karl-Friedrich Kowalewski, Caelán Max Haney, Gabriel Salg, Maximilian Dietrich, Hannes Kenngott, Ines Gockel, Thilo Hackert, Beat Peter Müller-Stich, Lena Maier-Hein, and Felix Nickel. “HeiPorSPECTRAL - the Heidelberg Porcine HyperSPECTRAL Imaging Dataset of 20 Physiological Organs”. In: *Scientific Data* 10.1 (June 24, 2023), p. 414. ISSN: 2052-4463. DOI: 10.1038/s41597-023-02315-8. URL: https://heiporspectral.org (cit. on pp. 215, 220, 221).

[329] Alexander Studier-Fischer, Silvia Seidlitz, Jan Sellner, Berkin Özdemir, Manuel Wiesenfarth, Leonardo Ayala, Jan Odenthal, Samuel Knödler, Karl Friedrich Kowalewski, Caelán Max Haney, Isabella Camplisson, Maximilian Dietrich, Karsten Schmidt, Gabriel Alexander Salg, Hannes Götz Kenngott, Tim Julian Adler, Nicholas Schreck, Annette Kopp-Schneider, Klaus Maier-Hein, Lena Maier-Hein, Beat Peter Müller-Stich, and Felix Nickel. “Spectral organ fingerprints for machine learning-based intraoperative tissue classification with hy-

perspectral imaging in a porcine model". In: *Scientific Reports* 12.1 (June 30, 2022), p. 11028. ISSN: 2045-2322. DOI: 10.1038/s41598-022-15040-w (cit. on pp. 80, 119, 125, 187, 188).

[330] Jonah J. Stulberg, Reiping Huang, Lindsey Kreutzer, Kristen Ban, Bradley J. Champagne, Scott R. Steele, Julie K. Johnson, Jane L. Holl, Caprice C. Greenberg, and Karl Y. Bilimoria. "Association Between Surgeon Technical Skills and Patient Outcomes". In: *JAMA Surgery* 155.10 (Oct. 1, 2020), pp. 960–968. ISSN: 2168-6254. DOI: 10.1001/jamasurg.2020.3007 (cit. on p. 28).

[331] Robert Sucher, Alvanos Athanasios, Hannes Köhler, Tristan Wagner, Maximilian Brunotte, Andri Lederer, Ines Gockel, and Daniel Seehofer. "Hyperspectral Imaging (HSI) in anatomic left liver resection". In: *International journal of surgery case reports* 62 (2019), pp. 108–111. DOI: 10.1016/j.ijscr.2019.08.025 (cit. on p. 29).

[332] Robert Sucher, Tristan Wagner, Hannes Köhler, Elisabeth Sucher, Hanna Quice, Sebastian Recknagel, Andri Lederer, Hans Michael Hau, Sebastian Rademacher, Stefan Schneeberger, Gerald Brandacher, Ines Gockel, and Daniel Seehofer. "Hyperspectral Imaging (HSI) of Human Kidney Allografts". In: *Ann Surg* 276.1 (Nov. 2020), e48–e55. DOI: 10.1097/SLA.0000000000004429 (cit. on pp. 4, 6, 29).

[333] James W Suliburk, Quentin M Buck, Chris J Pirko, Nader N Massarweh, Neal R Barshes, Hardeep Singh, and Todd K Rosengart. "Analysis of human performance deficiencies associated with surgical adverse events". In: *JAMA network open* 2.7 (2019), e198067–e198067. DOI: 10.1001/jamanetworkopen.2019.8067 (cit. on p. 28).

[334] Qiyang Sun, Alican Akman, and Björn W. Schuller. *Explainable Artificial Intelligence for Medical Applications: A Review*. 2024. DOI: 10.48550/arXiv.2412.01829. arXiv: 2412.01829 [cs.LG] (cit. on p. 219).

[335] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?" In: *IEEE Transactions on Medical Imaging* 35.5 (May 2016), pp. 1299–1312. ISSN: 1558-254X. DOI: 10.1109/tmi.2016.2535302 (cit. on p. 189).

[336] Mingxing Tan and Quoc Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 6105–6114. URL: https://proceedings.mlr.press/v97/tan19a.html (cit. on p. 126).

[337] Ran Tao, Janek Gröhl, Lina Hacker, Antonio Pifferi, Darren Roblyer, and Sarah E. Bohndiek. “Tutorial on methods for estimation of optical absorption and scattering properties of tissue”. In: *Journal of Biomedical Optics* 29.6 (2024), p. 060801. DOI: 10.1117/1.JBO.29.6.060801 (cit. on p. 85).

