A Transparent View on Machine Learning for Holographic Cytology

Stefan J. Röhrl

Technische Universität München
TUM School of Computation, Information and Technology

A Transparent View on Machine Learning for Holographic Cytology

Stefan J. Röhrl

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Gordon Cheng

Prüfende der Dissertation:

1. Prof. Dr.-Ing. Klaus Diepold
2. Prof. Dr. Oliver Hayden

Die Dissertation wurde am 19.02.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 02.09.2024 angenommen.

Stefan J. Röhrl. *A Transparent View on Machine Learning for Holographic Cytology*. Dissertation, Technical University of Munich, Munich, Germany 2024.

Chair of Data Processing, Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany, https://www.ce.cit.tum.de/ldv.

Acknowledgment

Completing a dissertation within such a broad interdisciplinary framework is a journey that owes its success not to the efforts of a single individual but rather to the collective support and inspiration of many. Therefore, I would like to express my sincere gratitude to all those who have contributed to the completion of this thesis.

First and foremost, I would like to express my sincere appreciation to my supervisor, **Prof. Dr. Klaus Diepold**, whose unconditional acceptance and unwavering guidance have been substantial in shaping my academic journey. His role as an advisor, protector, and source of encouragement has been invaluable. In addition, I am deeply grateful to **Prof. Dr. Oliver Hayden**, whose visionary approach and warmth of spirit were indispensable in the conception and execution of this research project. My passion for the Chair of Data Processing was ignited and nurtured by **Dr. Michael Zwick**, whose mentorship and support have been crucial since my earliest days.

To my colleagues at LDV, I extend my deepest gratitude and enduring friendship. **Martin Knopp**'s selfless assistance and compassionate demeanor have been a cornerstone of support throughout. The collaborative efforts of **Manuel Lengl**, **Simon Schumann**, and **David Fresacher** have enriched my professional and personal experiences. Thank you for all the funny moments and adventures. I want to thank **Michael Dötzer**, **Philipp Paukner**, and **Martin Gottwald** for putting up with me for so long as old hands and for always being available for new projects and table soccer matches. I am also indebted to **Alice Hein** for her insightful discussions and candid feedback. The camaraderie shared with **Luca Sacchetto**, **Matthias Kissel**, **Sven Gronauer**, and **Marisa Ripoll** has made the academic environment more light-hearted. **Ricarda Baumhoer** and **Ernst Sellmann**'s unwavering dedication has been the foundation upon which our chair thrives. Beyond mere colleagues, you have all become valued friends.

I am equally indebted to my colleagues on the CellFace project, especially **Christian Klenk**, whose dedication and companionship inspire me. As a like-minded person, he has carried the CellFace project on his shoulders for me and for everyone else. **Dominik Heim**, **Ellen Emken**, and **Julia Sistermanns** have provided invaluable support, both professionally and personally, which I greatly appreciate.

I want to thank the research partners and experts who have provided consolidated support for this endeavor, especially **Dr. Katja Peschke**, **Prof. Dr. Maximilian Reichert**, **Sabine Farschtschi**, **Prof. Dr. Michael Pfaffl**, **Dr. Matthias Ugele**, **Prof. Dr. Wolfgang Utschick**, **Dr. Johanna Erber**, **Dr. Sebastian Rasch**, **Dr. Hedwig Irl**, **Dr. Martin Schlegel**, and **Dr. Gregor Weirich**.

Finally, I would like to thank my family, my parents, **Ingrid** and **Helmut Röhrl**, and my brother **Michael** for carrying me. Thank you for the security you have given me. Your belief in me and your constant reassurance have made this possible. Thank you also to my best friends **Martin Götz** and **Johannes Kolmer**, who often had to do without me during this intense time. Above all, I would like to express my deepest and most sincere thanks to my partner **Sabrina**. You are my light and my guide, whom I will follow everywhere.

As one chapter draws to a close, I eagerly look forward to embarking on new endeavors, strengthened by the support of those who have accompanied me on this journey.

Abstract

Personalized medicine is a shared aspiration of patients and healthcare professionals. The ability to quickly and accurately diagnose diseases preemptively, independent of restricting defaults, holds great promise for improving the well-being of individuals. Label-free holographic cytology is emerging as a platform technology that offers a way to such flexible point-of-care diagnostics. Its effectiveness as an imaging technique requires the integration of powerful but also comprehensible algorithms from the fields of computer vision and machine learning. However, the opacity of modern machine learning methods poses a dilemma for researchers and clinicians.

Using the promising application of holographic cytology as an example, this work examines how a successful translation of learning algorithms into everyday clinical practice can be achieved. Based on an extended *Technology Acceptance Model*, the work evaluates how machine learning methods can form a solid data processing pipeline for the analysis of holographic cell images. Particular emphasis is placed on the explainability of these methods and their impact on trustworthiness when interpreted by humans. To put the developed concepts into practice, this work presents a catalog of design rules for the user-friendly construction of AI-infused biomedical tools. These guidelines promote transparency in interdisciplinary research on problems in holographic cytology.

The results of this work advocate for an understandable and equitable exchange of information and findings across specialist disciplines. This collaborative approach is essential for the sustainable translation of machine learning into the healthcare system, ensuring seamless integration and widespread adoption.

Zusammenfassung

Personalisierte Medizin ist der Traum vieler Patienten und Ärzte. Unabhängig von vordefinierten Methoden zu sein und Krankheiten schnell und präventiv diagnostizieren zu können, birgt ein großes Versprechen für die Verbesserung des Wohlbefindens des Einzelnen. Die markierungsfreie holographische Zytologie bietet eine Technologieplattform, um eine solch flexible und unmittelbare Form der Diagnostik zu verwirklichen. Als bildgebendes Verfahren benötigt sie dazu jedoch die Hilfe von leistungsfähigen, aber auch verständlichen Algorithmen aus dem Bereich der Computer Vision und des maschinellen Lernens. Die Undurchsichtigkeit der modernen maschinellen Lernmethoden stellt die Forscher und Mediziner jedoch vor ein Dilemma.

Am Anwendungsfall der vielversprechenden holographischen Zytologie untersucht diese Arbeit, wie eine erfolgreiche Translation von Lernalgorithmen in den klinischen Alltag dennoch gelingen kann. Basierend auf einem erweiterten *Technologieakzeptanzmodell* evaluiert die Arbeit, wie maschinelle Lernverfahren eine robuste Datenverarbeitungspipeline für die Analyse holographischer Zellbilder bilden können. Besonderes Augenmerk wird dabei auf die Erklärbarkeit der verwendeten Methoden und den Einfluss der menschlichen Interpretation auf deren Vertrauenswürdigkeit gelegt. Um die erarbeiteten Konzepte in die Praxis umzusetzen, stellt diese Arbeit einen Katalog von Gestaltungsrichtlinien für den nutzerfreundlichen Aufbau von KI-gestützten biomedizinischen Programmen vor. Diese fördern die Transparenz in der interdisziplinären Forschung zu Problemen der holographischen Zytologie.

Die Ergebnisse dieser Arbeit unterstützen den nachvollziehbaren und gleichberechtigten Austausch von Informationen und Erkenntnissen über Fachdisziplinen hinweg. Dieser kollaborative Ansatz ist für die nachhaltige Translation von maschinellem Lernen in das Gesundheitssystem unerlässlich und gewährleistet eine nahtlose Integration und breite Akzeptanz.

Contents

Acronyms

AI	Artificial Intelligence.
CNN	Convolutional Neural Network.
DHM	Digital Holographic Microscopy.
LIME	Local Interpretable Model-agnostic Explanations.
QPI	Quantitative Phase Imaging.
RF	Random Forest.
SHAP	SHapely Additive exPlanations.
SOM	Self-Organizing Map.
SVM	Support Vector Machine.
UEM	Usability Evaluation Metric.
VAE	Variational Autoencoder.
XAI	eXplainable Artificial Intelligence.

Chapter 1

Introduction

> “*Remember to let her into your heart,*
> *Then you can start to make it better.*”
>
> Paul McCartney, *Hey Jude*, 1968

1.1 Motivation

The Medical Tricorder. In the ever-evolving world of technological advancement, the line between science fiction and reality continues to blur. Once relegated to the realm of imagination, concepts manifest as tangible tools that shape our daily lives. From self-driving cars [11] to the wonders of 3D-printed food [101], the boundaries of innovation are constantly being pushed by ingenious minds. Among these transformative inventions is the *medical Tricorder*, a futuristic device reminiscent of Captain Kirk’s voyages aboard the Starship Enterprise [345]. Far from mere fiction, however, similar ideas are poised to revolutionize healthcare as we know it.

The *Tricorder*, a handheld device no larger than a smartphone, promises to provide in-depth analyses of blood and tissue samples. The device scans entire human bodies within seconds, providing insight into health conditions and detecting anomalies with unprecedented efficiency. Like existing diagnostic methods, the *Tricorder* prioritizes minimal invasiveness, ensuring patient comfort while delivering accurate results across a spectrum of medical contexts. But the heart of its innovation lies not in **what** it does but **how** it does it. Powered by highly customizable hardware and sophisticated software, the *medical Tricorder* represents the pinnacle of integration and precision. By surpassing previous assumptions about health states and species, it points to a new era of diagnostic capability. However, achieving this vision requires not only technological prowess but also a convergence of disciplines, from biology to computer science.

In the field of blood and tissue analysis, emerging technologies such as Quantitative Phase Imaging (QPI) offer a glimpse into the future of the *medical Tricorder*. By exploiting the physical properties of cellular structures, QPI enables health assessments without the need for cumbersome sample preparation or labeling [145, 160, 254, 283]. Combined with a microfluidic sample presentation, it is possible to digitize tens of thousands of cells within minutes and even reuse the cells afterward [200, 254]. This measurement principle not only streamlines the diagnostic process but brings science closer to the seamless functionality of a *Tricorder*. However, the QPI technology – closer explained in section 2.1 – shifts the complexity of a formerly bio-engineering problem into the domain of computer vision and data science.

Figure 1.1: A *medical Tricorder* for advanced cellular analysis illustrated by *DALL-E*.

Consequently, novel software solutions are essential to effectively process data generated by QPI [145, 230, 251]. Unlike conventional cytology, there is no backup from reliant *molecular labeling* [35]. Ambitious researchers are exploring the integration of machine learning algorithms to emulate the functionality of the *Tricorder's* software. The efficacy of machine learning in cytology is undeniable and offers promising scenarios for the fusion of label-free QPI and self-evolving algorithms [230, 251]. Often hailed as Artificial Intelligence (AI), these programs excel at pattern recognition and correlation analysis beyond human capabilities [40, 145]. But therein lies a crucial dilemma: while machine learning models provide unparalleled performance, their output may lack the desired intelligibility in sensitive matters such as personal health, posing a challenge to researchers and medical practitioners alike [133, 134, 261, 328].

As we move into a future where technology and healthcare converge, bridging the gap between innovation and understanding is a paramount concern [69, 133, 279]. This work is dedicated to systematically exploring the barriers to the integration of AI into biomedical research. It aims to comprehensively assess the feasibility of analyzing cellular structures without labeling while also addressing the critical role of human factors in the acceptance of AI-generated results. The overarching goal is to ensure equal consideration of both technical and human-centered aspects, thus facilitating the seamless assimilation of AI into biomedical products.

Apropos assimilation: While the concept of the *medical Tricorder* may seem distant, this work hopefully brings research closer to the dream of efficient and personalized healthcare. A possible realization of an advanced cellular analysis *Tricorder*, which provides direct insight into a patient's cardiovascular risk [55, 348], can be seen in Figure 1.1. At least, this is how a generative AI would render it based on the description provided.

Let AI into Your Heart. The transformative impact of AI in healthcare promises improved workflow efficiency, accelerated drug development, and reduction of medical errors, creating opportunities for healthcare systems and insurance companies. For practitioners and researchers, AI promises to provide invaluable support through reliable diagnoses, precise image analysis, and conclusive treatment recommendations. Yet the actual beneficiaries of this technological advance should be patients, who will experience immediate improvements in well-being through rapid health tracking, fitness monitoring, and personalized treatments [330].

Despite these promises, the full integration of AI into the heart of healthcare faces a significant challenge. The barrier to widespread adoption lies in the complex nature of machine learning models [18, 199, 284]. As powerful but incomprehensible technology expands into the sensitive areas of health, critical factors such as verifiability and trust come to the forefront [133, 134, 242]. Mistrust develops as these models offer approximations rather than universal certainty about the features influencing machine decisions [279]. Most models learn by optimizing probabilities based on examples, striving to make accurate statements for the majority of cases. Especially in deep learning, one of the most promising approaches for image analysis, transparent decision-making remains challenging [43, 269, 279, 362].

In contrast, conventional products operate within known physical or chemical principles. For example, a 40x objective consistently magnifies cells to 40 times their actual size, regardless of their origin. Molecular-level *fluorescent labeling* reliably illuminates cell organelles when exposed to the appropriate wavelength [240, 283]. These processes are unaffected by factors such as ethnicity [61, 88, 154, 252] or unwanted background structures [271]. Past examples have shown that machine learning can be vulnerable to such unintended influences [36, 154]. Their opacity makes it difficult to anticipate these obstacles. Even techniques aimed at constructing eXplainable Artificial Intelligence (XAI) may fall short of conveying the models' inherent mechanisms while remaining comprehensible [52, 94, 186].

Experts predict that at least 20% of all clinical applications will incorporate machine learning within the next few years [40, 105]. However, the prevailing perception of machine learning as a black box is a significant limitation that hinders its adoption in clinical settings [7, 104, 111, 325]. Some argue that current approaches are not yet ready for clinical use [366]. Between 2015 and 2020, authorities approved in certain fields less than 3% of more than 450 AI-based products, indicating a significant gap between potential and regulatory acceptance [105, 222]. In addition, a survey of over 6,000 citizens from different countries found that 70% lack confidence in AI in healthcare [96].

To move closer to the dream of *Healthcare 5.0* or even a *medical Tricorder*, methods must become reason-based and interpretable [219, 255, 285, 346]. Recognizing this challenge, lawmakers are seeking to establish guidelines for the development of trustworthy machine learning methods. Individuals should have a right to be informed about the underlying logic of a system [106, 110] and receive explanations of decision-making processes [72, 128]. Despite legislative efforts, the practical implementation of this right poses challenges, as explanation does not guarantee understanding [40, 279].

Building trust is an incremental process [172, 205, 213]. Thus, this work begins by testing machine learning in an *in vitro* diagnostic research project [274] instead of a high-risk clinical use case. As explained in the subsequent section, the goal is to explore how the translation of AI into medicine can overcome the technical hurdles and acceptance issues. It takes time and transparency for people to let AI into their hearts, both physically and mentally.

1.2 Research Scope and Objectives

All You Need Is AI. The demand for highly personalized medicine underscores the need to harness the power of machine learning and novel technologies. However, in the face of existing challenges, it is crucial to find solutions that not only enhance the performance of these technologies but also inspire confidence among users. Research has shown that the efficacy of the technology alone will not ensure its success [36, 269, 336]. Therefore, Davis et al. [62] introduced a widely used model to explain the acceptance of new technologies. Introduced in 1989, the *Technology Acceptance Model* posits that **usability** is critical in addition to technical **usefulness**. However, this model does not fully address the unique challenge posed by technologies that hide the logic behind their results from human observers [337]. Unlike traditional technologies, where a mechanic can understand the underlying processes in a vehicle even if the driver does not, machine learning presents a different scenario. In response, this work expands the existing model to include **trustworthiness**, which is recognized by experts [6, 249, 340] and legislators [77, 78, 128] as essential for the further development and adaptation of AI-infused products.

Recognizing the unique regulatory demands of the biomedical domain, this work embarks on a multidisciplinary research journey involving multiple stakeholders and technical challenges. The setting in a biomedical development project allows for a thorough exploration of user characteristics, algorithmic requirements, and their interaction under controlled conditions. (Note that Chapter 2 provides further details about the environmental and technological circumstances.) This work focuses not only on proof of concept or achieving the highest prediction accuracy but also on interpretation phases with human experts. Questions arise regarding how to explain machine learning results to biomedical audiences, what influences their interpretation, and how to facilitate the flow of information in scenarios involving black-box models.

These questions have a long history. Already 30 years ago, the bioinformatician Dean Sittig formulated fundamental requirements for successful AI integration into the biomedical context [305]. These include the establishment of "a unified controlled medical vocabulary", the creation of a "uniform, intuitive anticipating user interface", techniques to seamlessly integrate "new information management technologies into the infrastructure of organizations so that they can be used at the bedside or at the research bench", and the development of a "comprehensive *clinical decision support system*". However, the progress of AI technology in recent years has posed new challenges. The less interpretable these algorithms become, the more important it is to prioritize transparent and interdisciplinary communication in future technologies and research workflows. Collaboration among all stakeholders is essential to prevent the "ill-informed use of artificial intelligence" that has been identified as "driving a deluge of unreliable or useless research" [14]. Without concerted efforts, successful technology translation to the point-of-care will be infeasible [154].

Research Questions. The preceding discussion highlights numerous factors that might influence the successful adoption of a new technology. Therefore, the central research question of this work is:

What criteria determine the successful translation of an AI-driven data processing pipeline into holographic cytology? From this central question, three sub-questions emerge that align with the three translation criteria derived from the *Technology Acceptance Model*:

RQ 1. *Can existing machine learning methods be adapted to ensure stable and transparent processing of holographic cell images?*

This question involves investigating the applicability of existing models and strategies to the analysis of cells in the QPI domain. The aim is to identify the necessary design and technical adjustments to achieve reliable processing of cell images, resulting in effective segmentation and classification performance. This directly relates to the approach's **usefulness**. Objectives are the construction and hyperparameter tuning of a versatile data processing pipeline, as well as transfer learning and validation of machine learning models. Emphasis will be placed on models that prioritize transparency without compromising performance in overcoming technical challenges. The research includes a comparative analysis between classical image processing techniques and state-of-the-art machine learning methods, with a particular focus on whether there is an inverse relationship between interpretability and model performance [125, 133, 279, 355].

RQ 2. *What are the essential criteria for establishing trustworthiness in machine learning pipelines designed for holographic cytology?*

This question directly addresses the interpretability of the models used as a critical factor in assessing the **trustworthiness** of machine learning approaches. Trust can also be instilled through consistent demonstration of exceptional performance or a user-friendly interface [89, 213, 360]. However, the current model landscape often requires additional explanations of algorithmic behavior, namely XAI. In investigating this issue, special attention will be given to the interdisciplinary target audience, considering their diverse professional backgrounds and varying understanding of roles [53, 56, 246, 328]. A literature review is performed to compile criteria that determine the interpretability of a model and identify latent factors that influence interpretability in different situations. Finally, a user study of an illustrative XAI dashboard will assess how individuals perceive such explanations and how these explanations affect the perceived trustworthiness of the algorithms.

RQ 3. *What design rules are suitable for improving the usability of AI-driven user interfaces for holographic cytology?*

It is indispensable to establish practical guidelines to translate acquired knowledge into useful and reliable machine learning applications [74, 246]. Therefore, this research question focuses on understanding the dynamics of AI-based user interfaces and their impact on perceived **usability**. Since the biomedical target audience interacts primarily with the user interface, it must effectively communicate all of the aforementioned aspects and criteria. The success of AI translation depends on these last critical interactions. Novel rules for designing such tools will be evaluated through a prototype implementation of cytology analysis software that connects users with AI through an active learning [5, C, 132] approach. These domain-specific design rules will be compared to existing guidelines [6, 233, 299] and validated [118] through user testing throughout the development process.

1.3 Structure of This Work

After this introduction, Chapter 2 provides a detailed explanation of the previously outlined research framework and clarifies the contextual conditions of this work. It introduces the quantitative phase technology and its characteristics, organizing it within a cyclical research workflow. An overview section highlights key technological steps such as sample preparation and the AI-powered data processing pipeline. This is followed by a thorough explanation of the data processing steps, providing a comprehensive understanding of the necessary core tasks of machine learning for the examined biomedical use cases. Once the research context is established, Chapter 3 unfolds the theoretical background. It presents state-of-the-art machine learning models that have been successful in cell analysis. Subsequently, it explores the background of research on the explainability and interpretability of machine learning methods. This chapter also examines different stakeholders and evaluation procedures for interpretability and introduces the nuances of user interface design, especially when incorporating AI. The theory section concludes with an exploration of user interface evaluation and usability evaluation methods. Chapter 4 articulates the contributions within the three translation criteria by briefly summarizing the findings of the core publications constituting this work. Chapter 5 discusses the achieved results, places them in their scientific context, and gives an outlook on future scenarios for the integration of machine learning in personalized healthcare.

Note that this work is a cumulative dissertation. Therefore, the original core publications forming this work are printed in Appendix I - VI. Roman numerals indicate references to core publications [I]. In addition, related publications that are directly affiliated with this work but are not considered core publications are indicated by capital letters [A]. Standard references from literature are presented as usual in Arabic numerals [1].