[338] R Houston Thompson, Igor Frank, Christine M Lohse, Ismail R Saad, Amr Fergany, Horst Zincke, Bradley C Leibovich, Michael L Blute, and Andrew C Novick. “The impact of ischemia time during open nephron sparing surgery on solitary kidneys: a multi-institutional study”. In: *The Journal of urology* 177.2 (2007), pp. 471–476. DOI: 10.1016/j.juro.2006.09.036 (cit. on p. 29).

[339] Stojan Trajanovski, Caifeng Shan, Pim J. C. Weijtmans, Susan G. Brouwer de Koning, and Theo J. M. Ruers. “Tongue Tumor Detection in Hyperspectral Images Using Deep Learning Semantic Segmentation”. In: *IEEE transactions on bio-medical engineering* 68.4 (Apr. 2021), pp. 1330–1340. ISSN: 1558-2531. DOI: 10.1109/TBME.2020.3026683 (cit. on pp. 115, 117, 118, 155).

[340] Stojan Trajanovski, Caifeng Shan, Pim J. C. Weijtmans, Susan G. Brouwer de Koning, and Theo J. M. Ruers. *Tumor Semantic Segmentation in Hyperspectral Images using Deep Learning*. 2019. URL: https://openreview.net/forum?id=ryeAXGw79V (cit. on pp. 115, 117, 155).

[341] Minh H Tran and Baowei Fei. “Compact and ultracompact spectral imagers: technology and applications in biomedical imaging”. In: *Journal of biomedical optics* 28.4 (2023), pp. 040901–040901. DOI: 10.1117/1.JBO.28.4.040901 (cit. on p. 218).

[342] Stephen Trzeciak, R Phillip Dellinger, Joseph E Parrillo, Massimiliano Guglielmi, Jasmeet Bajaj, Nicole L Abate, Ryan C Arnold, Susan Colilla, Sergio Zanotti, Steven M Hollenberg, et al. “Early microcirculatory perfusion derangements in patients with severe sepsis and septic shock: relationship to hemodynamics, oxygen transport, and survival”. In: *Annals of emergency medicine* 49.1 (2007), pp. 88–98. DOI: 10.1016/j.annemergmed.2006.08.021 (cit. on p. 178).

[343] Sheng-Hao Tseng, Paulo Bargo, Anthony Durkin, and Nikiforos Kollias. “Chromophore concentrations, absorption and scattering properties of human skin in-vivo”. In: *Opt Express* 17.17 (Aug. 2009), pp. 14599–14617. DOI: 10.1364/oe.17.014599 (cit. on p. 17).

[344] Claudio Urrea, Yainet Garcia-Garcia, and John Kern. “Improving Surgical Scene Semantic Segmentation through a Deep Learning Architecture with Attention to Class Imbalance”. In: *Biomedicines* 12.6 (2024). ISSN: 2227-9059. DOI: 10.3390/biomedicines12061309. URL: https://www.mdpi.com/2227-9059/12/6/1309 (cit. on pp. 116, 155).

[345] István Vadász, Laura A Dada, Arturo Briva, Humberto E Trejo, Lynn C Welch, Jiwang Chen, Péter T Tóth, Emilia Lecuona, Lee A Witters, Paul T Schumacker, et al. "AMP-activated protein kinase regulates CO 2-induced alveolar epithelial dysfunction in rats and human cells by promoting Na, K-ATPase endocytosis". In: *The Journal of clinical investigation* 118.2 (2008), pp. 752–762. DOI: 10.1172/JCI29723 (cit. on p. 38).

[346] Joost Van De Weijer, Theo Gevers, and Arjan Gijsenij. "Edge-based color constancy". In: *IEEE Transactions on image processing* 16.9 (2007), pp. 2207–2214. DOI: 10.1109/TIP.2007.901808 (cit. on p. 60).

[347] Sudhir Varma and Richard Simon. "Bias in error estimation when using cross-validation for model selection". In: *BMC bioinformatics* 7 (2006), pp. 1–8. DOI: 10.1186/1471-2105-7-91 (cit. on p. 191).

[348] Ashish Vaswani. "Attention is all you need". In: *arXiv preprint arXiv:1706.03762* (2017) (cit. on pp. 10, 43).

[349] J -L Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PeterM Suter, and Lambertius G Thijs. "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure". In: *Intensive Care Medicine* 22 (1996), pp. 707–710 (cit. on p. 39).

[350] Jean-Louis Vincent. "Annual Update in Intensive Care and Emergency Medicine 2023". In: Cham: Springer Nature Switzerland, 2023. DOI: 10.1007/978-3-031-23005-9 (cit. on pp. 10, 42, 178).