Author Contributions. *At the time of publication, the authors of the core publications have taken a slightly different view of equal contribution to the papers than the promotion regulation. The main **scientific contribution** (>50%) of all core publications listed here lies with the **author of this dissertation**. Non-scientific contributions were nevertheless acknowledged and marked by mutual agreement by the research team in the individual papers. The author of this dissertation always has the full consent of all co-authors to use all the work presented here as a complete core publication in his dissertation. The supervisor confirms this by giving his consent to this cumulative dissertation. For a detailed explanation of the individual contributions, please consult the footnotes of each core publication.*

Tools. *Various common tools were used to improve the readability and language of this document, like Grammarly, DeepL, and different large language models. The actual content of this document was written entirely by the author of this dissertation and is his responsibility.*

Chapter 2

Overview and Research Environment

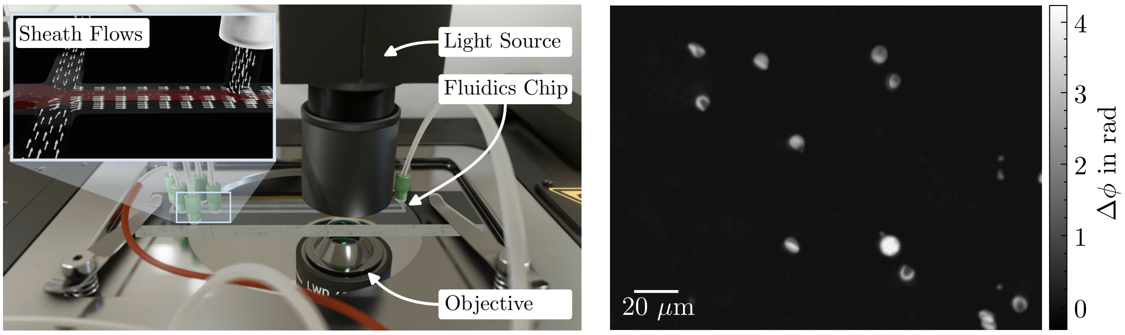
Given the vast scope of biomedical research, this work focuses on one promising platform technology, recognizing that making sweeping statements in this area must be treated with caution. The success and acceptance of essential computer vision techniques will play a key role in the successful translation of this technology into everyday clinical practice. Therefore, it is worthwhile to examine these techniques in detail. This chapter aims to provide an in-depth exploration by introducing the employed imaging method and comparing its advantages and disadvantages with conventional approaches. Subsequently, an overview of the interfaces and interrelationships within a cyclical interdisciplinary research workflow is presented. The core of this work unfolds with the exposition of the data processing components and the functions of the incorporated machine learning methods.

2.1 Holographic Cytology as New Platform Technology

Imaging and Microfluidics Platform. Digital Holographic Microscopy (DHM) has evolved significantly since the 1980s and is now widely used in industry and laboratories [331]. This technology is precious in cell analysis because most cellular structures appear transparent under conventional bright-field microscopes. The resulting quantitative phase images provide enhanced contrast without the need for sample staining, offering advantages such as reduced sample preparation effort and extended detection range. This approach requires less prior knowledge of the sample and its components compared to labeling procedures [57, 287]. While *hematology analyzers* automate *complete blood counts*, modern instruments face the challenge of providing deeper analysis, such as at the morphological level, in a short time and with high statistical power. Similarly, state-of-the-art *fluorescence microscopy* is precise but limited in spatial resolution and availability of appropriate fluorescent antibodies for specific cases [35]. The central hardware technology in this work, also known as QPI, allows the precise and simultaneous assessment of relevant biological and physical characteristics of the cell and its interior. This includes both phenotypic effects (size and optical volume) and temporal effects (drug response, granularity changes, aggregation) [230, 254].

A microscope based on QPI utilizes interference principles to measure both light transmission and phase shift ($\Delta\phi$). That allows for the inference of optical density in cellular structures without staining or *molecular labeling*, thus avoiding cell alteration or damage. This project uses an *off-axis diffraction* phase microscope from *Ovizio Imaging Systems*[†]

[†] https://ovizio.com



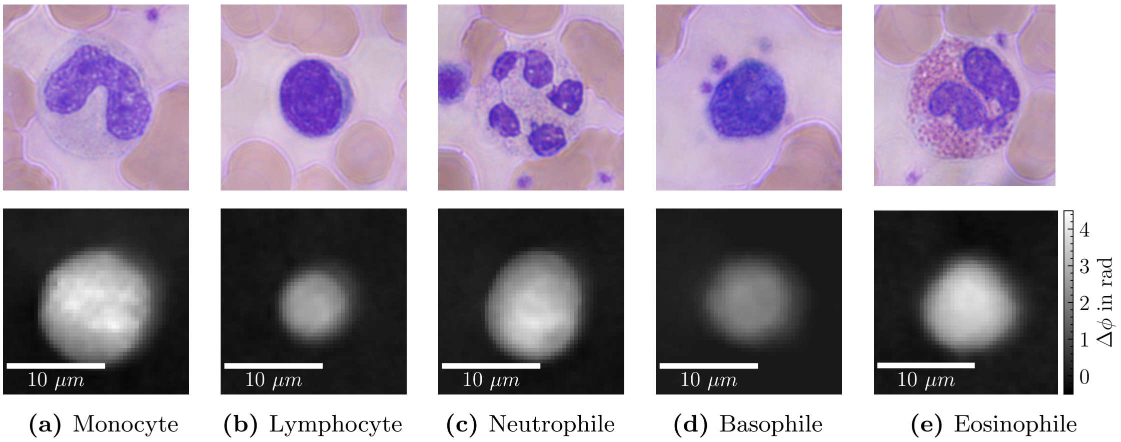
(a) DHM Microscope with Microfluidics **(b)** Raw Quantitative Phase Image

Figure 2.1: Digital Holographic Microscopy **(a)** allows the inference of optical density in cellular structures without staining. Image **(b)** displays the captured quantitative phase shift ($\Delta\phi$) at $\lambda = 528$nm as grayscale values.

combined with a microfluidics chip for label-free imaging of unprocessed cells in suspension under near *in vivo* conditions. The light source is a 528nm *Super-LED Köhler illumination*, directed onto a 50μm × 500μm *polymethylmethacrylate* microfluidics channel that uses sheath flows to focus the sample stream. Figure 2.1a shows a schematic of the employed setup and its components. As the light beam traverses the sample, the cell membrane, plasma, and nucleus cause a difference in the propagation time of the electromagnetic light wave due to their distinct optical densities. The resulting interference patterns (holograms) get projected onto a camera sensor, capturing 105 images per second. Further details of the optical setup can be found in [71, E, 333]. From here, physical [288] or AI-based [45, 273, 343] algorithms are able to reconstruct the values for amplitude and phase of the incident light and, thus, the nature of the cells from the recorded hologram images. All contained core publications employ phase images, as displayed in Figure 2.1b, as they optimally convey the internal structure and morphology of the observed cells.

To examine the sample as gently as possible and under near *in vivo* conditions [257], a microfluidic chip is used in the device, which brings the cells into the focal plane of the optical setup through *hydrodynamic* and *viscoelastic focusing* [9, 103, 109, D]. This ensures a significantly higher throughput than with microscope slides or wellplates, which in turn increases the statistical significance of the method. A comparatively low flow rate also minimizes the physical stress caused by shear forces [335]. The cells can even be used for further experiments or temporal monitoring of living cells [60, 158, 215, 216, 260]. As a result, the technique holds great promise for research, diagnosis, and treatment in a variety of biomedical applications [287]. However, the method cannot guarantee optimal focusing of all cells and requires downstream quality assurance mechanisms [20, 109].

Quantitative Phase Images - An Uncharted Territory. Modern blood analyzers are highly integrated machines that can distinguish and count cells using *immunohistochemical* or *fluorescent staining* [35, 210]. If a more detailed analysis is required that deviates from standardized procedures or if morphological changes in the cells are involved, the gold standard for the diagnosis of hematological diseases is the *Giemsa-stained blood smear* [16, 112, 254, 258]. Therefore, all current approaches require complex (albeit partially automated) sample preparation and staining reagents. For specific issues, this can also mean a high manual effort, a long processing time, and expensive use of specially trained personnel to examine just a few hundred cells, not to mention the associated inter-observer variations [35, 84, 162, 264].

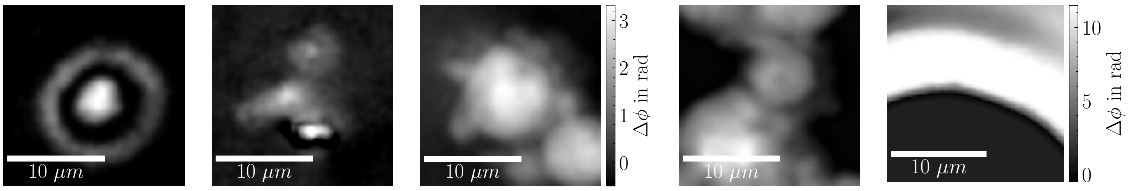


(a) Monocyte **(b)** Lymphocyte **(c)** Neutrophile **(d)** Basophile **(e)** Eosinophile

Figure 2.2: The five different subtypes of leukocytes. The upper row shows the cells on substrate with *Giemsa staining* under a conventional bright-field microscope. The bottom row contains the corresponding phase images in suspension using $\lambda = 528$nm illumination.

As stated before, QPI can circumvent these drawbacks by working label-free and, in combination with a microfluidics chip, capable of digitizing millions of cells at a fraction of cost and time. However, the matter is more complex than it may seem, as the phase images only show the phase shift ($\Delta\phi$), coded here in grayscale, as shown in the bottom row of Figure 2.2. There is no clear color assignment of the individual cell components. Dark values in phase images represent regions with low optical density; light values represent structures with high optical density, such as the cell nucleus. As a result, these pseudo-3D representations bear little visual resemblance to established images of tissue or blood cells. Jo et al. [145] summarize this effect of QPI on the biomedical field as "fast acquisition in the cost of low chemical specificity". Pathologists and laboratory personnel are trained on the purple-colored cell images or usually know the *clusters of differentiation* by heart during *immunophenotyping* [370]. However, they are entirely unfamiliar with grayscale phase images, making it difficult, if not impossible, for the human eye to distinguish the cell types [V]. To illustrate this challenge, Figure 2.2 compares the five types of leukocytes once using the established staining method on a glass slide in the upper row and once in suspension in the DHM in the bottom row.

Another challenge in imaging cells in the microfluidic channel is the blurred imaging of some specimens if the physical principles fail to optimally place them in the focal range of the microscope. Figure 2.3a shows an image of such a defocused cell, which quality assurance mechanisms must later filter out. Despite the minimal sample preparation and stress on the cells, fragments and other debris can always make it under the lens (2.3b). If the dilution is insufficient or disease-related effects occur in blood, cells can also form aggregates (2.3c) or clumps (2.3d), which makes their precise classification more difficult. With inexperienced handling or due to aging effects, air bubbles may even form in the system (2.3e). It is, therefore, clear that with this technology, many tasks previously assigned to pre-analytics, a physical principle, or humans now fall into the area of computer vision. The high number of cells makes purely manual quality assurance virtually impossible. The raw appearance of the cells takes trained biomedical researchers out of their comfort zone so that they appeal to the superhuman powers of machine-based object recognition and classification [145].



(a) Defocused **(b)** Debris **(c)** Aggregate **(d)** Cell Clump **(e)** Air Bubble

Figure 2.3: Potentially unwanted phenomena due to inadequate sample preparation, suboptimal focusing, inexperienced handling, or biological effects.

Digital Holographic Microscopy in Biomedical Applications. Various research groups worldwide have devoted years to addressing the diverse tasks and opportunities presented by QPI technology in biomedical applications. Researchers, clinicians, and engineers recognize the substantial advantage of label-free work in its high degree of flexibility [145, 254]. Eliminating sample preparation brings devices closer to the point-of-care without major time delays conceivable. The integration of QPI provides the opportunity to tailor diagnoses and treatments to individual patients with unprecedented speed and precision, marking a significant milestone in the monitoring of cardiovascular diseases, a leading cause of mortality worldwide [55, 86, 348]. This technology is also critical in the diagnosis of febrile patients, where timely confirmation of the underlying causes is paramount to initiating prompt treatment [15]. The potential for error in this process is not negligible and can have far-reaching consequences [183, 203]. In research, pharmaceutical facilities can study individual cells' behavior more comprehensively, enhancing the understanding of drug discovery and cytotoxic treatment [230, 287]. The technology applies to clinical samples and heterogeneous cell populations, allowing accurate tracking of various cellular events [230, 254]. It exhibits indifference to arbitrary cell types such as erythrocytes [157, 216, 332, 365], thrombocytes [156, D, E], leukocytes [73, 247, 333], tissues [259, 275], neurons [149, 204, 215] or sperm cells [34, 60, 116], making it capable of swiftly targeting a wide variety of diseases, including malaria [100, 159, 332], leukemia [247, I, 333], diabetes [182], Covid-19 [108, D, 235], carcinoma [23, 155, 237, G, 278], and many more [145, 230, 254].

Experts identify significant potential, especially in combining QPI and machine learning. New approaches reliably assume tasks related to segmenting and classifying cells and their internal structures. Nguyen et al. [230], Jo et al. [145], and Park et al. [254] provide an excellent overview of modern developments and new applications facilitated or made possible by QPI. Machine learning methods can enhance QPI techniques, such as tomographic reconstruction [60, 114, 211, 263, 359] or image enhancement [164, 351]. Generative approaches [273, 326, 327] can also address the issue of missing coloration, providing biologists with a familiar visual representation. Researchers "envision that the synergistic combination between QPI and AI could have far-reaching applications in biomedicine" [145]. This advantage originates from the versatility of data-driven approaches compared to techniques relying on direct *molecular labeling*. However, it is crucial to consider the potential risks associated with opaque machine learning. Consistently addressing this responsibility in the design of applications is essential, as "machine learning is poised to play an ever-increasing role in both the generation and interpretation of QPI data, and has already touched upon nearly every major application of QPI" [230].

The QPI microscopy presented here, in conjunction with the microfluidic setup, constitutes the hardware platform capable of revolutionizing numerous areas of cytology. It may

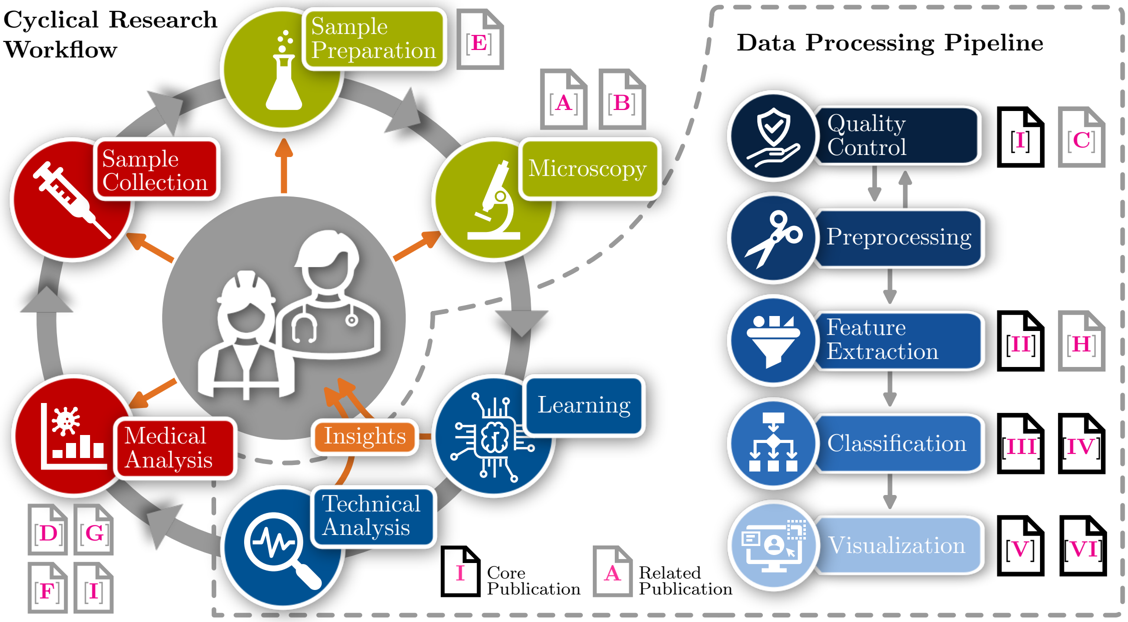


Figure 2.4: Integration of the data processing pipeline in the cyclical research workflow. Transparency in data processing provides valuable insights for all disciplines.

not work flawlessly as a *medical Tricorder* introduced in Chapter 1, but their label-free measurement principles are closely related. The hardware's role is to minimize interference in presenting biology, allowing the downstream software to utilize as much capacity as possible for detecting robust and generalized features in cells and cellular events. If humans can also comprehend these identified features, it opens ways for gaining new insights into some of the smallest building blocks of life.

2.2 Interdisciplinary Research Workflow

Interdisciplinary collaboration is essential to achieve a transformative revolution and ensure a sustainable increase in knowledge. Biologists do not need to acquire expert knowledge in AI, and vice versa. Instead, establishing transparent interfaces is essential to identify correlations and errors, thereby facilitating the generation of novel questions and discoveries. There needs to be more than a one-way service-provider relationship between disciplines; the active engagement of all stakeholders is crucial [40, 328, 329]. The current introductory phase of the presented platform technology favors a cyclical research workflow as drawn in Figure 2.4. Although it does not reflect a clinical workflow typical of established products, there are similarities, such as study design, standardized sampling techniques, and reference methods. Each stage in this iterative research process can solve problems and raise new questions. Given the interdependence of all disciplines, the seamless propagation of errors, effects, and conclusions, as well as the transparent presentation of relationships, is paramount. The following sections briefly highlight the process steps and provide information on their function and the stakeholders involved. Here, blood analysis for leukemia characteristics serves as an example application. The aim is to differentiate between the various types of leukocytes, as displayed in Figure 2.2, determine their frequency in the blood, and recognize and report morphological changes [121, II, VI, 333].

Sample Collection. Medical professionals carry out sample collection almost exclusively, performing it directly on the patient or healthy donor. The sample material can be tissue [122, 304], liquor [350], effusions [220], or blood. Blood is appropriate for most analyses, as its composition, the appearance of cells, and their interactions allow many statements about health [137]. That applies also to the investigation of leukemia. The specialists take venous blood with a standard cannula and fix it with an anticoagulant [E], which is usually already contained in the blood tube. Subsequently, a laboratory receives these blood tubes, where technical staff prepare them for further measurement, depending on the application.

Sample Preparation. Sample preparation is a delicate and uncertain step [16, 84, 162, 201, 283]. It involves bioengineers and laboratory technicians who aim to extract and stabilize as many raw biological effects and structures as possible for a reliant measurement. This inherently time-consuming and expensive pre-analytical stage necessitates dedicated laboratory resources. Hence, in the sense of portable *medical Tricorders*, it is unsurprising that researchers and clinicians alike are keen to simplify and speed up this step as much as possible. In the future, inexpensive software is expected to replace much of the manual, physical, and chemical procedures used to date. In the case of leukemia analysis, lab technicians focus on leukocyte isolation by lysing erythrocytes [339], followed by centrifugation. For QPI technology, the last preparation step is the dilution of the blood with a polymer solution and injection into the microscopy setup. The translational goal is to directly detect leukocytes in diluted whole blood and neglect lysis and other pre-analytics, creating a delay-free and highly integrable solution. This minimal sample preparation scenario also opens the door to digitizing extremely transient phenomena in the sample material, such as *microthrombotic events*. These aggregates, composed of leukocytes and thrombocytes, disintegrate within a few minutes [E, F], so there is currently no certified test for their routine diagnosis [107, D].

Microscopy. The diluted solution is now pumped into the microfluidic chip in the microscopy step. The chip uses *hydrodynamic* and *viscoelastic forces* [9, 103, 109, E] to focus the cells in as uniform a plane as possible, which the microscope can project sharply on the camera sensor. This process demands a high concentration of technicians, who ensure precise focusing, protection from excessive shear forces, maintenance of a noise-free background, homogeneous lighting, and the exclusion of air bubbles. (Help to partly automate this process could come from deep learning [136] or reinforcement learning [217] as well, but it may be far from immediate implementation.) By using a *self-referencing off-axis diffraction* phase microscopy setup, the optical path is robust against thermal and physical influences [71]. This step is uniform across cell types, with lower magnification objectives recommended only for larger tissue cells. The light rays that have passed through the leukocytes are captured by a 40x objective and brought to interference via the optical system. Special algorithms can reconstruct the amplitude and phase information from the recorded holograms [45, 273, 288, 343]. These form the basis for all further image processing steps.

Learning. The most computation-intensive part of the research process involves data scientists and machine learning experts. All incoming measurements, which may consist of up to 10,000 images, are checked for quality and converted into a standardized format [54, 285, 298, 318, 324]. Depending on the application, developers put together a pipeline that can analyze images in a targeted and robust manner [III]. In the leukemia example, segmentation algorithms need to reliably detect individual cells and separate them into image patches. The cells are then assessed for their characteristics in a feature extractor [119, II, H], raising the description of the cell to a higher level of abstraction. This step must capture changes in the appearance of the cell and still allow it to be clearly assigned to a specific leukocyte class. For these tasks, data scientists have a whole arsenal of classic feature extractors as well as machine learning methods at their disposal (see Section 3.2). The construction of a flexible and autonomous pipeline capable of reacting to new questions includes training and validation for each cell processing step. However, these data-driven approaches, now involving billions of parameters, are no longer deterministic and lack accessibility for human understanding [97, 347, 362]. To achieve the transparency of the workflow mentioned above, data scientists must solve a computer vision problem and design methods that are comprehensible in all involved disciplines.