[351] Jean-Louis Vincent, Yasser Sakr, Mervyn Singer, Ignacio Martin-Loeches, Flavia R Machado, John C Marshall, Simon Finfer, Paolo Pelosi, Luca Brazzi, Dita Aditianingsih, et al. "Prevalence and outcomes of infection among patients in intensive care units in 2017". In: *Jama* 323.15 (2020), pp. 1478–1487. DOI: 10.1001/jama.2020.2717 (cit. on p. 34).

[352] KR8872660 Walley. "Heterogeneity of oxygen delivery impairs oxygen extraction by peripheral tissues: theory". In: *Journal of applied physiology* 81.2 (1996), pp. 885–894. DOI: 10.1152/jappl.1996.81.2.885 (cit. on p. 179).

[353] An Wang, Mobarakol Islam, Mengya Xu, and Hongliang Ren. "Rethinking Surgical Instrument Segmentation: A Background Image Can Be All You Need". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Ed. by Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li. Cham: Springer Nature Switzerland, 2022, pp. 355–364. ISBN: 978-3-031-16449-1 (cit. on p. 157).

[354] Fei Wang and Albert J. P. Theuwissen. "Temperature Effect on the Linearity Performance of a CMOS Image Sensor". In: *IEEE Sensors Letters* 2.3 (2018), pp. 1–4. DOI: 10.1109/LSENS.2018.2860990 (cit. on p. 80).

[355] Hui Wang, Hong Ding, Zi-Yan Wang, and Kun Zhang. “Research progress on microcirculatory disorders in septic shock: A narrative review”. In: *Medicine* 103.8 (2024), e37273. DOI: 10.1097/MD.0000000000037273 (cit. on pp. 178, 179).

[356] Lihong Wang and Steven L Jacques. “Monte Carlo modeling of light transport in multi-layered tissues in standard C”. In: *The University of Texas, MD Anderson Cancer Center, Houston* 4.11 (1992) (cit. on p. 27).

[357] Lihong Wang, Steven L. Jacques, and Liqiong Zheng. “MCMLMonte Carlo modeling of light transport in multi-layered tissues”. In: *Computer Methods and Programs in Biomedicine* 47.2 (1995), pp. 131–146. ISSN: 0169-2607. DOI: 10.1016/0169-2607(95)01640-F (cit. on p. 27).

[358] Lihong V Wang and Hsin-I Wu. *Biomedical Optics: Principles and Imaging*. Wiley, Sept. 2012. ISBN: 978-0-470-17700-6 (cit. on p. 17).

[359] Pu Wang, Tyler M Berzin, Jeremy Romek Glissen Brown, Shishira Bharadwaj, Aymeric Becq, Xun Xiao, Peixi Liu, Liangping Li, Yan Song, Di Zhang, Yi Li, Guangre Xu, Mengtian Tu, and Xiaogang Liu. “Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study”. In: *Gut* 68.10 (2019), pp. 1813–1819. ISSN: 0017-5749. DOI: 10.1136/gutjnl-2018-317500 (cit. on p. 221).

[360] Yu Winston Wang, Nicholas P. Reder, Soyoung Kang, Adam K. Glaser, and Jonathan T.C. Liu. “Multiplexed Optical Imaging of Tumor-Directed Nanoparticles: A Review of Imaging Systems and Approaches”. In: *Nanotheranostics* 1 (2017), pp. 369–388. DOI: 10.7150/ntno.21136. URL: https://www.ntno.org/v01p0369.htm (cit. on p. 21).

[361] Joan Webster and Abdullah Alghamdi. “Use of plastic adhesive drapes during surgery for preventing surgical site infection”. In: *Cochrane Database Syst Rev* 2015.4 (Apr. 2015), p. CD006353. DOI: 10.1002/14651858.CD006353.pub4 (cit. on p. 157).

[362] Markus Weigand, Maximilian Dietrich, and Mathias Pletz. *Sepsis: Pathophysiologie, Diagnose und klinisches Management*. Vol. 1. Walter de Gruyter GmbH & Co KG, 2022. ISBN: 978-3-11-067336-4. DOI: 10.1515/9783110673395 (cit. on pp. 32, 34, 37–39, 41, 178).

[363] Bernard L Welch. “The generalization of STUDENT’Sproblem when several different population varlances are involved”. In: *Biometrika* 34.1-2 (1947), pp. 28–35 (cit. on pp. 192, 256).

[364] Manuel Wiesenfarth, Annika Reinke, Bennett A. Landman, Matthias Eisenmann, Laura Aguilera Saiz, M. Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. “Methods and open-source toolkit for analyzing and visualizing challenge results”. In: *Scientific Reports* 11.1 (Jan. 2021), p. 2369. ISSN: 2045-2322. DOI: 10.1038/s41598-021-82017-6. URL: https://www.nature.com/

articles/s41598-021-82017-6 (cit. on pp. 67, 74, 130, 136, 137, 161, 170, 247, 248, 253).

[365] Ross Wightman. *PyTorch Image Models*. 2019. DOI: 10.5281/zenodo.4414861 (cit. on pp. 187, 189).

[366] Sebastian J Wirkert, Hannes Kenngott, Benjamin Mayer, Patrick Mietkowski, Martin Wagner, Peter Sauer, Neil T Clancy, Daniel S Elson, and Lena Maier-Hein. "Robust near real-time estimation of physiological parameters from megapixel multispectral images with inverse Monte Carlo and random forest regression". In: *International journal of computer assisted radiology and surgery* 11 (2016), pp. 909–917. DOI: 10.1007/s11548-016-1376-5 (cit. on pp. 26, 27, 65, 67).

[367] Sebastian J. Wirkert, Anant S. Vemuri, Hannes G. Kenngott, Sara Moccia, Michael Götz, Benjamin F. B. Mayer, Klaus H. Maier-Hein, Daniel S. Elson, and Lena Maier-Hein. "Physiological Parameter Estimation from Multispectral Images Unleashed". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*. Ed. by Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne. Cham: Springer International Publishing, 2017, pp. 134–141. ISBN: 978-3-319-66179-7. DOI: 10.1007/978-3-319-66179-7_16 (cit. on pp. 3, 26, 27, 65, 67).

[368] Sebastian Josef Wirkert. "Multispectral image analysis in laparoscopy – A machine learning approach to live perfusion monitoring". PhD thesis. Karlsruhe Institute of Technology, Jan. 18, 2018 (cit. on pp. 26, 27).

[369] Misganaw Tadesse Woldemariam and Worku Jimma. "Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review". In: *BMJ Health & Care Informatics* 30.1 (2023). DOI: 10.1136/bmjhci-2022-100704 (cit. on p. 178).

[370] Andrew Wong, Erkin Otles, John P Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, et al. "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients". In: *JAMA internal medicine* 181.8 (2021), pp. 1065–1070. DOI: 10.1001/jamainternmed.2021.2626 (cit. on pp. 10, 178).

[371] Lianlian Wu, Jun Zhang, Wei Zhou, Ping An, Lei Shen, Jun Liu, Xiaoda Jiang, Xu Huang, Ganggang Mu, Xinyue Wan, Xiaoguang Lv, Juan Gao, Ning Cui, Shan Hu, Yiyun Chen, Xiao Hu, Jiangjie Li, Di Chen, Dexin Gong, Xinqi He, Qianshan Ding, Xiaoyun Zhu, Suqin Li, Xiao Wei, Xia Li, Xuemei Wang, Jie Zhou, Mengjiao Zhang, and Hong Gang Yu. "Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy". In: *Gut* 68.12 (2019), pp. 2161–2169. ISSN: 0017-5749. DOI: 10.1136/gutjnl-2018-317366 (cit. on p. 221).

[372] Pavel Yakubovskiy. *Segmentation Models Pytorch*. 2020. URL: https://github.com/qubvel/segmentation_models.pytorch (cit. on p. 126).

[373] Yuji Yamaguchi and Vincent J Hearing. "Melanocytes and their diseases". In: *Cold Spring Harb Perspect Med* 4.5 (May 2014). DOI: 10.1101/cshperspect.a017046 (cit. on p. 180).

[374] Wenjun Yan, Lu Huang, Liming Xia, Shengjia Gu, Fuhua Yan, Yuanyuan Wang, and Qian Tao. "MRI Manufacturer Shift and Adaptation: Increasing the Generalizability of Deep Learning Segmentation for MR Images Acquired with Different Scanners". In: *Radiol Artif Intell* 2.4 (July 2020), e190195. DOI: 10.1148/ryai.2020190195 (cit. on p. 80).

[375] Zhenyu Yang, Xiaoju Cui, and Zhe Song. "Predicting sepsis onset in ICU using machine learning models: a systematic review and meta-analysis". In: *BMC infectious diseases* 23.1 (2023), p. 635. DOI: 10.1186/s12879-023-08614-0 (cit. on pp. 178, 190).