Technical Analysis. The issue of transparency and communication of the results is part of the technical analysis, usually including a statistical evaluation and various visualizations. The highest priority is explaining the findings to provide interdisciplinary insights and increase knowledge (see Section 3.3). These are, among others, the comparison with a reference measurement [D], a list of the most significant features [29, 196], or the quantification of the uncertainty contained in the algorithmic analyses [C, 139]. The leukemia use case involves indicating the relative frequency of cell classes and visually explaining the cell components responsible for individual classifications. A representation of the shift in the morphological characteristics of the specimens in a typical scatterplot can be highly informative for medical professionals [V, 332, 333]. Adequate communication and consideration of the professional backgrounds of all those involved is crucial (see section 3.3.3). Data scientists and developers must ensure that this is reflected in the design of user interfaces, as the systems should also be operable by non-technical users later (see section 3.4). Generally, this process step contributes to an interpretation of the data and results from a technical point of view. It is an offer to all human researchers and is intended to help them detect sources of error, explore data, make well-founded statements, and uncover interesting new research aspects.

Medical Analysis. The research workflow comes full circle with the offer of the technical interface towards the *Medical Analysis*. Results are classified based on principles familiar to the biomedical domain, establishing a connection to the clinical picture or a broader study. The interpretation by doctors and biologists can then have far-reaching consequences [36, 183, 203, 352]. For a leukemia patient, this can act as a guiding and, thus, life-prolonging early diagnosis. The reliable and rapid evaluation can be a decision-making aid for the doctor [154, 318]. With this information, a research project can test hypotheses, set the course for further sampling, influence the planning of new experiments, trigger additional questions, or identify different application areas. Transparent evaluations are crucial for building trust [40, 105, 110, 133, 186, 242], emphasizing the duty of the data science and machine learning community to ensure explainability and detect or exclude unwanted behavior [36, 183, 203, 352], especially as tasks shift from hardware to software.

2.3 Data Processing Pipeline

As with any other field, the integration of AI-supported data processing continues to expand into the realm of biomedical products. This is particularly necessary in order to handle the increasing complexity shift from hardware to software, which facilitates the realization of many new applications. However, this transition brings new technical challenges and increases the responsibilities of developers. Consequently, there is a need for a paradigm shift that emphasizes not only predictive technical performance but also descriptive meaning [69, 133, 223, 242, 325]. Section 3.3 gives further details on this aspect. To better understand the pipeline steps drawn in blue in Figure 2.4, the following sections briefly explain the involved tasks.

Quality Control. In the context of biological samples, quality assurance steps that recognize the inherent imperfections of pre-analytical processes are essential. Maintaining a consistent standard is critical for reliable data processing [285, 306, 318]. Statistical evaluation of all images from the entire measurement helps identify significant deviations and provides immediate feedback to the laboratory staff [298]. A concept-based analysis should identify systematic errors such as background structures, microfluidic channel boundary effects, air bubbles, or blank images [I]. Incidents in this area are immediately reported to the measurement staff, facilitating immediate repetition of experiments or enabling predictive maintenance. These quality control measures remain primarily consistent regardless of the specific application, as in the example of the analysis of leukemia samples. Once systematic errors have been eliminated, the next step is to preprocess the images.

Preprocessing. This step is closely linked to the previous one and aims to standardize the digitized cells and prepare them for processing by machine learning methods [54, 285, 298]. Subtraction of the static background frees the images from residual confounding factors. All data representations, including hologram, phase, and amplitude, are stored in a uniform container format [324] along with their corresponding statistical values [298]. It is then essential to identify image regions containing relevant objects and structures, a process known as segmentation. The associated algorithms form two classes: *instance segmentation* [115], which involves identifying and distinguishing individual objects within an image, and *semantic segmentation* [179], which consists of grouping areas of equal meaning under a shared label. Solving this task is rarely straightforward [304, 309], especially in scenarios involving complex object arrangements such as overlapping, shadows, blur, insufficient resolution, or lack of texture. Hence, in microscopic cytology and potentially in the broader field of computer vision, this is one of the most common challenges and must be addressed before further analysis of the objects [76, 187]. In the case of leukocyte detection, simple *Threshold Segmentation* is often sufficient to separate cells and filter out noise, debris, and small cell fragments. However, more flexible methods like *Active Contours* [41, 93] or neural networks [123, 276, 358] can also be used to perform this task.

Feature Extraction. Feature extraction aims at a robust and expressive description of cells. After segmentation, cells are available as individual image patches, ready for analysis using various techniques. Hand-crafted morphological features [68, 121, 188, 226, 229, 235, 237, 245, 247, 275, 332] and well-established image transformations [82, 270] describe cells deterministically, but they may be insufficient to capture the whole nature of these organisms [145, 331]. These features are basically mathematical operations applied to a cell‘s contour and internal pixels. This involves translating known optical representations of cells from biological textbooks [13] into mathematical formulations to concisely characterize cells by numerical values. The cell contours [319] play a key role in assessing cell circumference and roundness, providing insight into cell elasticity [253, 335] and can reveal the effects of excessive shear forces (*Quality Control*) [D, 253, 335]. In addition, the pixels belonging to the cell’s interior contribute to the calculation of optical height, optical volume, estimation of dry mass, and approximation of cell homogeneity

[117, 121, 216, 260, 333]. These parameters are closely related to conventional descriptions derived from bright-field microscopy [127, 297]. Recently, however, Convolutional Neural Networks (CNNs) have shown superior performance and generalization capabilities for optimal adaptation to specific queries [4, 23, 46, III, 148, E, 243, IV]. Unfortunately, this advantage comes at the cost of reduced comprehensibility in the decision-making process. The development of automated feature extraction faces challenges posed by different object groups (leukocytes, erythrocytes, tissue cells, and parasites), substantial variability in object size, and the nature of the research objectives (e.g., malaria or leukemia detection). The primary goal is to identify features that provide generalization and robustness while preserving interpretability.

Classification. In the classification phase, the decision depends on the features or, in the case of a neural network, directly on the output of the low-level filter layers. The leukocytes under consideration must be assigned to one of the five subgroups of leukocytes typically found in human blood (compare Figure 2.2). Algorithmically transparent methods, such as Random Forests (RFs) [29] and Support Vector Machines (SVMs) [58], use decision trees or a projection of the cell feature space to classify the cells. Opaque deep learning models with many parameters can achieve equivalent or excelling performance [105, 190, 242] by propagating classification errors back through the layers during training [181, 311]. Supervised learning methods, including SVMs or *Linear Discriminant Analysis* [206], can be used as well as unsupervised approaches such as clustering or Self-Organizing Maps (SOMs) to group inherent properties of cell classes [H]. Alternatively, biologists may opt for manual gating [III, 312, 333], in which the computed features are plotted in different dimensions, and the resulting point clouds, called populations, are assigned to a cell class using hand-drawn gates. The variety of approaches to this step ranges from more or less autonomous to traceable. There are debates about whether non-traceable results should be allowed in the workflow or whether transparent approaches should be preferred [18, 187, 279]. There is also an ongoing discussion about the potential trade-off between classification performance and interpretability [125, 133, 223, 242, 279, 355].

Visualization. The role of visualization is to elevate all of the preceding processing steps to an appropriate level of abstraction and to ensure that the resulting metrics, parameters, and results are understandable to an interdisciplinary audience [56, 168, 357, 362, 364]. Depending on the application and methodology, various visual representations such as tables, box plots, scatter plots, and confusion matrices find their benefit. However, more detailed means of inspection are required to provide insight into the decision-making process throughout the pipeline. Therefore, additional plots include segmentation masks, 2D feature embeddings, confidence plots, and visual pattern explanations [IV]. In the context of leukemia analysis, this stage contains the statistical evaluation of cell frequency. Techniques such as Local Interpretable Model-agnostic Explanations (LIME) [271] shed light on the specific substructures that support or contradict the classification of a particular cell. Global visualization of morphological changes in cell populations compared to healthy donors provides meaningful details for subsequent diagnostic and treatment steps [V].

Chapter 3

Background and Related Work

The previous chapter has outlined the numerous tasks that need to be mastered and has placed these in the overarching research process. Label-free QPI microscopy opens up a world of cytology close to *in vivo* conditions, potentially bringing flexibility to the monitoring of cell kinetics [216, 260, 268, 365] and intercellular events [D, 235] into routine clinical diagnostics. However, it is also clear that the pressure on software developers and their methods is increasing due to the necessary restriction of sample preparation. Machine learning methods are intended to help realize a data-driven processing pipeline. Still, according to scientists and legislators, this poses a high risk that the algorithms will behave differently than intended [36, 111, 154, 215, 225] or that people will receive correct information but not gain any knowledge [14]. This chapter revisits the *Technology Acceptance Model*, which serves as the scientific framework for the acceptance and translation of the presented ideas. Based on this, the three translation criteria form the structure for the theoretical background and related work. The first section introduces previous approaches to solving the computer vision problem of cell analysis. That follows methods that should offer a way out of the black-box dilemma and the state of research on measuring explainability and interpretability. Putting the acquired knowledge into practice, the chapter ends with a summary of previous findings on the design and evaluation of user interfaces. Only these can ensure adequate communication between researchers and the algorithms.

3.1 Technology Acceptance Model

The translation of a new method into practice, which has already proven viability and safety, depends on its acceptance by the stakeholders [242, 325, 329, 330, 364, 366]. Therefore, this work follows the *Technology Acceptance Model* by Fred Davis 1989 [62], which outlines the central role of **usability** and **usefulness** in acceptance. In the area of machine learning methods, **trustworthiness** emerges as an additional crucial aspect. In the context of the QPI technology presented in this work, successful translation depends on meeting these key criteria in conjunction with machine learning methods. The criteria of usability, usefulness, and trustworthiness, as posited by Panagoulias et al. [249], stand out as paramount to the likelihood of method adoption. The following list briefly introduces these three criteria.

- **Usefulness:** Davis defines usefulness as “the degree to which a person believes that using a particular system would enhance his or her job performance.” Researchers often express concerns with numerous systems that are highly useful but have poor usability or trustworthiness [36, 269, 336]. Conversely, a system considered useless because of bugs or slowness is unlikely to become a true innovation. The perceived usefulness of a system is enhanced by consistent and accurate performance. The

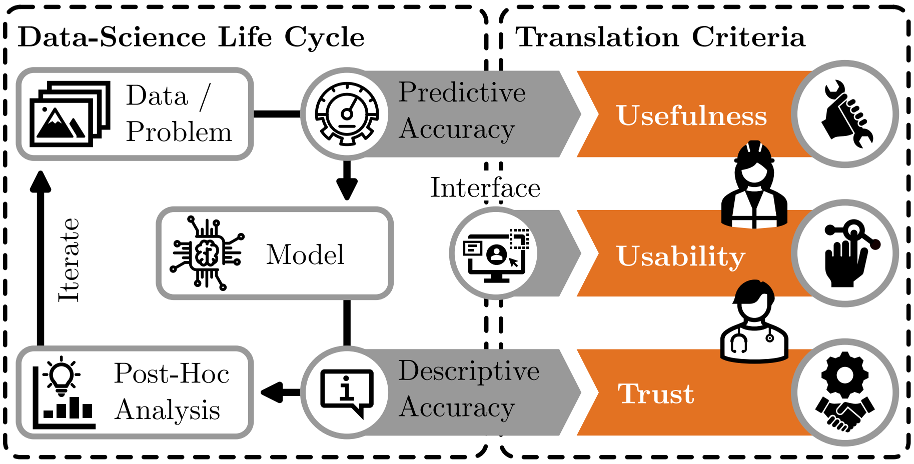


Figure 3.1: Data-Science Life Cycle in combination with the *Technology Acceptance Model*. Adapted from [62, 223]

effectiveness of a technology is demonstrated by the synergistic alignment of its functions with the most appropriate user applications [330]. In terms of Murdoch et al. [223], this criterion would correlate to the *Predictive Accuracy* of a model.

- **Trustworthiness:** Trust is a central issue in the context of artificial intelligence. The *European Commission's High-Level Expert Group on Artificial Intelligence* emphasizes in its report that "for a system to be trustworthy, we must be able to understand why it has behaved in a certain way" [128]. In the medical domain, Singh et al. [303] specifically define trust in the context of a clinical diagnostic system, stating that it must be "transparent, understandable, and explainable" to gain the acceptance of physicians, regulators, and patients. It is critical to recognize that trust is inherently subjective [186]. Murdoch et al. [223] would refer to this criterion as the *Descriptive Accuracy* of a model. The Ethics Guidelines [128] incorporate further aspects into the evaluation of trustworthy artificial intelligence systems, like privacy and accountability. It also includes considerations of fairness, sustainability, robustness, and security, with implications that extend into the realm of usefulness (see above).

- **Usability:** Usability refers to the user's interaction with technology, and Davis [62] and Bevana et al. [24] characterize this dimension as "perceived ease of use". Ease of use is defined by the ability of the system to be operated effortlessly, resulting in the achievement of desired outcomes. A technology that is easy to use, requiring minimal effort, is typically adopted more likely than one that is difficult to use but provides equivalent benefits. Beginning in the early years of usability research in 1980 [24], the concept has undergone several reinterpretations but always emphasizes the effectiveness of the interaction between users and systems. Currently, it is often referred to as *user experience* [300] and is a primary concern within the research field of human-computer interaction.

To establish a visual relationship between the three criteria, Figure 3.1 illustrates the data-science life cycle, emphasizing the machine learning interpretation process [223]. The *Predictive Accuracy* of a model originates from its performance on a test set or in daily use, typically referred to as **usefulness**. **Trustworthiness** comes from a model's *Descriptive Accuracy*, manifested in the consistency between the model behavior and its explanation. This work integrates the contributions of **usability** into this process by adding a user interface that facilitates interaction with the AI-infused technology. These three critical criteria guided the formulation of research questions that actively influence the successful adaptation of any new technology, whether it be a self-driving car, an *in vitro* diagnostics device, or a *medical Tricorder*, regardless of its promise.

3.2 Machine Learning for Cytological Image Analysis

3.2.1 Shifting Complexity from Hardware towards Computer Vision

Unlike other biomedical imaging techniques that examine entire body parts, high-throughput flow cytology is characterized by the acquisition of millions of digitized cells [148, 187, 235, 353]. The challenge lies in obtaining associated ground truth labels and segmentation [187, 218]. Human-assisted labeling [2, 32, C, 132] becomes essential, where a human oracle [79] primarily identifies samples that are difficult to label, given the impracticality of manually classifying a large number of cells. However, human classification is limited by the lack of contrast between cells, as expressed by Figure 2.2. Also, there is potential disagreement in finding a commonly accepted segmentation [87, 187, 197, 286]. Achieving optimal focus is also subjective. To facilitate supervised learning, collaboration with other disciplines, e.g., enabling *immunomagnetic separation* [333], is necessary to acquire an initial training dataset. Uniform data standards are needed as variations between devices and reference measurements pose hurdles [54, 152, 201, 298, 306]. Paramount is the challenging integration of data- and resource-intensive machine learning methods into clinical IT systems [39, 105].

The primary goal is to ensure that the developed solutions serve as decision support for patient care and research. Gaining the trust of the biomedical community, which adheres to rigorous approval processes before adopting new technologies, is essential [134, 210]. Since 2015, the number of papers showcasing the use of deep neural networks in medical image processing has rapidly grown [18, 81, 140, 187]. However, translating innovative ideas into routine clinical practice is daunting, underscoring the need for conscientious groundwork rather than imposing approaches on the research community. For this reason, these surveys [144, 251, 252, 295, 353] give a comprehensive overview of the application of machine learning methods, in particular deep neural networks, to medical image analysis. The following sections highlight major solutions and their applications in biomedical contexts.

3.2.2 Related Publications and Applications

Out-of-Distribution Detection. Several approaches are available for *out-of-distribution detection*. The most obvious variant is to define allowed ranges for specific morphological characteristics of a cell, such as its size or mass distribution [H, 333]. Fragments or platelets are then considered outliers and rejected so that the model is not forced to assign them a leukocyte class illogically. However, this method proves to be very inflexible to adjustments of the microscope or sample preparation. This procedure discriminates against heterogeneous cell classes and requires manual adaptations to evaluate new cell types. Additionally, the quality of the achievable filter is limited by the method used to gate morphological features. This may result in lower-than-expected performance when out-of-distribution samples exhibit similar characteristics to in-distribution samples. Methods that can make more autonomous and generalizing decisions would be preferable.

One-class classification [221] introduces an additional class that serves as a reservoir for all outliers, such as erythrocytes that survive the lysis process, cell fragments, or other unwanted blood components in the analysis of leukocytes (compare Section 2.2).

Researchers strive for an object description that allows detecting events deviating at a certain magnitude from the "normal" state. This task can be accomplished using *kernel Principal Component Analysis* [130] or one-class SVM [289, 322]. Ruff et al. [280] propose the *Deep Support Vector Data Description*, which the authors later refined in a variant that aims to make outlier classification comprehensible using heatmaps [192]. Objects that deviate significantly from the norm can also be identified based on reconstruction errors, e.g., with autoencoders [47, 368]. Similarly, SOMs [31, 75, I] or *Nearest Neighbor* algorithms [30, 163] can provide a quantification error metric to specifically detect outliers. This method has been successfully tested on blood cells by several research groups [265, 361]. Post-hoc outlier detection is possible by determining the uncertainty in the neural network classification [C, 139, 167, 228, 293], whether through *softmax* output or other distance measures. Their calibration is critical to avoid over- or underconfident networks [113, I]. For identifying unknown outlier types, the *Outlier Exposure* method of Hendrycks et al. [124] is an additional option.

Since human expert knowledge is costly in the biomedical environment and the ground truth is relatively sparse, approaches that can learn from unlabeled data are wanted. These so-called self-supervised approaches include *Contrastive Learning* [48, 320], which uses image transformations to generate correlated views of a cell. These positive pairs are then used to learn similarities and push out highly deviant samples. The *Self-supervised Outlier Detection Framework* by Sehwag et al. [291] is another technique that uses representation learning. Not requiring a ground truth dataset sounds promising, but so far these algorithms have mainly been tested on macroscopic images. The image transformations used, such as *jittering* or *random cropping*, are unsuitable for the low-contrast phase images. Note that, as crucial as outlier detection is in keeping the data processing pipeline free of interference, it is essential to remember that investigated objects are biological material. Any deviation from the norm may be a hidden biomarker.

Segmentation of Blood and Tissue Cells. The segmentation step involves identifying individual cells in the image and determining contiguous pixel regions corresponding to a cell instance. Kulwa et al. [171] distinguish between classical and machine learning methods, the latter being available to the community since the mid-2000s [187, 208]. Among the classical techniques, thresholding, in particular locally adaptive thresholding [150] or Otsu thresholding [321, 354], is often used for biological objects. Segmentation algorithms that work on edges or contiguous regions [120] show advantages when dealing with cell clusters or aggregates. More recent publications rely on CNNs for the segmentation of images of cells in flow cytometry [212] as well as for cells on substrate [282, 309]. Combining different architectures and learnings from macroscopic image processing is advisable since biological microscopy images have unique characteristics [70, 282].

Aggregates consisting of different components, such as leukocytes and platelets, are reliably handled by the *Watershed* algorithm [143, 227, 229]. The *U-Net* [80, 276] and other *Fully Convolutional Networks* [358] as well as the *Mask R-CNN* [123], are prominent choices for biological image segmentation, with the latter also able to classify detected objects directly. All CNN-based algorithms show exceptional performance on holographic images [131, 147, 174, 212], and continue to be refined for more precise segmentation of biological images [194, 207, 369]. To this extent, it is also possible to transfer knowledge from the stained imaging approaches onto the label-free phase domain, facilitating the segmentation of subcellular structures [49, 165, 244].

Recent advances include transformer architectures for reliable segmentation of tomographic and histological images [91, 334]. The built-in *attention* mechanism in these architectures is also valuable for generating visual explanations [42, 241, 356].

It is noteworthy that generating ground truth and assessing segmentation accuracy is challenging [218]. The resolution of cells is typically insufficient to determine whether one pixel is on the cell membrane or in the background. Methods that align human labels may improve the ground truth quality [87, 187, 197, 286]. Consequently, it is impractical to insist on an identical overlap between the ground truth and the result of the segmentation algorithm. The priority is to preserve the detailed morphology of the cells in the segmentation mask [21], ensuring that cells are not erroneously split, merged, or subjected to the introduction of artificial edges. Also, semi-supervised [315] or generative approaches [8] might help here.

Feature Extraction & Classification. "One especially promising application of machine learning methods for QPI studies is in the classification and identification of cells and tissues" [230]. Several approaches have been explored in literature to accomplish the task of classification. Today, the evaluation of macroscopic images, such as MRI and CT scans, shows a tendency towards large CNNs compared to their slower adoption in the QPI microscopy community. Although the development in this direction is a few years behind, it is expected to soon reach a comparable level in cytology [18, 81, 140] (see Figure 3.2a).

In oncology, especially in the monitoring and evaluation of cancer tissue cells, using handcrafted morphological features for classification is standard practice [21, 188, 332, 333]. Typically, predictions are made using SVM [177, 178], while RF and *Linear Regression* are also used, as exemplified in the grading of prostate cancer [229]. In recent years, there has been an increasing dominance of CNNs in this discipline [18], particularly in the differentiation of breast cancer cells [46]. Ben Baruch et al. [23] demonstrated the flexibility of freely combinable networks by fusing stationary and variable image streams to improve the discrimination of different cancer cell lines. In addition, generative approaches have been used to address the challenge of training CNNs with a vast amount of parameters. These methods allow CNNs to perform even when trained on a dataset of fewer than 300 images [278].

In hematology, particularly leukemia research, the differentiation of leukocytes relies on examining their appearance in blood smears. Biologists resort to distinct visual characteristics such as granularity and the structure of the nucleus to reliably categorize cells [13, 16, 359]. In contrast to *flow cytometry*, QPI is an imaging method that leads to the prevalence of morphological features for classification. Typically, these features are mathematical summaries of cell-occupied pixels in the phase images [117, 254, 332]. For example, the cell contour provides insight into perimeter and roundness, while the combination of pixel values within the cell allows mathematical estimation of optical volume and texture characteristics such as granularity. Another possibility to describe the cells are model-based computational rules inspired by phenotypic and physical leukocyte properties [189]. SVMs [237, 245], *Gradient Boosting* [226], *k-Nearest Neighbors* [359], or RF [247] use the presented morphological features to classify unknown cells into appropriate leukocyte classes based on a similar feature distribution. However, research trends are shifting toward neural networks because manually generated cell descriptions struggle to capture the nature of cells accurately [44, 148, 243, H]. Unlike human-derived rules, neural networks adapt their feature extraction during training, providing greater flexibility.

Immunothrombosis, also a phenomenon within hematology, entails the interaction between leukocytes and thrombocytes. This interplay between the human immune system and the coagulation system yields crucial insights into inflammation [313], sepsis [86], COVID-19 [1, 107, D, 235], and cardiovascular diseases [86, 313]. During this process, cells undergo partial activation [156] and adhere to each other for several minutes before the agglomeration dissolves again [E]. The resulting assemblies of various cells and cell types form complex structures that require thorough unraveling. Morphological features offer limited utility in this context. If the goal is solely to determine the size of platelet-aggregates, these features, as outlined by Nishikawa et al. [235] or Khan et al. [156], prove sufficient. However, techniques such as *U-Net* [108] or *Mask R-CNN* [D, E] emerge as the preferred methods for a more in-depth understanding of microthrombi composition.

Parasites and bacteria have diverse characteristics, appearances, and behaviors, raising numerous scientific questions that underscore the complexity of data processing. In the case of malaria, where the *P. falciparum* parasite infects erythrocytes, the goal is to identify the disease and determine the stage of parasite development without relying on markers. Using morphological features from QPI microscopy, the disease profile can be inferred using methods such as *Linear Discriminant Analysis* [146], SVM [100], or manual gating [332]. A study in 2018 by Poostchi et al. [258] shows that SVM, in particular, is often used for this purpose. It is closely followed by CNNs, *k-Nearest Neighbors*, and even *Threshold Deciders*. This arsenal of techniques helps uncover the intricate features and stages of parasites and bacteria in the context of infections. Related publications demonstrate here the versatility and applicability of various data processing methods for this use case [142, 146].

3.2.3 Current Challenges

Scientists, including those referenced in various surveys [145, 230, 254], strongly believe that the combination of label-free holographic microscopy and machine learning has great potential to revolutionize laboratory and clinical practice. Continuous advances in computer vision models inspire new ideas to improve the segmentation and classification of phase images. Existing approaches are continually being refined to address specific challenges in QPI microscopy. However, caution is warranted as many studies, while demonstrating feasibility, are not yet ready for real-world translation [105]. The scarcity of clinical studies makes it difficult to validate the generalizability and robustness of the approaches [209, 230].

Still, much of the current work relies on hand-crafted morphological features, which are considered accepted and reliable in the community [21]. Cells are filtered by this rigid description using fixed value ranges [332, 333] or classified via manual gating [312]. However, this practice may prove too inflexible to account for biological variability and pathological changes in broader trials [H]. Since creating and adapting the necessary computational rules require strong assumptions and massive domain knowledge [145], experts assume that deep learning will soon wholly replace morphological features [258]. However, others emphasize the need for rigorous evaluation to determine the suitability of a data-driven approach for each application [145].

Consistent workflow standardization is a notable challenge, particularly in sample preparation [E, 162] and hardware [175]. Yet, in software, inconsistencies persist between preprocessing methods for neural networks [54, 201, 298], data storage, and transmission [152, 306, 318]. Unlike traditional *flow cytometers* or *hematology analyzers*, QPI technology lacks quality standards and certified operational limits [210]. Yet, a machine learning model's effectiveness ultimately depends on the data quality [285]. Furthermore, unresolved data privacy issues arise when training samples can be reconstructed from models [367].

While introducing machine learning methods in cytology offers numerous advantages, it also poses inherent dangers similar to those observed in handling macroscopic images and other data sets. Problems such as poor data quality, overfitting, and limited generalizability have been demonstrated, particularly in oncology by Amorim et al. [7]. Consequently, these shortcomings lead to unrecognized confounders, important overlooked subgroups, or the ignorance of rare events [187]. The opacity of the decision pathways using uninterpretable features or neural networks introduces the risk of making false predictions with undue confidence, especially in the presence of subtle adversarial noise [85]. Machine learning models can be susceptible to biases caused by imbalanced datasets due to factors such as a lack of donors, challenges in obtaining ethical approvals, or the rarity of certain diseases, as Litjens et al. [187] point out. Undesirable effects such as *shortcut learning* [92] may also contribute to *discrimination bias* [154].

The limited human understanding of black-box approaches' overall behavior [111], coupled with the potential for undesirable and unpredictable model behavior [36, 111, 154, 225, 272], raises concerns. Such uncertainties are detrimental not only to scientific confidence in this emerging platform technology but also to patient health. As a result, proactive measures to mitigate these risks are essential.

Jo et al. [145] recommend the early involvement of experts from different fields to effectively address these challenges, as equitable collaboration is essential for reliable design and development of new approaches. Implementing user-centered design [129, 238, 290], human-in-the-loop processes [132, 252], and XAI [18, 40, 246, 325] will be critical to overcoming the complexity associated with current trends in machine learning. Moen et al. [218] "recommend integrating tool building with biological discovery. Deep learning is a data science, and few know data better than those who acquire it. In [their] experience, better tools and better insights arise when bench scientists and computational scientists work side by side – even exchanging tasks – to drive discovery." These measures not only increase the transparency and interpretability of machine learning models but also contribute to the responsible and ethical use of these technologies in the evaluation of cytological specimens. Hence, the following section describes which techniques are available for this purpose and how their use can be evaluated.

3.3 Interpretability of Machine Learning Methods in Interdisciplinary Research

In laboratory-based biomedical tests, the fundamental metrics are the *specificity* and the *sensitivity* of the respective underlying assay. The gold standard method, which achieves 100% in both metrics, sets a benchmark for reliability [176]. Novel test methods stand out by their time- or cost-effectiveness, but their proximity to this gold standard assesses their credibility. Similar metrics are also used in computer vision to evaluate the performance of data processing pipelines and machine learning algorithms. Published works often report improvements in *precision* and *recall* percentages for specific applications on designated datasets, claiming to be the new state of the art. The pure focus on *Predictive Accuracy* [223] may be appropriate for *in vitro* diagnostic products, which rely on the deterministic physical and chemical processes governed by natural laws. However, this attitude presents a distinct challenge in the context of opaque learning procedures [134, 279]. As explained in the previous section, the principles guiding these cases' decision-making are undisclosed, posing unique risks as humans lack insight into the fundamental mechanisms driving these outcomes [36, 272].

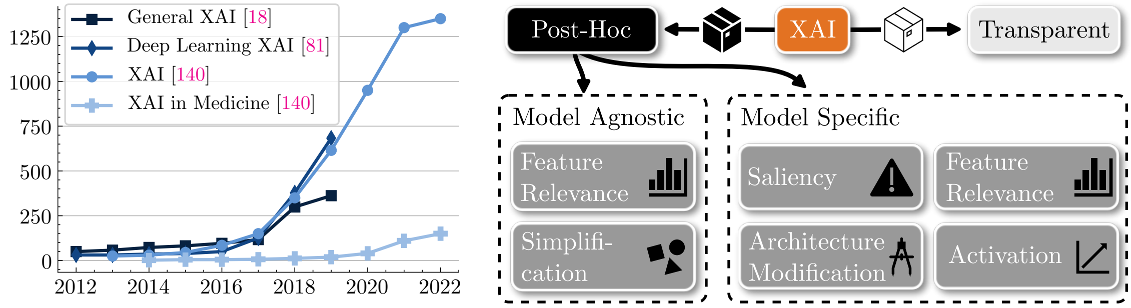
3.3.1 The Quest for Intelligible Machine Learning

Many scientists criticize the exclusive emphasis on accuracy when assessing learning methods, which can lead to unintended machine behavior [36, 269, 271, 336]. To ensure that correct predictions are not mere chance occurrences due to unknown confounders, scientists advocate for a deeper understanding of the decision-making processes in machine learning pipelines [106, 256, 277, 290, 318, 340]. The adverse effects of inaccurate predictions and unintended developments in artificial intelligence have prompted legislative intervention. The European Union's ethics guidelines [128] see transparency and, thus, the explainability of internal processes as an integral part of maintaining trustworthy technologies. The *General Data Protection Regulation* [77], which has been in force since 2018, even stipulates a "right to explanation" for the respective data subject.

Although experts agree on the mandatory requirement for explainability, there is still no consensus on the best methods to achieve it. This lack of agreement leads to discussions about meeting legal requirements and what demands the developers of such systems must place on themselves. Efforts to push the development of explainable approaches have increased, as seen in the publication of guidelines [6, 294, 299]. Some lure developers with the argument of efficiency enhancement through explanations, identification of inconsistencies, and user feedback [132, 255, 347]. In the recently published *AI Act* of the European Union [78], applications in the healthcare sector are classified as high-risk technology and are subject to ambitious requirements. That again illustrates why there is a technology acceptance problem [62, 249, 347] without a well-founded justification of technical decisions. It also raises the question of whether approaches can develop their full potential without the trust of interdisciplinary users and researchers despite their superior technical performance [242].

Given the significant implications involved, it is vital to approach the "right to explanation" with caution to ensure its fundamental purpose of providing clarity to individuals is not compromised [279]. To achieve clarity, the recipient of an explanation plays a crucial role in interpreting the offered knowledge chunks [69, 328]. Although explainability and interpretability are often used interchangeably, they are distinct aspects of the same concept. Models can be difficult to explain due to their inherent complexity, which creates a trade-off between interpretability and completeness, as noted by Gilpin et al. [97]. To distinguish between these terms, this work uses the definition proposed by Holzinger et al. [134]. According to their conceptualization, interpretation is the "mapping of an abstract concept into a domain that the human expert can perceive and comprehend." On the other hand, explanation involves identifying a "collection of features of the interpretable domain, that have contributed to a given example to produce a decision."

In the context of the interdisciplinary research workflow outlined in Section 2.2, the explanation refers to the outputs generated during the *Learning* stage. These outputs can take the form of numeric, textual, or visual information. They serve as a proposition for all participating scientists, allowing for interpretation in both *Technical Analysis* and *Medical Analysis*. There are differing opinions on the use of opaque models in these applications. Some suggest avoiding them altogether [279], while others advocate for the exclusive use of causal models [134], acknowledging that complete interpretability may be unattainable [186]. However, research efforts continue on multiple fronts, recognizing the absence of a unified standard [18, 111, 223]. A combination of factors and techniques will likely be necessary to achieve interpretability, serving as a confidence-building measure [66, 242, 249]. It is crucial to adapt the level of abstraction to align with the stakeholder's role and background domain. Stakeholders involved in the development, implementation, and evaluation of technologies will shape and survey their usage. The following section



(a) XAI Publications (data from [18, 81, 140]) **(b)** Excerpt from the XAI taxonomy by [18]

Figure 3.2: The field of XAI is gaining importance due to the lack of transparency in modern machine learning methods. AI and the need for explanation are slowly entering the medical domain.

summarizes the key techniques essential for managing the complexity, given the multitude of potential explanation techniques.

3.3.2 Explainability Techniques for Machine Learning

Section 3.2 discussed the continuous improvement in the *Predictive Accuracy* of machine learning, particularly neural networks in segmentation and classification. Despite this progress, a significant research gap remains to improve their applicability in critical areas of life. To make these systems more trustworthy, there is a growing emphasis on XAI to elucidate otherwise opaque decision-making processes. Figure 3.2a shows this almost exponential trend. Numerous taxonomies that categorize XAI methods based on their use or technological aspects have emerged. A notable attempt to consolidate these approaches is presented by Barredo Arrieta et al. [18], a taxonomy to which this work is aligned. Figure 3.2b illustrates this taxonomy's main branches, distinguishing between inherently **transparent** models and those requiring **post-hoc** explanations for interpretation. The latter category is further subdivided into **model-agnostic** and **model-specific** methods. Several other taxonomies, such as those proposed by Lipton [186], Guidotti [111], and Kamakshi [338], place significant emphasis on whether the explanation takes place within a local framework or allows for global insights into the model. Given the specific focus of this work on cell image analysis, a universal review of XAI methods is out of scope. Interested readers are encouraged to consult other sources [111, 133, 325]. What follows, however, is a brief exploration of essential XAI methods and their application in biomedical projects, particularly with imaging techniques in cytology.

Transparent Methods. Transparent models, by design, eliminate the need for additional explanation. Interpretation for non-technical users relies, for instance, on visualizations or textual explanations. Models such as *Linear Regression* or *Decision Trees*, already used for cancer cell assessment [229], demonstrate simplicity. *Generalized Additive Models* [40, 119, 196] facilitate the identification of interactions between different features (predictors) and find application in all likelihood-based regression models. In the medical field, they have proven effective in assessing the risk of pneumonia [37]. In the field of neural networks, the application of *Bayesian Deep Learning* [139] provides a means to transparently communicate uncertainty within a network. This technique has been successful in detecting oral cancer [308]. In addition, *Bayesian Deep Learning* finds applicability in the detection of cancer cells in urine [151] and the segmentation of volumetric image data [342].

Post-Hoc Explanations. Post-hoc explanations are explanatory methods that aim to retrospectively translate a model's internal processes into descriptions, metrics, or visualizations that are more understandable to human observers than the machine learning model itself.
When the adjustments depend on the model's architecture or internal processes, the explanations are referred to as **model-specific**.

- **Guided Back-Propagation** [311] modifies the architecture of a neural network to propagate only non-negative gradients, providing insight into the network paths that contribute to decision-making and those considered irrelevant. With that technique, Wang et al. [344] demonstrated that their CNN focuses on regions containing mitochondria directly related to cellular metabolism. In different microscopy setups, Nishimura et al. [236] showed the contribution to cell classification on a pixel level by *Guided Back-Propagation*.

- A closely related method is **Deconvolution** [362], which works on output rather than input gradients. A parallelly constructed *DeConvNet* [363] projects CNN-extracted features back to input layers using unpooling and rectification. This technique, classified as activation-based by Barredo Arrieta et al. [18], was used by Sui et al. [316] to localize "high-grade cancer regions" in tissue samples.

- **Layer-Wise Relevance Propagation** [10] uses special propagation rules for a similar effect. The influence of a neuron on subsequent layers is determined using a relevance metric and propagated to the network input. Successful applications of *Layer-Wise Relevance Propagation* include macroscopic medical images [19, 28] and microscopic cell images [27].

- **DeepLIFT** employs a comparable approach in its explanations, relying on the product of gradient and input, similar to *Layer-Wise Relevance Propagation*. The method of Shrikumar et al. [301] introduces a reference activation and defines the deviation of a single neuron from this reference as its contribution. Significance is conveyed by back-propagation, which is similar to *Guided Back-Propagation*. *DeepLIFT* has demonstrated its utility in detecting COVID-19 in X-ray images [12] and multiple sclerosis cases [195].

- **Gradient-weighted Class Activation Mapping** [292] does not operate at the pixel level but on connected regions in the feature map of the last convolutional layer, avoiding the fully connected layers. Despite potential limitations in semantic information propagation, *Gradient-weighted Class Activation Mapping* provides convincing results, as confirmed by experiments on leukocytes [99, 173]. *Gradient-weighted Class Activation Mapping* reveals that the network relies on distinctive intracellular granules for its decision. Nagao et al. [224] thus determine meaningful cell organelles to observe the individual phases in the life cycle of a cell. The method also supports radiologists in identifying COVID-19 [250].

- **Saliency Maps** provide an alternative approach, using internal gradients to identify neurons "belonging" to a particular class. Areas of high saliency highlight image regions that contribute significantly to decisions for a respective class. Simonian et al. [302] call this weakly supervised object localization. Ferriera et al. [83] use this technique with a *VGG-16* model for lesion detection in *Pap smear* samples. However, it is insufficient in some cases for other imaging techniques [161].

If the explanation is entirely independent of the underlying model, it falls into the category of **model-agnostic** methods.

- One of the best-known representatives of such an explanation for image data are **Local Interpretable Model-agnostic Explanations (LIME)** [271]. As the name implies, LIME is specifically designed for single examples. This technique involves learning a linear proxy model, which can be categorized as simplification and rule-based learning. LIME finds application in explaining *VGG-16* models specialized for Parkinson's disease detection [202], or custom CNNs aimed at detecting lymph node metastases [248].
- Subsequently, **SHapely Additive exPlanations (SHAP)** [198] were introduced for machine learning methods, drawing on an older method in game theory. SHAP allows statements about the relevance of features or image regions, taking into account whether the feature was included or omitted during learning. Skillful sampling strategies and integration into the training process eliminate the need for costly model retraining [314]. Combinations of LIME and SHAP are known as *KernelSHAP*. When *DeepLIFT* is used as an approximation method in SHAP, it is referred to as *DeepSHAP* [153], which allows SHAP values to be determined over an entire neural network. For example, Gotkas et al. [102] use tree-based SHAP to improve the explanation of human mesenchymal stem cell classification in high-throughput scenarios. SHAP has also proven successful in visualizing important features in malaria detection [266] or in the categorization of cysts and tumor cells using CNNs [25].

Transparency Paradox. This section concludes by highlighting a notable disagreement within the data science community regarding a potential paradox between *Predictive* and *Descriptive Accuracy* [18, 223]. Some assert the existence of a trade-off [242], claiming that accuracy often has an inverse relationship with the interpretability of a method [133, 355]. Others, however, argue that the evidence for such a trade-off is weak [125] or nonexistent for post-hoc explanation approaches [223]. Instead, they propose the idea that interpretability plays a positive role in improving models, fostering a mutually beneficial relationship [279]. Care should be taken not to compare linear models with highly nonlinear ones or simple mathematical edge detectors with convolutional filter layers with thousands of parameters [125]. The comparison should be limited to models that require internal modifications, *model-based interpretability*, as Murdoch [223] calls it, to improve transparency. Figure 3.3 illustrates the according impact on the individual model classes. Otherwise, there is a risk of comparing disparate elements, like apples and oranges.

3.3.3 Evaluation of Explainability and Interpretability

In software development for low-risk applications such as infotainment systems, simple unit and integration tests often ensure security. However, the application of machine learning methods in a clinical context presents unique challenges [69, 78]. These challenges include many models' inherent uncontrollability and statistical learning behavior. In addition, the variability of the biological samples complicates their application. In particular, the potential consequences for life and well-being are significant when used in critical healthcare scenarios [154, 225, 318, 329]. For this reason, the European Union requires human oversight [128], which in turn necessitates interpretability and transparency of the machine learning methods. However, despite their recognized importance, there remains a

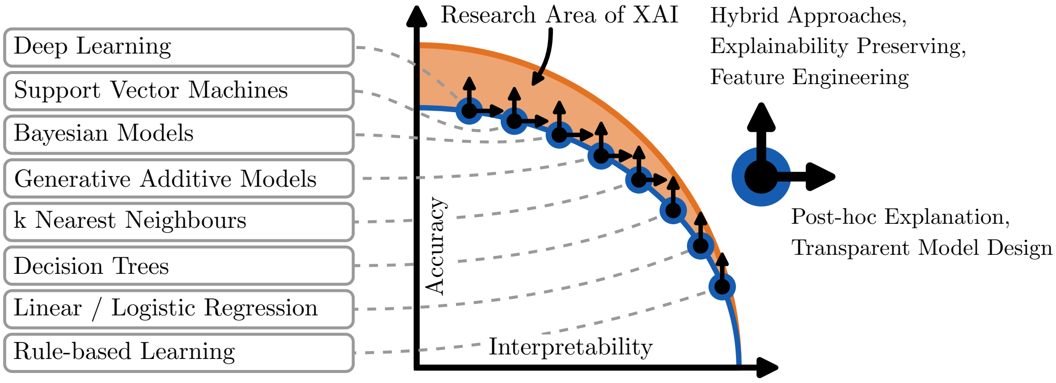


Figure 3.3: XAI Paradox: Is there a potential trade-off between model accuracy and interpretability? Adapted from [18]

lack of consensus within the scientific community on the precise definitions of interpretability and explainability in the context of machine learning. Furthermore, the establishment of standardized metrics and methodologies for measuring these values is an ongoing challenge.

Latent Dimensions of Interpretability. Interpretability and explainability indirectly influence the translation criteria, operating through the user's understanding and perception. Doshi-Velez and Kim [69] term these influential factors as latent dimensions of interpretation, visualized in Figure 3.4. Various techniques generate explanatory statements for a model or its results. However, these statements have no direct contribution to the perceived trustworthiness. A subjective interpretation layer receives these explanations and modifies their information content based on various factors. Below are some known modifiers:

- **Domain and Background:** The specialist background of a user or operator plays a pivotal role [69]. In the context of the research tool discussed herein, the primary users are researchers in data science, biomedical engineers, biologists, and physicians. The evolution of technology may broaden the user base to include specialists or patients from diverse societal domains [328]. The available domain knowledge, such as *AI literacy* [98, 249] or *computer literacy* [318], significantly shapes an individual's capacity to comprehend explanations.