[376] Ji Yi and Vadim Backman. "Imaging a full set of optical scattering properties of biological tissue by inverse spectroscopic optical coherence tomography". In: *Optics letters* 37.21 (2012), p. 4443. DOI: 10.1364/OL.37.004443 (cit. on p. 19).

[377] Jonghee Yoon, James Joseph, Dale J Waterhouse, A Siri Luthman, George S D Gordon, Massimiliano di Pietro, Wladyslaw Januszewicz, Rebecca C Fitzgerald, and Sarah E Bohndiek. "A clinically translatable hyperspectral endoscopy (HySE) system for imaging the gastrointestinal tract". In: *Nature Communications* 10.1 (Apr. 2019), p. 1902. DOI: 10.1038/s41467-019-09484-4 (cit. on p. 217).

[378] Alice C Yu, Bahram Mohajer, and John Eng. "External validation of deep learning algorithms for radiologic diagnosis: a systematic review". In: *Radiology: Artificial Intelligence* 4.3 (2022), e210064. DOI: 10.1148/ryai.210064 (cit. on p. 10).

[379] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 6022–6031. ISBN: 978-1-7281-4803-8 (cit. on pp. 154, 157).

[380] Anam Zahra, Rizwan Qureshi, Muhammad Sajjad, Ferhat Sadak, Mehmood Nawaz, Haris Ahmad Khan, and Muhammad Uzair. "Current advances in imaging spectroscopy and its state-of-the-art applications". In: *Expert Systems with Applications* 238 (2024), p. 122172. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.122172 (cit. on p. 3).

[381] Chenglong Zhang, Zhimin Zhang, Dexin Yu, Qiyuan Cheng, Shihao Shan, Mengjiao Li, Lichao Mou, Xiaoli Yang, and Xiaopeng Ma. "Unsupervised band selection of medical hyperspectral images guided by data gravitation and weak correlation". In: *Computer Methods and Programs in Biomedicine* 240 (2023), p. 107721. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2023.107721 (cit. on p. 218).

[382] Yating Zhang, Xiaoqian Wu, Li He, Chan Meng, Shunda Du, Jie Bao, and Yongchang Zheng. "Applications of hyperspectral imaging in the detection and diagnosis of solid tumors". In: *Translational Cancer Research* 9.2 (Feb. 2020). Publisher: AME Publishing Company. ISSN: 2219-6803, 2218-676X. DOI: 10.21037/tcr.2019.12.53. URL: https://tcr.amegroups.com/article/view/34678 (cit. on p. 148).

[383] Zhongheng Zhang, Lin Chen, Ping Xu, Qing Wang, Jianjun Zhang, Kun Chen, Casey M. Clements, Leo Anthony Celi, Vitaly Herasevich, and Yucai Hong. "Effectiveness of automated alerting system compared to usual care for the management of sepsis". In: *npj Digital Medicine* 5.1 (July 19, 2022), p. 101. ISSN: 2398-6352. DOI: 10.1038/s41746-022-00650-5 (cit. on p. 206).

[384] J. M. Zhao and L. H. Liu. "Radiative Transfer Equation and Solutions". In: *Handbook of Thermal Science and Engineering*. Ed. by Francis A. Kulacki. Cham: Springer International Publishing, 2017, pp. 1–46. ISBN: 978-3-319-32003-8. DOI: 10.1007/978-3-319-32003-8_56-1 (cit. on p. 26).

[385] Qingyu Zhao, Ehsan Adeli, and Kilian M. Pohl. "Training confounder-free deep learning models for medical applications". In: *Nature Communications* 11.1 (Nov. 2020), pp. 1–9. ISSN: 2041-1723. DOI: 10.1038/s41467-020-19784-9 (cit. on p. 109).

[386] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. "A survey of large language models". In: *arXiv preprint arXiv:2303.18223* (2023) (cit. on pp. 10, 43).

[387] Evgeny Zherebtsov, Viktor Dremin, Alexey Popov, Alexander Doronin, Daria Kurakina, Mikhail Kirillin, Igor Meglinski, and Alexander Bykov. "Hyperspectral imaging of human skin aided by artificial neural networks". In: *Biomedical optics express* 10.7 (2019), pp. 3545–3559. DOI: 10.1364/BOE.10.003545 (cit. on p. 218).

[388] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. "Random Erasing Data Augmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.07 (Apr. 2020), pp. 13001–13008. ISSN: 2374-3468 (cit. on pp. 154, 157).

[389] Zhi-Hua Zhou. *Machine learning*. Springer nature, 2021. DOI: 10.1007/978-981-15-1967-3 (cit. on pp. 44, 48, 50).