- **Role:** Despite a clearly defined background, individuals may assume different roles in evaluating technology [328]. A person could be a developer striving for optimal algorithmic performance or a test engineer focusing on application security [22]. This role variance necessitates diverse explanation techniques. For instance, a developer might prioritize performance measures, while a regulator seeks insights into internal processes and causal relationships. Similar effects also hold for the biomedical and other domains [329].

- **Application:** The nature of the task at hand modulates the perception of interpretability [69, 329]. In exercises like leukocyte classification, where the sheer volume makes manual examination impractical, understanding general algorithmic behavior suffices. However, tasks involving rare circulating tumor cells demand a diligent interpretation of individual instances for certainty.

- **Explanation Method:** The choice of explanation method significantly influences interpretation. The type of explanation (numerical, textual, visual, etc.) can vary

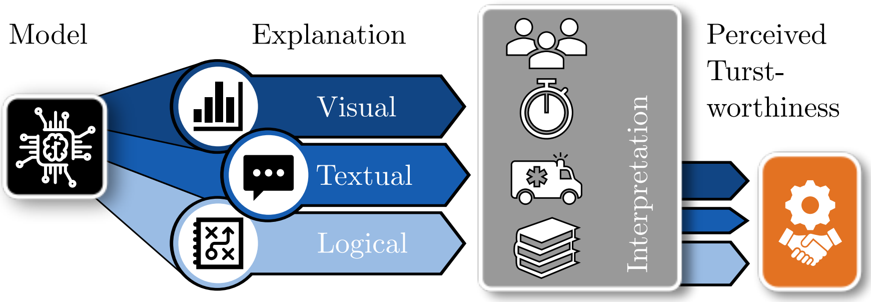


Figure 3.4: The latent dimensions of interpretability shape the perception of XAI methods. While the explanations were equally meaningful from a technical point of view, they appear to be scaled based on a subjective interpretation. Inspired by [69]

in effectiveness based on the required interpretation, contingent upon the level of abstraction inherent in the explanation [50, 69].

- **Time Constraint:** The time available for interpretation is critical in understanding a situation [69]. Different scenarios, such as an emergency room versus a research tool, entail distinct time constraints [329]. The level of abstraction in the explanatory method must align with the respective time frame to facilitate an appropriate interpretation.

This list does not claim to be complete, as several other factors and confounders can influence the interpretation of explanatory offers. The reader is referred here to reading [22, 26, 69, 307, 329].

Explainability Metrics. The interpretive foundation for stakeholders relies on the set of explanations outlined in Section 3.3.2. In general, measuring the success of an explanation approach is not trivial [66]. The evaluation of these explanations against specific requirements, as outlined by Tonekaboni et al. [329], often occurs qualitatively, given the inherent reliance on human judgment. Nevertheless, certain evaluations may be quantifiable or boolean. The impact of an explainability metric varies depending on the scenario and the interpreter.

- **Domain-appropriate Representation:** The value of an explanation is measured by its alignment with the requirements of the current application. It should present only information that is appropriate in terms of scope, relevance, and direct applicability. This criterion seeks to answer the question: Is the explanation currently helpful? [69, 329]

- **Potential Actionability:** Explanations should have the ability to positively influence subsequent decisions by providing valuable information. Integration into the clinical workflow is critical for actionability, with the importance of the explanation depending on the time available for evaluation and the potential impact on diagnosis or treatment. [329]

- **Consistency:** Explanations must behave deterministically and injectively [329]. Any changes in the model statements should be directly reflected in the explanations, regardless of design variations. A model-agnostic explanation also gains in this property in the context of the previously introduced taxonomy.

- **Comparability:** Comparability of explanations implies two conditions: they should be comparable across models (model-agnostic) and facilitate comparison with past situations or experiences, possibly through reference measurements or historical data. Consistency remains a crucial consideration, interfering with the comparability metric. [40, 325]

- **Anomaly Detection:** Explanations should enable the detection of anomalies in the data, such as outliers, quality variations, or hidden biomarkers. The goal is to attribute model predictions to these effects, thereby increasing their visibility and preventing noise from obscuring them. [307, 349]

- **Uncertainty:** "Trustworthy medical AI systems need to know when they don't know" [110]. A good explanation method must clearly communicate uncertainty in both the model and the explanation [110, 134, 167, 293, 329], which affects the validity of other criteria and is consistent with regulatory requirements [128].

- **Fairness:** Fairness is a critical metric in many fields. A reliable explanatory method should consistently identify and express the presence of bias in data or predictions, thereby mitigating the risk of discrimination. [61, 111, 154, 242, 251]

- **Completeness:** As explanations simplify complex models or use proxy models, assessing how complete the actual model is represented becomes vital. A good explanation strikes a balance, conveying essential features without overwhelming the user with cognitive overload [69, 97, 186]. Algorithmically transparent models are an exception here. These are understandable in a way that they can be presented in their full scope [18, 186, 307].

- **Locality:** For effective decision-making in individual cases, an explanation should provide a local justification that allows systematic errors to be detected at their root. This type of explanation is advantageous for careful analysis, especially in the case of rare events. [69, 97, 186]

- **Globality:** In contrast to local explainability, understanding the general behavior of a model is necessary for assessing the safety of many applications [69]. While it may be impractical to examine every sample, knowing the general trends of the model is critical to ensuring reliability in the face of anomalies or biases [347, 349]. In the presence of such disturbances, the method must communicate these uncertainties transparently.

- **Interactivity:** Interacting with an explanation greatly facilitates its interpretation. Features such as rotating and zooming graphics, adjusting color schemes, and controlling the visibility of elements increase user engagement and facilitate optimal cognitive absorption, especially in the final stages of decision-making. [132, 197, 307, 349]

- **Predictive Accuracy:** *Predictive Accuracy* is twofold: the explanation itself should be highly accurate, free of numerical singularities or discrepancies, and its integration into the model should not compromise predictive performance [242, 307, 349]. On the contrary, it should have the potential to improve *Predictive Accuracy* through error correction and human-in-the-loop approaches [132, 279].

This compilation of explainability metrics is tailored to the specific relevance of the biomedical research tools under consideration. For a more comprehensive evaluation, Sokol and Flach [307] provide a more encompassing catalog of functional, operational, usability, security, and validation requirements applicable to the assessment of explainability methods.

Evaluation Process. Finally, as with any measurement, defining the associated measurement protocol or study design is necessary. Many sources refer to the work of Doshi-Velez and Kim [69], which aims to rigorously scientifically investigate interpretability regardless of its field of application. The authors describe three distinct evaluation approaches to determine the impact of explanation methods on interpretability and, consequently, on the overall trustworthiness, usefulness, and usability of a model.

- **Functionally-grounded Evaluation:** This approach, considered the most objective, eliminates the need for a human evaluator. However, it does require a formal description of the interpretation, which is challenging at the beginning of the project. This type of evaluation is typically done after experience with human testers and is similar to unit testing in software development. It is appropriate when monitoring performance changes or other types of regularization in a model that has already been classified as interpretable [C].

- **Human-grounded Evaluation:** In cases where modifiers' impacts on the interpretability of an explanation are not sufficiently determined, human critics play a key role in the evaluation. To simulate real-world scenarios, a simplified task that closely mirrors the actual task is presented. The assessment of explainability should be independent of the model's predictive power. There are three recommended study designs: In *binary forced-choice*, users have to choose between a few explanatory approaches, depending on which one they personally find most beneficial [V]. In a *forward simulation*, the testers are given a specific explanation approach and must imitate the model's decision. However, the decision does not have to be correct. In a *counterfactual scenario*, the testers are presented with input (possibly faulty), output, and the corresponding explanation. The testers must then specify what needs to be changed in this combination in order to change the model output to the expected one.

- **Application-grounded Evaluation:** Considered the most realistic, this type of study observes users or stakeholders in real-life situations, providing insights into practical utility. While not universally applicable, it proves feasible in many use cases. For example, a human-in-the-loop scenario for cell segmentation correction can involve laboratory experts working on real clinical samples without disrupting clinical workflow [VI, 317, 329]. However, feasibility depends on the availability of sufficient domain experts.

Despite the justification for all three evaluation methods, the current state of the research project only allows for the feasibility of human-grounded scoring. *Functionally-grounded Evaluation* is precluded due to the lack of confirmed proxy tasks and the need to determine modifier weights through human feedback. *Application-grounded Evaluation*, while more realistic, faces challenges in implementation due to difficulties in securing an adequate number of domain experts despite the availability of clinical samples.

3.4 User Interface Design for Biomedical Applications

3.4.1 Usability of AI-Induced Biomedical Systems

The interaction between users and technology, especially in certified clinical applications, is primarily facilitated by a user interface, as users rarely interact directly with the underlying hardware or software. The goal of the user interface is to promote the optimal handling of the technology. The last of the three translation criteria, **usability**, describes how well this handling works and how much effort is required [24, 62]. With regard to the explanation method for machine learning procedures and their interpretability, it is not enough to simply conjure up a visualization out of thin air [242]. The manner, time, and location in which a particular information or feedback is displayed are as important as the content itself [74, 90].

The integration of computers into clinical practice has exponentially increased the volume of diagnostic data available. New imaging techniques contribute to generating extensive data sets that require seamless integration into daily clinical workflows through digital data processing [54, 110, 133, 187]. Challenged to keep pace with this rapid technological evolution, developers and designers strive to find efficient ways to present new data through appropriate user interfaces. The overarching goal has always been to support medical professionals in their decision-making processes by facilitating multimodal information integration. However, examining some existing programs reveals shortcomings such as complex interfaces, shallow learning curves, and unclear functionality that hinder an optimal user experience and impede successful translation [135, 262].

A contemporary challenge is the increasing autonomy of data processing pipelines operating on large and diverse datasets, leading to more personalized diagnoses and abstract decisions. In addition, the growing influence of machine learning introduces the need for careful diagnosis not only for patients but also for algorithms, changing the basis of trust in medical decision-making. Combining these challenges, the graphical user interface becomes a critical element, requiring a clear and attractive presentation to ensure that both human and machine decisions are well-founded [277]. Entire books are dedicated to the integration of machine learning in the medical domain [267], but there is no mention of how to make it easy to use. Some developers and researchers want to counteract this trend with user-centered design.

User-centered design [238, 300] prioritizes human individuals, emphasizing their needs and prior knowledge. In the area of biomedical research tools, a modern user interface must meet the needs of an interdisciplinary audience [74, 214, 239, 329]. Given the diverse backgrounds of users in various domains, different requirements and vocabularies must be accommodated [26, 69, 97, 305]. As a result, researchers advocate tailored design guidelines for specific application domains and more personalized explanatory approaches for machine learning [126, 249]. Hybrid explanations and especially visualizations play an essential role here [168, 357].

In response to the rapid advancement of machine learning methods and the resulting requirements, numerous guidelines have emerged for the successful integration of these methods into software tools [6, 26, 74]. While there are existing best practices tailored for clinical applications [105, 294, 357], the field of standardized frameworks for AI in medicine remains an ongoing process [59, 191, 310]. This work uses these recommendations to formulate design rules for a biomedical research tool focused on cell analysis.

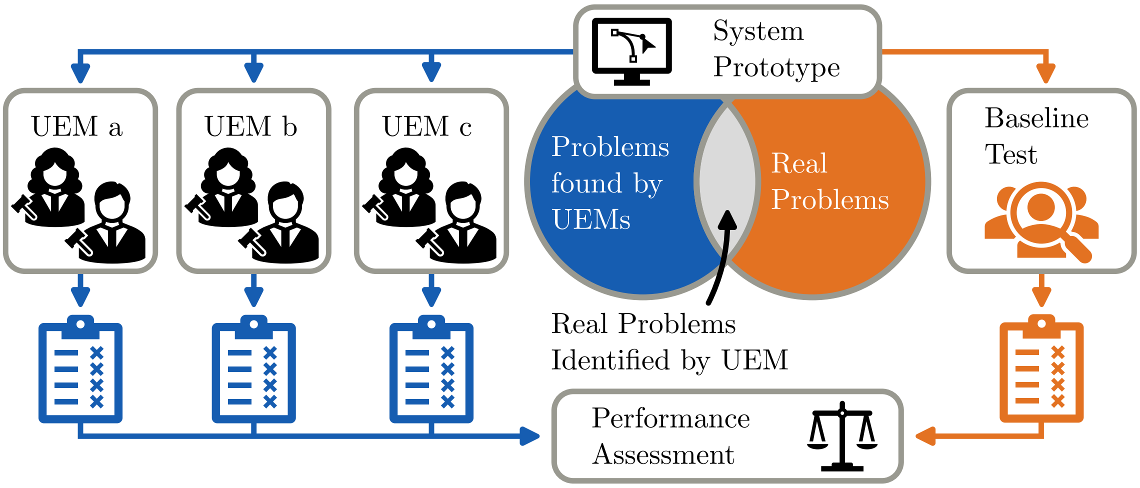


Figure 3.5: Assessment of Usability Evaluation Metrics: Different UEMs a-c have to prove their *Validity* and *Thoroughness* compared to a baseline user test. Adapted from [118]

3.4.2 Usability Heuristics

Analogous to evaluating a system's trustworthiness, the assessment of its usability lacks objectivity and is mainly tested via user studies. A reliable understanding of quality can be achieved if the participants are selected appropriately and their number is sufficient [184]. Usability, defined as the minimum effort required to operate software [62], is identified through users revealing usability problems while interacting with the user interface. As stated in the literature, the detection rate of usability problems per user ranges between 15% and 45%, although these detections often overlap and exhibit merely asymptotic behavior [118, 184].

For the target audience of researchers and clinicians in the presented application, factors such as availability and cost escalation make it impractical to comprehensively inspect all system functions in a single test. In addition, the assessment of long-term effects and the familiarization phase require repeated user participation [6], making it more pragmatic to integrate a cost- and time-efficient testing method into the software development cycle.

A remedy lies in purely heuristic evaluations [238, 299], which avoid the need for expensive domain experts. A small number of usability experts, even a single person, can evaluate the user interface under development based on predefined heuristic rules. The interface is examined for rule violations, and feedback on the score and severity of violations is provided to the developers. This iterative process significantly reduces workforce and costs and can be performed independently of the actual end user. Evaluators can range from *novices* to *single experts*, including *double experts* with experience in both usability and the target domain [231]. Noyes and Baber suggest that *novices* can identify about 22%, *single experts* 41%, and *double experts* 60% of actual usability problems with a robust heuristic rule set [238]. However, heuristic evaluations cannot entirely replace user studies and must be validated by real user feedback.

The effectiveness of heuristic rules is crucial, and Nielsen and Molich's list of ten general usability heuristics [233] from 1990 is still a standard work. While these heuristics offer universality, they may not address specific problems that arise in interdisciplinary settings or those posed by artificial intelligence. The need for domain- and application-specific heuristic rule sets becomes apparent, prompting a re-evaluation of their suitability for the development of new tools [126, 249].

There are also procedures for testing Usability Evaluation Metric (UEM). Hartson et al. [118] point out that there is a competition, drawn in Figure 3.5, between novel UEMs, such as domain-specific rule sets, and existing sets. The rule set that allows usability experts to accurately predict most problems is considered the better one. An asymptotic user study at the same stage of development serves as a reference, measuring the effectiveness of the UEM in terms of *Validity* (precision), *Thoroughness* (recall), and *Consistency* [118, 169]. It is crucial for effective UEMs to reliably detect severe usability problems early in the development process.

Chapter 4

Contributions

4.1 Adaptation of Machine Learning Methods for Holographic Cytology

The first core contributions of this work focus on improving the translation criterion of **usefulness**. While several effective machine learning approaches exist for image processing, their optimization is compulsory for the analysis of holographic cytology data. It is crucial to emphasize that this optimization should not be pursued in isolation but rather in consideration of the other two translation criteria, trustworthiness and usability. True innovation can only occur if all three criteria are met. This section systematically addresses the first research question:

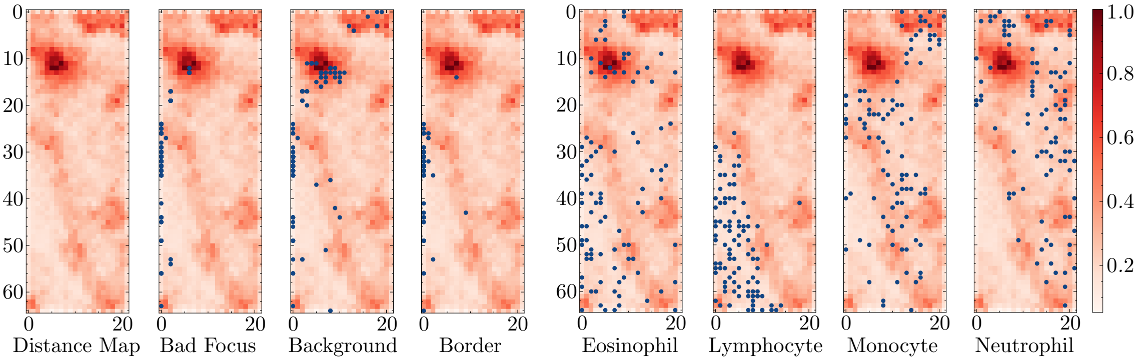
RQ 1. *Can existing machine learning methods be adapted to ensure stable and transparent processing of holographic cell images?*

Drawing insights from the results of the core publications [I], [II], and [III], the discussion moves on to the strategies used to adapt machine learning methods. Note that this analysis extends to other core and related publications to provide a comprehensive understanding of the topic.

Core Publication [I] Outlier Detection using Self-Organizing Maps for Automated Blood Cell Analysis.[*] This core publication focuses on quality control in the context of holographic blood cell imaging. Despite careful sample preparation and microfluidic focusing, as detailed in Section 2.1, inherent biological and hardware limitations remain. To ensure stable data processing, the contribution underscores the need to implement effective quality assurance and outlier detection methods. Failure to do so could have a significant impact on neural networks, introducing inconsistencies during training or disrupting the alignment of other regressors due to high-leverage features [54, 201, 285, 298]. In particular, the publication presents an innovative approach to outlier detection that uses the established morphological features [247, 332, 333] dynamically and data-driven through SOMs, as opposed to traditional hand-crafted filter rules [332, 333].

SOMs, initially proposed by Kohonen in 1981 [166], are a form of unsupervised artificial neural network capable of dimensionality reduction and clustering based on data similarity.

[*] **Author Contributions:** I initiated the research idea together with Alice Hein. I was the lead author of the manuscript and had substantial responsibility for dataset curation, feature engineering, experiment planning, and validation. I was also responsible for all figures, layout, and revisions based on reviewer feedback. I did all the presentation and defense efforts for the paper. Lucie Huang contributed to the technical implementation. Equal contribution is to be understood here as an appreciation of work outside of science, e.g. in technical or craft activities. The main share of scientific work (> 50%) lies with the author of this dissertation.



(a) Different groups of outliers and defects **(b)** The four inlier classes of leukocytes

Figure 4.1: The SOM algorithm places outliers in sparsely populated areas of the lattice (dark), while inliers are placed in densely populated areas (bright). This also reveals clusters in an unsupervised manner.

Trained on input data points, SOMs yield a lattice of neurons that preserves the distribution of the input data in a topologically meaningful way. The dataset used in this publication consists of three subsets of phase images of leukocytes: an unfiltered dataset of 447,541 images, an inlier dataset of 82,056 images, and an outlier dataset of 10,136 images.

The performance of the SOM is evaluated on both outlier and unfiltered datasets using the average quantization error as the metric for outlier detection. The evaluation identifies samples with quantization errors that exceed a threshold derived from two standard deviations of all quantization errors. The SOM demonstrates remarkable effectiveness in detecting outliers within the dataset, achieving an accuracy rate of 99.6%. Visualization of the SOM's ability to distinguish inliers from outliers by quantization error ranges reveals irregular shapes for the detected outliers (see Appendix I). Further analysis of the SOM's distance map in Figure 4.1 confirms its ability to identify dense clusters of inliers in the lighter regions and locate outliers in less dense and, therefore, darker regions or the lattice border.

This approach offers distinct advantages, allowing not only the detection of outliers but also the formation of clusters corresponding to specific types of defects or unwanted objects (compare Figure 4.1a). Because it is unsupervised, this method is capable of grouping and detecting previously unknown types of outliers. For example, identifying relevant cell aggregates for evaluating COVID-19 [III, 4, 235] can be facilitated. Automatic clustering of leukocytes into subclasses, as shown in Figure 4.1b, provides a quick overview of their statistical distribution during quality assurance, allowing early bias detection. It is important to note that although this method is useful for quality control, it does not replace highly specialized analysis by a classifier. In summary, the core publication confirms the suitability of SOMs for outlier detection in holographic blood cell images, demonstrating high accuracy on a test set of outliers. This SOM-based method represents a more generalizable and robust approach compared to current manual filtering methods while maintaining transparency in its design.

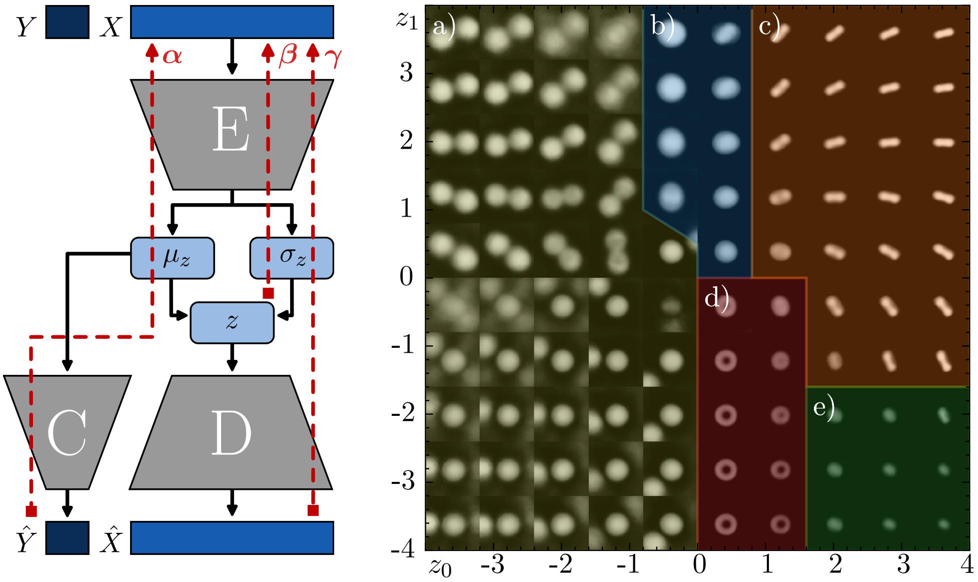
Core Publication [II] Explainable Feature Learning with Variational Autoencoders for Holographic Image Analysis.[*] Motivated by the goal of improving communication about cellular structures at a more human level [40], this core publication focuses on identifying abstract features for describing leukocytes. The aim is to foster interdisciplinary dialogue through high-level features that overcome the limitations of low-level morphological features, which often fail to robustly and meaningfully convey cell characteristics to classifiers. Regulatory constraints also request the return of quality control to human oversight [128], a challenge mitigated by the higher level of abstraction. In addition, the need for interpretability through visual representation across a broad interdisciplinary research spectrum, which SOMs do not fully address, is highlighted.

This core publication presents a modified Variational Autoencoder (VAE) tailored for explainable feature learning in holographic image analysis. This classifying VAE architecture, drawn in Figure 4.2a, aims to make quantitative phase representations more transparent and interpretable, facilitating communication about cellular events, leukocyte classification, and outlier detection. Training and test datasets include *whole blood* samples, isolated leukocytes, and defocused cells.

The resulting latent space from the classifying VAE rendered in Figure 4.2b serves as an intuitive map, allowing researchers to pre-filter cells based on specific characteristics. For focus detection, the latent space in Appendix II shows linear separability between well-focused and defocused cells, with an accuracy of approximately 96%. Discrimination between erythrocytes and leukocytes in *whole blood* samples registers an accuracy of roughly 97%. A *four-part differential* analysis of leukocyte fractions achieves a classification accuracy of 74%. In particular, the potential to improve accuracy by allowing for a high-dimensional latent space is recognized but balanced against the priority of maintaining interpretability and human accessibility of a 2D representation.

The proposed approach provides a concise overview of large datasets, streamlining quality assurance and data cleaning through an intuitive and visual interface. While not claiming perfect accuracy, the method offers a practical tool for gaining visible and understandable insights into holographic image data. The generative capabilities of this approach lay the groundwork for an eye-level exchange in interdisciplinary research, allowing discussion of cells in their natural appearance rather than relying on low-level morphological approximations. Despite potential inaccuracies, the method serves as a practical resource for newcomers to cell analysis, providing a transparent view of the general behavior of the underlying techniques. Overall, the interpretability and visual accessibility of the classifying VAE help make holographic image analysis more accessible and insightful to researchers across disciplines.

[*] **Author Contributions:** I developed the research idea and was the leading author of the manuscript. I acquired and curated the dataset and designed, executed, and validated the experiments. I was also responsible for all figures, layout, and revisions based on reviewer feedback. I did all the presentation and defense efforts for the paper. Lukas Bernhard contributed the Python library. Equal contribution is to be understood here as an appreciation of work outside of science, e.g. in technical or craft activities. The main share of scientific work (> 50%) lies with the author of this dissertation.



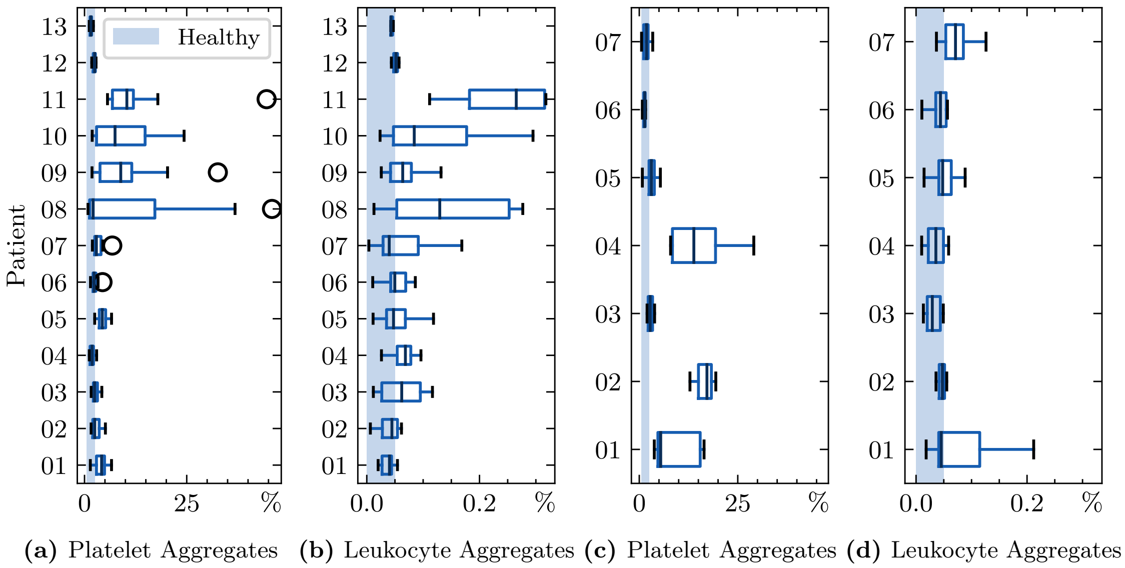
(a) Classifying VAE Architecture **(b)** The latent space offers intuitive means of selection

Figure 4.2: The architecture of a classifying VAE **(a)** combines the *classification error* ($\boldsymbol{\alpha}$), the *Kullback-Leibler Divergence* ($\boldsymbol{\beta}$), and the *reconstruction error* ($\boldsymbol{\gamma}$). The classifier **C** forces the encoder **E** to enhance the separation of the individual classes in the latent space. The areas contain a) aggregates, b) leukocytes, c) tilted erythrocytes, d) plain erythrocytes, and e) platelets.

Core Publication [III] Composition Counts: A Machine Learning View on Immunothrombosis using Quantitative Phase Imaging.[*] This core publication introduces a novel processing pipeline for the detection and quantitative analysis of blood cell aggregates, specifically focusing on platelet and leukocyte-platelet aggregates. Thrombotic events, often triggered by inflammatory conditions like sepsis and COVID-19, involve a close relationship between inflammation and hemostasis, known as immunothrombosis. The publication investigates formations of platelet and leukocyte-platelet aggregates as potential predictive biomarkers for risk assessment for COVID-19 and sepsis propagation. However, the complex analysis of these aggregates requires mastery of several aspects, including instance segmentation of the aggregates, reliable classification, and counting of individual components. Consequently, the application lends itself to a thorough comparison of data processing techniques from both classical computer vision and machine learning domains.

The performance of the proposed pipeline is systematically evaluated in various test cases, demonstrating its robustness even under challenging conditions, gradually getting closer to real-world scenarios. Alongside machine learning methods such as *U-Net* and *Mask R-CNN*, the pipeline also incorporates transparent methods such as *Watershed* segmentation and RF. In particular, the *Mask R-CNN* approach proves to be the most effective for the detection, segmentation, and classification of cell aggregates. When comparing the performance of opaque models to algorithmically transparent approaches, a clear difference becomes apparent. Classical approaches, such as combinations of traditional *Watershed* with manual

[*] **Author Contributions:** David Fresacher and I share 50% of the scientific contribution. Together with Christian Klenk, we conceived the project. I was the main author of the manuscript, while David Fresacher contributed the machine learning models. Christian Klenk and Johanna Erber collected the medical data. David and I were equally responsible for all figures, layout, and revisions based on reviewer feedback.



(a) Platelet Aggregates **(b)** Leukocyte Aggregates **(c)** Platelet Aggregates **(d)** Leukocyte Aggregates

Figure 4.3: To determine the severity of COVID-19 **(a)** & **(b)** or sepsis **(c)** & **(d)**, microthrombotic events serve as a predictive biomarker. A high proportion of platelets aggregated with other platelets or leukocytes indicates a more severe course of the disease.

gating or *Watershed* with RF, achieve an F_1-score of 0.78 even by incorporating expert knowledge. These classical approaches can identify the type of microthrombotic event observed in 73% of cases. In contrast, opaque end-to-end approaches using *U-Net* or *Mask R-CNN* architectures achieve higher F_1-scores, up to 0.91, and demonstrate an improved ability to correctly classify 97% of the events detected. This direct comparison indicates some sort of trade-off between *Predictive Accuracy* and *Descriptive Accuracy*.

Experimental results underscore the pipeline's effectiveness in detecting and analyzing blood cell aggregates in clinical samples, with a particular focus on activated platelets, platelets spiked into *whole blood*, and activated *whole blood*. While the method proves reliable in identifying platelet aggregates, it has certain limitations in detecting leukocyte-platelet aggregates. The study then applies the proposed method to clinical samples from patients with sepsis and COVID-19, revealing elevated levels of platelet and leukocyte-platelet aggregates in the majority of patients, suggesting the potential utility of these aggregates as biomarkers for immunothrombotic events and their related diseases. Especially patients with subsequent severe progression showed an extremely high proportion of aggregation, as seen in Figure 4.3. The constructed pipeline offers significant advantages over existing methods [235], as it can be used to break down not only the size of aggregates but also their composition. As shown in a derived study [D], aggregate composition is a valuable predictor of disease progression.

Acknowledging limitations, the publication emphasizes the need for a larger clinical trial to increase statistical power and validate the diagnostic potential of the proposed biomarkers. Factors such as the short lifespan of blood cell aggregates and the choice of anticoagulant are recognized as areas requiring further investigation [E].

Summary. Based on core publication [I], it is clear that implementing effective quality assurance and outlier detection methods is critical to ensuring stable data processing. SOMs show remarkable effectiveness in identifying outliers within the data set. In addition, SOMs offer advantages such as automatically clustering outliers into subclasses, fostering the detection of previously unknown types of outliers. This method provides a more generalizable and robust approach compared to traditional manual filtering methods while maintaining transparency in its design. Publication [II] introduces a modified VAE tailored for explainable feature learning. This approach aims to make quantitative phase representations more transparent and interpretable, facilitating communication about cellular events, leukocyte classification, and outlier detection. The resulting latent space from the classifying VAE serves as an intuitive map, allowing researchers to pre-filter cells based on specific appearances. While not claiming perfect accuracy, this method provides practical tools for gaining visible and understandable insights into holographic image data, making it more accessible and insightful to researchers across disciplines. Finally, publication [III] integrates and compares techniques into a novel processing pipeline for detecting and quantitatively analyzing blood cell aggregates. The pipeline incorporates both machine learning and transparent methods, demonstrating robustness under challenging conditions. It reveals potential utility as biomarkers for immunothrombotic events and related diseases, fostering clinical trials [D]. Further successful machine learning models and their performance can be found in core publication [IV]. In conclusion, existing machine learning methods can be effectively adapted to ensure stable and transparent processing of holographic cell images. Techniques such as SOMs, modified VAE, and novel processing pipelines offer promising solutions for quality assurance, outlier detection, and biomarker identification in holographic cell imaging. These publications are notable for their extensive analysis of nearly half a million cells, a significantly larger dataset compared to many similar studies that typically examine only a few hundreds [23, 148, 229, 245, 275, 278] or thousands [44, 46, 73, 226, 237] of cells. Achieving statistical significance on such a large scale is rare in comparable studies [235]. However, further research and validation, particularly in clinical settings, are needed to realize the full potential of the presented approaches.

4.2 Trustworthy Machine Learning in Interdisciplinary Research

Having confirmed the technical stability and suitability of specific learning methods for holographic cell image analysis, the focus shifts to investigating and potentially improving their **trustworthiness**. Previous work has highlighted a trade-off between *Predictive* and *Descriptive Accuracy*, demonstrating the superhuman performance of black-box models on quantitative phase images. Consequently, this section addresses the second research question and provides insights based on core publications [IV] and [V].

RQ 2. *What are the essential criteria for establishing trustworthiness in machine learning pipelines designed for holographic cytology?*

To answer this question, this work compiles criteria from various literature sources and integrates them into a comprehensive model. Given the extensive nature of the criteria listed in Section 3.3.3, and with a primary focus on establishing a new platform technology in biomedical research, the next steps require an evaluation and selection of modifiers. These latent dimensions of interpretability weigh the influence of explanation methods aiming for transparency. The domain is clearly determined by the research areas involved. These are data science, biomedical engineering, biology, and medicine. Stakeholders are primarily considered in their research activities, including the roles of practitioners and potential end-users. While regulatory or patient roles are currently absent, these might come into play the closer the technology reaches the point-of-care. Areas of application include hematooncology, tumor and other tissues, as well as immunothrombosis. Explanatory methods mainly include graphical visualization and statistical analysis, as further described in the succeeding core publications. Notably, at this stage in the development of the research tool, there is no specified time limit for the interpretation of an explanation.

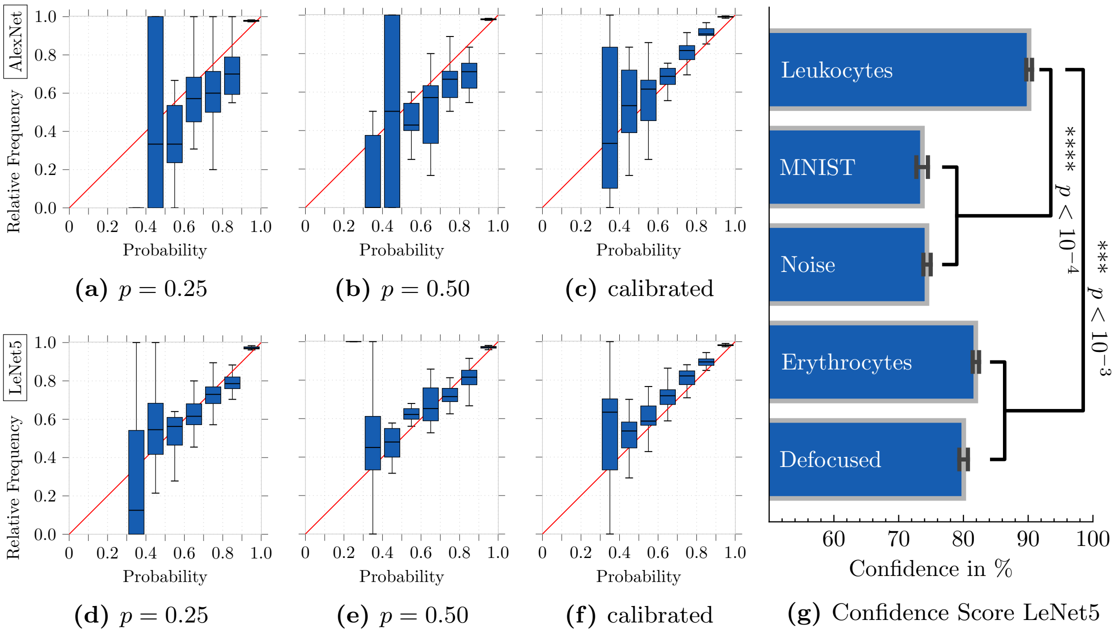
Core Publication [IV] Towards Interpretable Classification of Leukocytes based on Deep Learning.[*] In pursuit of trustworthy AI-driven systems that effectively communicate their limitations, this core publication addresses uncertainty calibration and communication in the context of holographic cytology. Focusing on interpretability, the work uses local visual explanations to uncover patterns used by different neural networks in cell recognition, as this ability is denied to the human eye in the QPI representation. The primary goal is to demonstrate that neural networks can learn to identify cell features in a manner similar to human biologists examining a blood smear, thereby establishing trust and making algorithmic general behavior predictable despite its black-box nature.

The core publication explicitly investigates the interpretability of deep learning models for leukocyte classification, comparing two relatively small architectures, *AlexNet* [170] and *LeNet5* [180]. Thereby, it emphasizes confidence estimation and visual explanation techniques. *Variational inference* is employed and compared to a *frequentist* approach for confidence calibration, with no detrimental effect on *Predictive Accuracy*. Contradictingly, *variational inference* even increases the robustness and precision of the networks. As Figure 4.4 reveals, *temperature scaling* proves an effective method for recalibrating confidence estimation, reducing overconfidence in slightly larger models such as *AlexNet*. With a robust classification accuracy of 96%, these values serve as quality measures for certification when integrated into a biomedical assay.

The investigation employs several visual explanation techniques to comprehend the cell properties that guide the network statements. These include LIME, *Guided Back-Propagation*, *Occlusion*, and *Gradient-weighted Class Activation Mapping*. *Meta-Aggregations* of these explanations are used to uncover the detection strategies employed by the networks. The distinct explanation clusters can be seen in Figure 4.5. Clear patterns emerge for the respective leukocyte classes but deviate from the biological characteristics described in textbooks (compare section 2.1). Although the general behavior of the algorithms can be assessed, these findings do not seamlessly bridge the gap to the biological domain as perceived by humans.

The trustworthiness of the networks is demonstrated not only by reliable leukocyte classification but also by robustness to outliers such as defocused data or unknown objects. The bar plot 4.4g shows that the networks' confidence decreases in the presence of outliers, allowing their specific detection and exclusion. Remarkably, leukocytes are primarily classified with such high confidence, prompting a reconsideration of the validity of the ground truth. The networks highlight instances of probably mislabeled cells, as shown in Figure 11 in the Appendix IV. That underscores the ability of the networks to provide highly accurate classifications by learning to perceive cells as they truly are.

[*] **Author Contributions:** I initiated this project and generated the data sets. I was the lead author of the manuscript and designed, performed, and validated the experiments. I was also responsible for all figures, layout, and revisions based on reviewer feedback. I did all the presentation and defense efforts for the paper. Johannes Groll implemented the machine learning models. Equal contribution is to be understood here as an appreciation of work outside of science, e.g. in technical or craft activities. The main share of scientific work (> 50%) lies with the author of this dissertation.



(a) $p = 0.25$ **(b)** $p = 0.50$ **(c)** calibrated

(d) $p = 0.25$ **(e)** $p = 0.50$ **(f)** calibrated **(g)** Confidence Score LeNet5

Figure 4.4: Reliability diagrams [65] **(a)-(f)** show the degree to which a model deviates from optimal calibration (red diagonal). Different dropout values p or temperature scaling improve the confidence calibration. A well-calibrated *LeNet5* shows significantly high confidence **(g)** for leukocytes compared to other cell types or outliers.

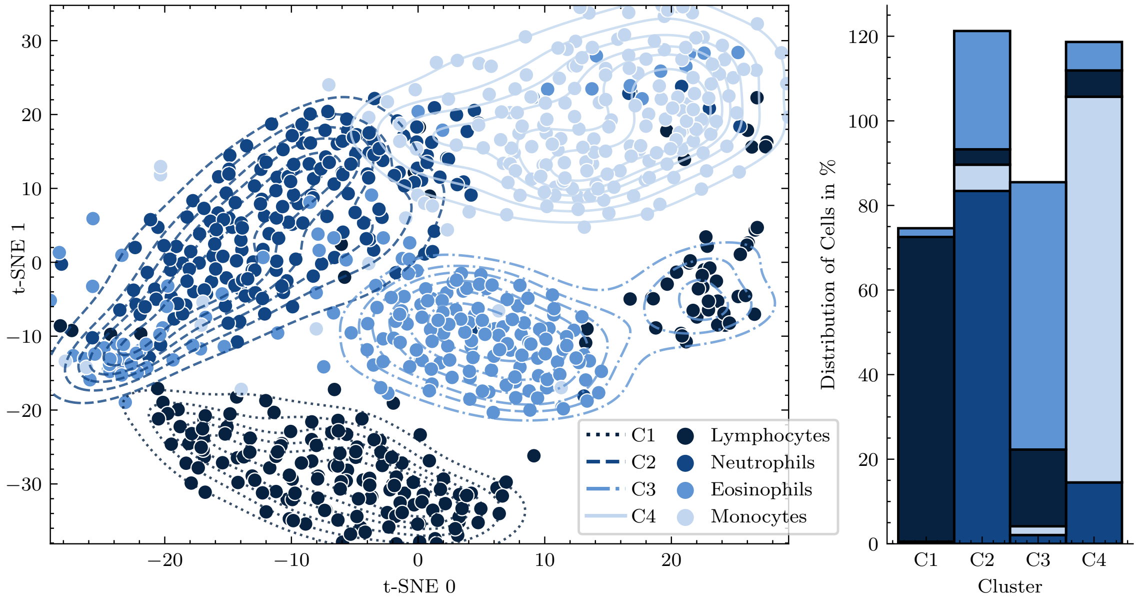


Figure 4.5: Clustering of *t-SNE* embedded LIME explanations from *AlexNet* using a *k-means* algorithm. The clusters could not match exactly all four classes. Therefore, the cumulative sum for some clusters is higher than 100% in the right bar plot.

Core Publication [V] Explainable Artificial Intelligence for Cytological Image Analysis.[*] Equipped with calibrated models and corresponding visualizations, the subsequent core publication introduces a prototype XAI dashboard. This web interface, composed of several microservices, provides a modern platform for interacting with trained models and their results, as presented in the previous section. Inspired by related work [67], the dashboard incorporates different explanation methods to accommodate the preferences of individual test subjects. Design adaptations are also implemented to align with familiar concepts in biomedical research. The elements of the web interface will be evaluated through a user study involving participants from all relevant domains to assess contributions to trustworthiness.

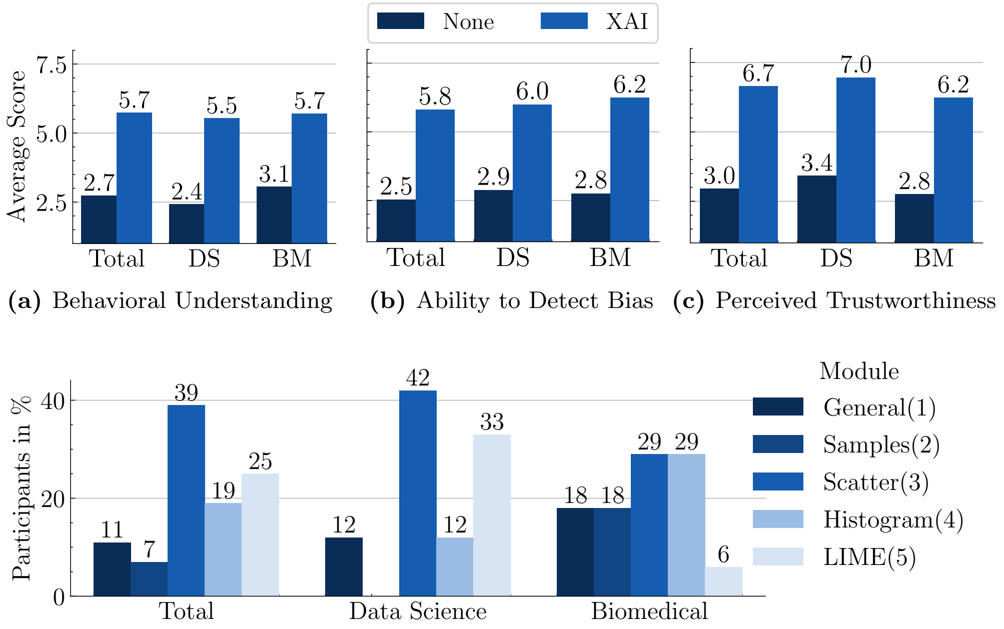
The developed prototype includes the following modules:

- **General Training and Validation Information** (Module 1): Provides background information on the algorithm used, its performance on a validation dataset, and a summary of the training dataset.
- **Samples of Classified Cells** (Module 2): Displays cell samples from the prediction results, allowing visual inspection of classified cells based on phase images.
- **Morphological Features in a Scatter Plot** (Module 3): Presents a scatter plot of individual cells, taking into account morphological features for an overall analysis of the result.
- **Morphological Feature Distribution Histogram** (Module 4): Visualizes the numerical distribution of features grouped by the individual cell classes.
- **Revealing Relevant Areas of an Image using LIME** (Module 5): Uses the LIME library to reveal relevant parts of the image to the neural networks, helping to understand the model's decision-making.

The interactive XAI dashboard is assessed using *Human-grounded Evaluation*, which measures the impact of different XAI methods on user perception and judgment in a slightly adapted *binary forced-choice* scenario [69]. Evaluation criteria include the modules' contribution to **behavioral understanding** of the algorithm, their **ability to detect bias**, and their perceived **trustworthiness**. The study involves two user groups, data scientists and biomedical researchers, with a total of 57 participants.

Compared to an unexplained performance report (Module 1), the results in Figure 4.6 show a remarkable improvement in understanding, bias detection, and trustworthiness when using the XAI dashboard. A combination of XAI modules proves to be more effective than individual modules. However, users tend to overestimate the trustworthiness of the algorithm compared to their perceived understanding of its behavior and bias detection (compare Figure 4.7a and 4.7b with 4.7c). Furthermore, the bar plot 4.6d demonstrates that certain modules appeal to specific user groups. Data scientists appreciate the LIME module, while biomedical researchers rate it as their least favorite. The image examples, on the other hand, appeal exclusively to biomedical personnel. Only the scatterplot module, inspired by established cytology tools, emerges as a rather generally accepted explanation.

[*] **Author Contributions:** I came up with the research idea and provided the revised machine learning models and data. I also planned and organized the study and the cohorts. I was the lead author of the manuscript. I was also responsible for all figures, layout, and revisions based on reviewer feedback. I did all the presentation and defense efforts for the paper. Hendrik Maier implemented the dashboard prototype and conducted user testing.



(a) Behavioral Understanding **(b)** Ability to Detect Bias **(c)** Perceived Trustworthiness

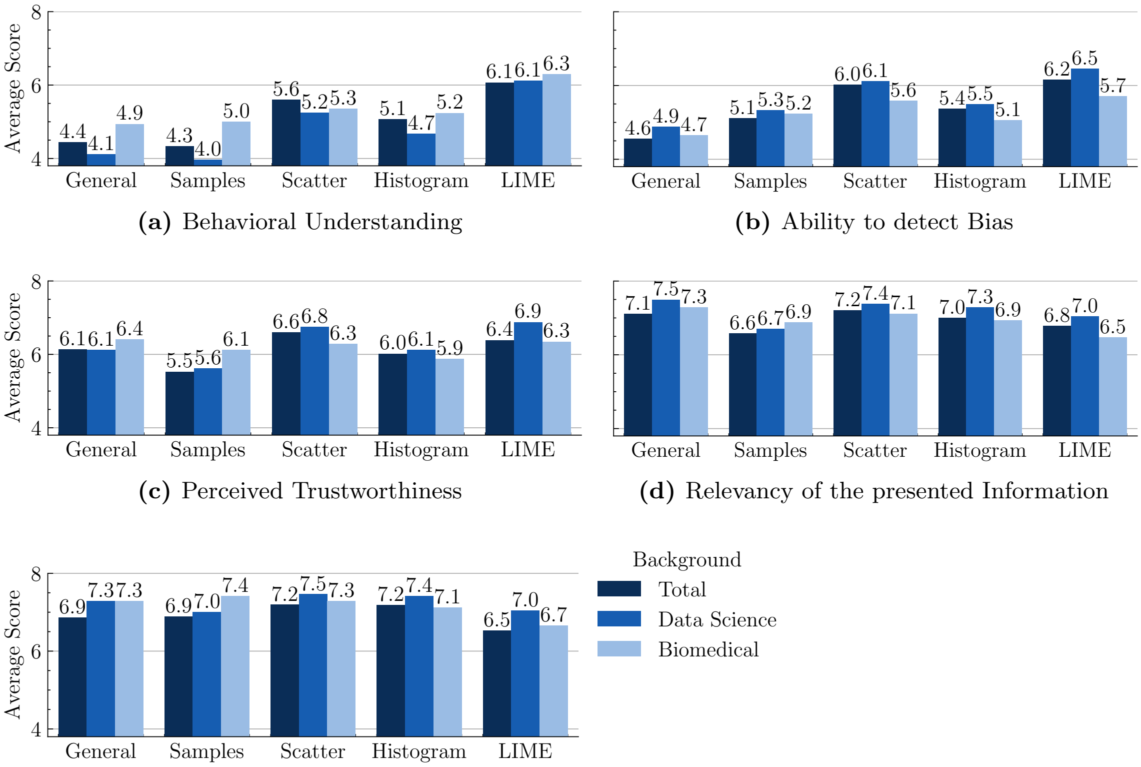
(d) Participants were asked to pick their favorite XAI module.

Figure 4.6: Participants benefit from the XAI dashboard regardless of their background in biomedical or data science. However, their distinct preferences **(d)** show the importance of providing a multimodal portfolio. (Legend: data science = DS, biomedical = BM)

These findings underscore the need for domain-specific explanations and diverse approaches to foster collaborative interdisciplinary research.

The publication concludes that while the XAI dashboard improves interpretability and confidence in machine learning models, there is still room for improvement in users' understanding of algorithm behavior and bias detection. Transparency about model accuracy and limitations is emphasized to avoid user misinterpretation. The need for explainability in machine learning remains high, especially in interdisciplinary research and *clinical decision support systems*.

Summary. Core publication [IV] highlights the criteria of interpretability and uncertainty communication. Techniques such as *variational inference* and visual explanations improve the understanding and predictability of deep learning models, especially in leukocyte classification. The calibrated models show resilience to outliers and unknown objects, ensuring reliable performance essential for accurate cell type identification. The subsequent publication [V] uses these optimized models to investigate further the perceived trustworthiness conveyed by various explanation techniques. It also concentrates on the latent dimensions of interpretability, which modify the relevance and quality of the explanations. An XAI dashboard tailored to the preferences of users, such as data scientists and biomedical researchers, provides intuitive insights into model decisions, building trust and understanding. Evaluating this dashboard with a diverse user group emphasizes the importance of domain-specific explanations and collaborative interdisciplinary research. However, this also implies that the research question cannot be answered universally but must be individually re-evaluated in new situations. Nevertheless, distilling the large number of explanation metrics and latent dimensions to the biomedical use case provides better support for further investigation. In summary, the essential criteria for establishing



(a) Behavioral Understanding

(b) Ability to detect Bias

(c) Perceived Trustworthiness

(d) Relevancy of the presented Information

(e) Comprehension of the presented Information

Figure 4.7: The participants rated the XAI dashboard modules individually. Most of them found the explanations presented to be highly relevant **(d)** and comprehensible **(e)**. However, there is a risk of over-trusting the algorithms **(c)**, as the participants experienced only moderate improvements in behavioral **(a)** or bias detection **(b)** insights.

trustworthiness in machine learning pipelines for biomedical research applications include interpretability, confidence calibration, robustness to outliers, bias detection, user-centered interface design, and human-based evaluation. By addressing these criteria, machine learning pipelines can provide transparent and reliable results, facilitating their adoption in biomedical research and clinical decision-making.

4.3 Design Rules for AI-driven Biomedical Interfaces

The third and final translation criterion examined is **usability**, which is often the determining factor in the adaptability of a tool or its rejection by users due to perceived difficulty. The user interface is of central importance and must meet several requirements. Design considerations extend beyond visual aesthetics to include temporal and logical processes within the program [185]. Excessive ambiguity hinders proper program use, leading to user frustration, reduced program utility, and potential impact on trustworthiness [234].

A recommended approach to improve the quality of the user interface and ensure optimal usability is user-centered design [125, 129, 238, 290]. This methodology aims to minimize the gap between the offered solution and the practical usability of the program. Achieving this involves iterative improvements to an initial user interface prototype based on user feedback obtained through *formative evaluation* [118], as discussed in Section 3.4.2. However, due to the challenges posed by the specific target audience and the limited scope of the application, conducting user feedback is resource-intensive. Therefore, a more practical approach is to perform *formative evaluation* promptly using design rules and reserve human user involvement for the *summative assessment* [118] at the end of the development cycle.

Design rules are an early warning system and guardrail, facilitating user-centered development despite indirect user involvement. Hence, the subsequent core publication addresses the third research question:

RQ 3. *What design rules are suitable for improving the usability of AI-driven user interfaces for holographic cytology?*

Core Publication [VI] Rethinking Usability Heuristics for Modern Biomedical Interfaces.[*] The focus of this core publication is to improve AI-driven interface usability in biomedical research, specifically in blood cell segmentation and classification. The publication compares three sets of usability heuristics rule sets: **Nielsen's general usability heuristics** [233], **guidelines for human-AI interaction** [6], and a newly developed set tailored to **biomedical AI** interfaces. The assessment of these UEMs [118] includes expert reviews and user testing conducted on a prototype interface designed for blood cell analysis.

To significantly influence an appropriate design process of a biomedical tool and to witness the application of design rules, an independent software prototype is actively developed. This prototype addresses the lack of ground truth data by using an active learning approach [C, 132, 296] that minimizes human effort to improve usability. The software prototype guides the user through different views to control the program and its functions, providing both generic and direct interaction with the AI technology. The prototype facilitates a seamless workflow for segmenting quantitative phase images and classifying cells via a web interface. Using a human-in-the-loop process [69, 132], the software learns from human intervention, corrects errors, and evolves with minimal initial training data.

The characteristics of the biomedical application and the needs of the user group were determined through extensive interviews and a literature review. Based on these insights, the publication compiles a catalog of 15 heuristic design rules critical to the development of AI-driven biomedical user interfaces, with detailed sources listed in Table 4.1.

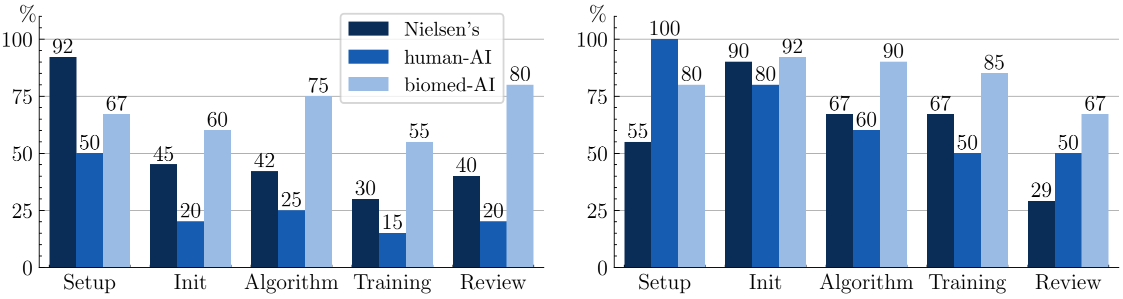
Evaluation of the three UEMs shows that while Nielsen's general heuristics excel at identifying usability problems in AI-light areas, they struggle in AI-heavy areas influenced by machine learning. Figure 4.8 visualizes the decreasing values for *Validity* and *Thoroughness* of the predicted usability problems for AI-driven parts of the user interface. The human-AI interaction guidelines face challenges in this specific domain, with generally low *Validity* and *Thoroughness* scores compared to the other rule sets. Conversely, the newly developed biomedical AI heuristics perform well, especially in domains that enforce human-AI interaction. As shown in Figure 4.9, every rule was able to identify usability problems, highlighting the successful adaptation to the biomedical field.

The publication confirms the importance of domain-specific heuristics in biomedical interfaces due to the unique challenges. It underlines the effectiveness of the developed biomedical AI heuristics in detecting critical usability issues. The intention is to apply

[*] **Author Contributions:** I initiated the scientific base, the development of the prototype and provided the data sets. I planned the interviews and supervised the user and expert study. I was the lead author of the manuscript and was also responsible for all figures, layout, and revisions based on reviewer feedback. I did all the presentation and defense efforts for the paper. Christian Janotte implemented the prototype, conducted the interviews, and performed the user testing.

Table 4.1: Heuristic Rules for Embedding AI in Biomedical Research Tools

	#	Name	Short Description
Structure	1	Streamline main task	Focus on the main task that a system was created for and make the system easy to learn [185].
	2	Provide full control	Provide global control of important model parameters and the data pipeline [5][95].
	3	Orientation	Always show users where they are, what is currently going on and what they can do next [234].
Interaction	4	Guide attention	Keep the users focused on their task and only alarm them in urgent cases [138][318].
	5	Provide comparisons	Let users compare among similar data or parameters when they need to judge an outcome or make a decision.
	6	Show impact	Users need to see how their actions influence the system and its performance [141].
	7	User over System	Allow users to correct errors of the AI efficiently at all times and even turn off the AI if needed [138].
Presentation	8	Familiar language	Use non-technical language if possible. Pay attention to use correct terminology for medical concepts [281].
	9	Precise language	Avoid ambiguous wording for labels and commands that could trigger confusion [234].
	10	Familiar look	Use ways of presentation for the interface that users know from other tools.
	11	Appeal	Give the users the feeling of using a state-of-the-art and high-quality product.
Explainability	12	Explain data	Foster the interpretability of the data and how it differs from other data sources [64].
	13	Explain processing	There needs to be a high-level explanation for the overall procedure that is performed by the system [133].
	14	Explain reasoning	There has to be an explanation why and how the system derived a certain result or prediction [133].
	15	Strengths/Limitations	Show what the strengths and weaknesses of the system are and what expectations are realistic [128].



(a) Thoroughness

(b) Validity

Figure 4.8: Quality assessment of the heuristic rule sets on the individual views of the prototype compared to a user test. While the general heuristics perform well on more generic views, the AI-driven views are better handled by the biomedical AI heuristics.

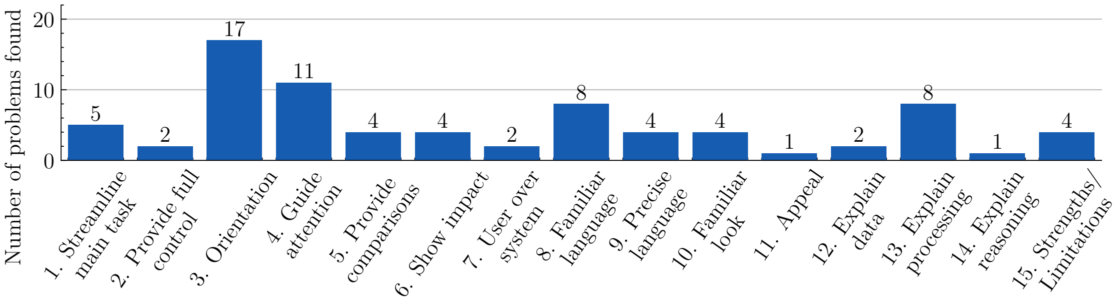


Figure 4.9: Potential usability problems detected by the newly developed biomedical AI heuristics

the developed design rules to other biomedical interfaces to further advance AI-based technologies in research and healthcare.

Summary. The results of this publication provide valuable insights into design rules suitable for improving the usability of AI-driven user interfaces in modern and interdisciplinary cytology. One of the main contributions is an intuitive tool for the segmentation and classification of cells, which can be optimally adapted to the needs in this domain through interactive and user-centered design. Due to resource constraints, formative evaluation using heuristic design rules is a practical approach to improving usability early in the development cycle. Design rules act as guardrails, facilitating user-centered development despite limited direct user involvement. The outstanding contribution is a set of 15 heuristic design rules tailored for biomedical AI-infused interfaces, which are critical for improving the usability of blood cell segmentation and classification tasks. The validity of these rules is underlined by their comparison to existing usability heuristics, highlighting the effectiveness of domain-specific biomedical AI heuristics in detecting critical usability issues. While general usability heuristics perform well in AI-light domains, they struggle in AI-heavy domains influenced by machine learning.

Chapter 5

Discussion

This research explores the dynamic interplay between data-driven machine learning and the field of cytology within the medical-biological domain. By examining three key criteria – **usefulness**, **trustworthiness**, and **usability** – this work aims to facilitate the successful translation of these advanced technologies. The focus is on illuminating significant influences in each criterion, leveraging the powerful yet complex nature of artificial intelligence. To this end, the work demonstrates a viable path for transforming the emerging platform technology of quantitative phase microscopy into a true innovation through synergy with interpretable machine learning methods. However, achieving this synergy requires extensive and further in-depth research and studies. The following sections highlight the ongoing research gaps and prospects.

5.1 Limitations

The Usefulness of a Fast-Paced Technology in a High-Risk Sector. Machine learning models are evolving rapidly, creating challenges for their effective use in clinical settings. Publications praise remarkable accuracy and speed in image processing, often exceeding the practical skills of pathologists and lab technicians. However, they also exceed the level of human comprehension. The robust strengths of machine learning must be reliably harnessed; otherwise, these achievements will remain aspirational. The future features *Machine Learning Operations* [323] prominently, underscoring the critical role of curating AI models for biomedical applications. Issues related to the curation, lifecycle, and real-world applicability of these models require more extensive and large-scale experimentation beyond the scope of the current work [154].

Addressing these concerns will require diversifying datasets with a broader range of donors, cell types, and hardware setups to represent real-world conditions at the point-of-care accurately. Transparency is paramount, demanding a clear exposition of potential biases in learning methods, ideally directly traceable to human biologist approaches. Integrating the ever-growing array of data-driven learning architectures into this endeavor remains a challenge, especially given the current focus on comparatively small architectures presented in this work. Furthermore, the search for additional applications of digital holography in conjunction with machine learning methods must continue, underlining the exceptional versatility of these technologies as their major selling point.

Given the undeniable potential of this technology, it is crucial to establish binding data standards and calibration procedures for AI-driven cytology [54, 59, 285, 298, 310]. Stable adoption depends on smooth integration into clinical infrastructure [105, 154, 305, 318], facilitating the construction of *clinical decision support systems* with the presented technology. The clinical relevance of AI-based quantitative phase imaging can be increased only by seamless integration in biomedical workflows and validation by multicenter studies [209].

Demystifying Trustworthy AI. Various results show that trustworthy machine learning can be achieved and is not destined to remain a myth [186]. When provided with interpretable explanations, users experience a significant increase in trust in machine learning models. However, for black-box models, there is a persistent gap between the complexity of these powerful models and their simplified explanation techniques [105, 299]. The question arises: Can we bridge this knowledge gap, similar to how we handle other technologies in our daily lives, without a complete understanding of their inner workings? While this work explores specific explanation approaches, there is room to explore newer and potentially more effective visual explanations to improve model interpretability [18, 40, 129, 307, 325].

Limitations of the work include the need for a broader range of methods to measure interpretability, with consideration of metrics not yet applied. Clinical applicability, particularly in a point-of-care setting, remains to be established. Moving on to an *Application-grounded Evaluation* [69] of interpretability, paradigms governing the tool under study may evolve as it transitions to a clinical context.

Additionally, this work underscores the importance of building a foundation in a research environment before venturing into real-world applications and avoids immediate experimentation in clinical settings. To accurately assess the trustworthiness of algorithms, further studies under controlled conditions with larger subject samples and control groups are essential. The goal is to comprehensively model psychological effects in human-machine communication [50] and to understand the latent dimensions of machine learning interpretability [69].

Finally, this work highlights the importance of *AI literacy* in building a rational trust relationship with machine learning. Users tended to place excessive trust in algorithms despite uncertainty about biases or limited understanding of algorithmic functions. Subjects with domain knowledge are more skeptical of the new technology, highlighting the need to provide both general and background information in explanations [51, 67]. Ensuring that the presence of an explanation does not disproportionately increase user confidence is critical [V], and the accountability is on developers to maintain transparency and avoid misleading users [53]. Legislation and science mandate communication of system accuracy and limitations [110], raising the question of whether it is acceptable for some aspects of AI to remain undisclosed [38, 193].

Overlooking the Enabling Role of Usability. Although ergonomic aspects of software are often neglected, they critically impact user experience, with even subtle details having a significant effect. It is often assumed that a visually appealing graphical user interface is a sign of high-quality internal functionality [63, 89]. Users become impatient when calculations take a long time, and the program communicates its internal state poorly. Conversely, suspicion arises when the analysis is too fast, raising concerns about the machine's diligence with valuable data [3, 33, 232]. Usability, like machine learning and interpretability, has many facets.

The prototypes presented in this study serve as examples of biomedical tools. To effectively validate the design guidelines, a broader evaluation of numerous such tools is imperative. Diversification of the learning algorithms integrated into the user interface is necessary to comprehensively assess their impact on the biomedical domain. The design rules developed have potential applications in the design of software for various biomedical use cases, particularly in the clinical integration of phase-contrast microscopic examinations of tissues [G, 304]. Additional studies are needed to evaluate other aspects of usability, such as feasibility, efficiency, error rates, and overall user satisfaction [337], as the current study situation does not allow for conclusive statements. A more in-depth study of long-term effects is also essential [6].

This work underlines the importance of incorporating recently published guidelines for the development of biomedical user interfaces [6, 249, 299, 329]. There is an urgent need for domain-specific usability recommendations to facilitate the seamless use of data-driven and semi-autonomous technology in the biomedical research environment.

5.2 The Advent of New Platform Technologies

A Sustainable Cyclic Research Workflow. To establish a sustainable cyclical research workflow integrating machine learning into medical research, closer collaboration within the research community is essential [218, 305, 329]. While traditional linear research projects follow a planned sequence of data collection and statistical evaluation, complex systems such as deep learning applied to intricate living organisms require caution due to inherent uncertainties. Purely linear research may produce results with limited generalisability [17, 341]. Although machine learning promises fast and accurate results, it often sacrifices the essential knowledge gain that is crucial for sustainable research [14].

This work emphasizes that interdisciplinary collaboration is essential throughout the research workflow, as questions and problems can arise at any stage and any domain. Examining software tools in research offers the advantage of using human-grounded metrics without exposing the research to the risks of premature clinical application. The influence of machine learning can be systematically investigated under controlled conditions, fostering a broad understanding from multiple perspectives and eliminating the hierarchical service provider dynamic between disciplines [218].

With a significant number of studies, it becomes possible to better understand and model latent dimensions in the interdisciplinary interpretation of artificial intelligence. These insights may enable *Functionally-grounded Evaluations* of human and machine behavior and, in turn, unit test-like predictions of the success and safety of a given methodology [69]. Similarly, Kelly et al. [154] require metrics for *post-market surveillance* of AI-infused systems to permanently track their performance and alert responsible experts. Emphasizing the importance of interpretability in machine learning, explanation techniques should not be seen as mere add-ons; rather, algorithms and their application need to be redesigned for human interpretation [132, 330]. It is crucial to establish a common, understandable vocabulary [305] across all disciplines involved, which requires an extra effort to include different technical backgrounds and to understand the roles of researchers. This inclusiveness ensures transparency in the workflow and an unhindered flow of information. Finally, aligning the mental models of the people involved with the machine learning models is essential for successful integration [50].

Let AI Under Your Skin. The integration of label-free QPI microscopy with machine learning undoubtedly has transformative potential. Personalized medicine takes a significant step forward by removing the need for precise, pre-defined targets and putting data exploration in the hands of computer vision. The simplicity, adaptability, and speed of these approaches promise a new form of *in vitro* diagnostics, accessible either at the bedside or directly in the doctor's office. As a reliable early warning system and highly customizable health monitor, it could become a real revolution. Realizing this concept comes a bit closer to the vision of a universal *medical Tricorder*. However, innovators must consider the translation criteria outlined in this work.

Anticipating future application scenarios for AI and QPI provides food for thought. A fully autonomous AI takeover seems highly unlikely, given the unjustifiable risks. Instead, a *conditional control scenario* [330], where machines take over tasks suitable for safe automation, is more plausible. Human specialists will remain an essential backup and last resort for decision-making and emergency intervention. Even on the futuristic Starship Enterprise, a human doctor controls the medical equipment and makes the final decisions. Establishing open and inclusive interaction between research disciplines will be the key to fostering trust. However, gaining the trust of physicians and patients requires a leap of faith [205, 213]. Therefore, navigating the path through interdisciplinary research and software use is proving effective in building a solid foundation for the pillars of translation. Responsible use of AI can be carefully and continuously tested, ensuring its safe introduction into the clinical workflow, as evidenced by recent clinical approvals [330]. It is up to all stakeholders to take responsibility for this technology's sustainable development and the acceptance of their "skin in the game". First, AI must be allowed to look under our skin, permitting a gradual progression towards optimal use, before it gains access to our hearts.

> “*The minute you let her under your skin,*
> *Then you begin to make it better.*”
>
> Paul McCartney, *Hey Jude*, 1968

Bibliography

References

[1] M. Ackermann et al. “Pulmonary Vascular Endothelialitis, Thrombosis, and Angiogenesis in Covid-19”. In: *New England Journal of Medicine* 383.2 (2020), pp. 120–128.

[2] D. Acuna, H. Ling, A. Kar, and S. Fidler. “Efficient Interactive Annotation of Segmentation Datasets with Polygon-RNN++”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 859–868.

[3] E. Adar, D. S. Tan, and J. Teevan. “Benevolent Deception in Human Computer Interaction”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 1863–1872.

[4] A. H. Alharbi, C. V. Aravinda, M. Lin, P. S. Venugopala, P. Reddicherla, and M. A. Shah. “Segmentation and Classification of White Blood Cells Using the UNet”. In: *Contrast Media & Molecular Imaging* 2022 (2022), pp. 1–8.

[5] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza. “Power to the People: The Role of Humans in Interactive Machine Learning”. In: *AI Magazine* 35.4 (2014), pp. 105–120.

[6] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz. “Guidelines for Human-AI Interaction”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, pp. 1–13.

[7] J. P. Amorim, P. H. Abreu, A. Fernandez, M. Reyes, J. Santos, and M. H. Abreu. “Interpreting Deep Machine Learning Models: An Easy Guide for Oncologists”. In: *IEEE Reviews in Biomedical Engineering* 16 (2023), pp. 192–207.

[8] A. Arbelle and T. R. Raviv. “Microscopy Cell Segmentation via Adversarial Neural Networks”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging*. IEEE, 2018, pp. 645–648.

[9] M. Asghari, M. Serhatlioglu, B. Ortaç, M. E. Solmaz, and C. Elbuken. “Sheathless Microflow Cytometry Using Viscoelastic Fluids”. In: *Scientific Reports* 7.1 (2017), p. 12342.

[10] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation”. In: *PLoS ONE* 10.7 (2015), e0130140.

[11] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixão, F. Mutz, L. De Paula Veronese, T. Oliveira-Santos, and A. F. De Souza. “Self-Driving Cars: A Survey”. In: *Expert Systems with Applications* 165 (2021), p. 113816.

[12] S. M. Baik, K. S. Hong, and D. J. Park. “Deep Learning Approach for Early Prediction of COVID-19 Mortality Using Chest X-ray and Electronic Health Records”. In: *BMC Bioinformatics* 24.1 (2023), p. 190.

[13] B. J. Bain. *A Beginner’s Guide to Blood Cells*. 3rd Edition. John Wiley & Sons, 2017.

[14] P. Ball. “Is AI Leading to a Reproducibility Crisis in Science?” In: *Nature* 624.7990 (2023), pp. 22–25.

[15] E. P. Balogh, B. T. Miller, J. R. Ball, and The National Academies of Sciences, Engineering, and Medicine. *Improving Diagnosis in Health Care*. 1st Edition. National Academies Press, 2015.

[16] J. J. Barcia. “The Giemsa Stain: Its History and Applications”. In: *International Journal of Surgical Pathology* 15.3 (2007), pp. 292–296.

[17] G. Barlevy. “On the Cyclicality of Research and Development”. In: *American Economic Review* 97.4 (2007), pp. 1131–1164.

[18] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI”. In: *Information Fusion* 58 (2020), pp. 82–115.

[19] P. R. A. S. Bassi, S. S. J. Dertkigil, and A. Cavalli. “Improving Deep Neural Network Generalization and Robustness to Background Bias via Layer-Wise Relevance Propagation Optimization”. In: *Nature Communications* 15.1 (2024), p. 291.

[20] D. J. Beebe, G. A. Mensing, and G. M. Walker. “Physics and Applications of Microfluidics in Biology”. In: *Annual Review of Biomedical Engineering* 4.1 (2002), pp. 261–286.

[21] A. V. Belashov, D. A. Gorbenko, A. A. Zhikhoreva, T. N. Belyaeva, E. S. Kornilova, I. V. Semenova, and O. S. Vasyutinskii. “The Development of Segmentation Algorithms in Holographic Microscopy and Tomography for Determination of Morphological Parameters of Cells”. In: *Technical Physics Letters* 45.11 (2019), pp. 1140–1143.

[22] V. Belle and I. Papantonis. “Principles and Practice of Explainable Machine Learning”. In: *Frontiers in Big Data* 4 (2021), p. 688969.

[23] S. Ben Baruch, N. Rotman-Nativ, A. Baram, H. Greenspan, and N. T. Shaked. “Cancer-Cell Deep-Learning Classification by Integrating Quantitative-Phase Spatial and Temporal Fluctuations”. In: *Cells* 10.12 (2021), p. 3353.

[24] N. Bevana, J. Kirakowskib, and J. Maissela. “What Is Usability”. In: *Proceedings of the 4th International Conference on Human-Computer Interaction*. 1991, pp. 1–6.

[25] M. Bhandari, P. Yogarajah, M. S. Kavitha, and J. Condell. “Exploring the Capabilities of a Lightweight CNN Model in Accurately Identifying Renal Abnormalities: Cysts, Stones, and Tumors, Using LIME and SHAP”. In: *Applied Sciences* 13.5 (2023), p. 3125.

[26] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley. “Explainable Machine Learning in Deployment”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 2020, pp. 648–657.

[27] A. Binder et al. “Towards Computational Fluorescence Microscopy: Machine Learning-Based Integrated Prediction of Morphological and Molecular Tumor Profiles”. In: *arXiv* Preprint (2018), p. 1107.1153.

[28] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter. “Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer’s Disease Classification”. In: *Frontiers in Aging Neuroscience* 11 (2019), p. 194.

[29] L. Breiman. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32.

[30] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. “LOF: Identifying Density-Based Local Outliers”. In: *ACM SIGMOD Records* 29.2 (2000), pp. 93–104.

[31] P. L. Brockett, X. Xia, and R. A. Derrig. “Using Kohonen’s Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud”. In: *The Journal of Risk and Insurance* 65.2 (1998), p. 245.

[32] S. Budd, E. C. Robinson, and B. Kainz. “A Survey on Active Learning and Human-in-the-Loop Deep Learning for Medical Image Analysis”. In: *Medical Image Analysis* 71 (2021), p. 102062.

[33] R. W. Buell and M. I. Norton. “The Labor Illusion: How Operational Transparency Increases Perceived Value”. In: *Management Science* 57.9 (2011), pp. 1564–1579.

[34] A. Butola, D. Popova, D. K. Prasad, A. Ahmad, A. Habib, J. C. Tinguely, P. Basnet, G. Acharya, P. Senthilkumaran, D. S. Mehta, and B. S. Ahluwalia. “High Spatially Sensitive Quantitative Phase Imaging Assisted with Deep Neural Network for Classification of Human Spermatozoa under Stressed Condition”. In: *Scientific Reports* 10.1 (2020), p. 13118.

[35] S. Bystryak, R. P. Bandwar, and R. Santockyte. “A Flow-through Cell Counting Assay for Point-of-Care Enumeration of CD4 T-cells”. In: *Journal of Virological Methods* 271 (2019), p. 113672.

[36] F. Cabitza, R. Rasoini, and G. F. Gensini. “Unintended Consequences of Machine Learning in Medicine”. In: *Journal of the American Medical Association* 318.6 (2017), p. 517.

[37] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. “Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 2015, pp. 1721–1730.

[38] D. Castelvecchi. “Can We Open the Black Box of AI?” In: *Nature* 538.7623 (2016), pp. 20–23.

[39] K. Chadaga, S. Prabhu, V. Bhat, N. Sampathila, S. Umakanth, and R. Chadaga. “A Decision Support System for Diagnosis of COVID-19 from Non-COVID-19 Influenza-like Illness Using Explainable Artificial Intelligence”. In: *Bioengineering* 10.4 (2023), p. 439.

[40] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar. “A Review of Trustworthy and Explainable Artificial Intelligence (XAI)”. In: *IEEE Access* 11 (2023), pp. 78994–79015.

[41] T. Chan and L. Vese. “Active Contours without Edges”. In: *IEEE Transactions on Image Processing* 10.2 (2001), pp. 266–277.

[42] H. Chefer, S. Gur, and L. Wolf. “Transformer Interpretability Beyond Attention Visualization”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 782–791.

[43] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin. “This Looks Like That: Deep Learning for Interpretable Image Recognition”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 2019.

[44] C. L. Chen, A. Mahjoubfar, L.-C. Tai, I. K. Blaby, A. Huang, K. R. Niazi, and B. Jalali. “Deep Learning in Label-free Cell Classification”. In: *Scientific Reports* 6.1 (2016), p. 21471.

[45] H. Chen, L. Huang, T. Liu, and A. Ozcan. “Fourier Imager Network (FIN): A Deep Neural Network for Hologram Reconstruction with Superior External Generalization”. In: *Light: Science & Applications* 11.1 (2022), p. 254.

[46] H. Chen, Q. Dou, X. Wang, J. Qin, and P. Heng. “Mitosis Detection in Breast Cancer Histology Images via Deep Cascaded Networks”. In: *Proceedings of the Conference on Artificial Intelligence* 30.1 (2016).

[47] J. Chen, S. Sathe, C. Aggarwal, and D. Turaga. “Outlier Detection with Autoencoder Ensembles”. In: *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 2017, pp. 90–98.

[48] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. “A Simple Framework for Contrastive Learning of Visual Representations”. In: *Proceedings of the 37th International Conference on Machine Learning*. 2020, pp. 1597–1607.

[49] E. M. Christiansen et al. “In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images”. In: *Cell* 173.3 (2018), 792–803.e19.

[50] M. Chromik. “Human-Centric Explanation Facilities: Explainable AI for the Pragmatic Understanding of Non-Expert End Users”. PhD thesis. Ludwig-Maximilians-Universität München, 2021.

[51] M. Chromik. “Making SHAP Rap: Bridging Local and Global Insights Through Interaction and Narratives”. In: *Human-Computer Interaction*. Vol. 12933. Springer International Publishing, 2021, pp. 641–651.

[52] M. Chromik, M. Eiband, F. Buchner, A. Krüger, and A. Butz. “I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI”. In: *26th International Conference on Intelligent User Interfaces*. ACM, 2021, pp. 307–317.

[53] M. Chromik, F. Fincke, and A. Butz. “Mind the (Persuasion) Gap: Contrasting Predictions of Intelligent DSS with User Beliefs to Improve Interpretability”. In: *Companion Proceedings of the 12th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. ACM, 2020, pp. 1–6.

[54] F. Ciompi, O. Geessink, B. E. Bejnordi, G. S. De Souza, A. Baidoshvili, G. Litjens, B. Van Ginneken, I. Nagtegaal, and J. Van Der Laak. “The Importance of Stain Normalization in Colorectal Tissue Classification with Convolutional Networks”. In: *2017 IEEE 14th International Symposium on Biomedical Imaging*. IEEE, 2017, pp. 160–163.

[55] R. Conroy. “Estimation of Ten-Year Risk of Fatal Cardiovascular Disease in Europe: The SCORE Project”. In: *European Heart Journal* 24.11 (2003), pp. 987–1003.

[56] A. Cooper. “The Inmates Are Running the Asylum”. In: *Software-Ergonomie*. Vol. 53. Vieweg+Teubner Verlag, 1999, pp. 17–17.

[57] Y. Cotte, F. Toy, P. Jourdain, N. Pavillon, D. Boss, P. Magistretti, P. Marquet, and C. Depeursinge. “Marker-Free Phase Nanoscopy”. In: *Nature Photonics* 7.2 (2013), pp. 113–117.

[58] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* 13th Edition. Cambridge University Press, 2012.

[59] S. Cruz Rivera et al. “Guidelines for Clinical Trial Protocols for Interventions Involving Artificial Intelligence: The SPIRIT-AI Extension”. In: *Nature Medicine* 26.9 (2020), pp. 1351–1363.

[60] G. Dardikman-Yoffe, S. K. Mirsky, I. Barnea, and N. T. Shaked. “High-Resolution 4-D Acquisition of Freely Swimming Human Sperm Cells without Staining”. In: *Science Advances* 6.15 (2020), eaay7619.

[61] A. Datta, M. C. Tschantz, and A. Datta. “Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination”. In: *Proceedings on Privacy Enhancing Technologies* 2015.1 (2015), pp. 92–112.

[62] F. D. Davis. “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology”. In: *Management Information Systems Quarterly* 13.3 (1989), p. 319.

[63] A. De Angeli, A. Sutcliffe, and J. Hartmann. “Interaction, Usability and Aesthetics: What Influences Users’ Preferences?” In: *Proceedings of the 6th Conference on Designing Interactive Systems.* ACM, 2006, pp. 271–280.

[64] J. De Fauw et al. “Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease”. In: *Nature Medicine* 24.9 (2018), pp. 1342–1350.

[65] M. H. DeGroot and S. E. Fienberg. “The Comparison and Evaluation of Forecasters”. In: *The Statistician* 32.1/2 (1983), p. 12.

[66] P. Dharshini, R. Kumar K, S. Venkatesh, K. Narasimhan, and K. Adalarasu. “An Overview Of Interpretability Techniques For Explainable Artificial Intelligence (XAI) In Deep Learning-Based Medical Image Analysis”. In: *9th International Conference on Advanced Computing and Communication Systems.* IEEE, 2023, pp. 175–182.

[67] C. Diehl, A. Martins, A. Almeida, T. Silva, Ó. Ribeiro, G. Santinha, N. Rocha, and A. G. Silva. “Defining Recommendations to Guide User Interface Design: Multimethod Approach”. In: *Journal of Medical Internet Research Human Factors* 9.3 (2022), e37894.

[68] M. A. do R. B. F. Lima and D. Cojoc. “Monitoring Human Neutrophil Differentiation by Digital Holographic Microscopy”. In: *Frontiers in Physics* 9 (2021), p. 653353.

[69] F. Doshi-Velez and B. Kim. “Towards A Rigorous Science of Interpretable Machine Learning”. In: *arXiv* Preprint (2017), p. 1702.08608.

[70] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. “The Importance of Skip Connections in Biomedical Image Segmentation”. In: *Deep Learning and Data Labeling for Medical Applications.* Vol. 10008. Springer International Publishing, 2016, pp. 179–187.

[71] F. Dubois and C. Yourassowsky. “Off-Axis Interferometer”. US9207638B2. 2015.

[72] J. M. Durán and K. R. Jongsma. “Who Is Afraid of Black Box Algorithms? On the Epistemological and Ethical Basis of Trust in Medical AI”. In: *Journal of Medical Ethics* (2021), medethics-2020–106820.

[73] K. Eder, T. Kutscher, A. Marzi, Á. Barroso, J. Schnekenburger, and B. Kemper. "Automated Detection of Macrophages in Quantitative Phase Images by Deep Learning Using a Mask Region-Based Convolutional Neural Network". In: *Label-Free Biomedical Imaging and Sensing*. SPIE, 2021, p. 54.

[74] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, and H. Hussmann. "Bringing Transparency Design into Practice". In: *23rd International Conference on Intelligent User Interfaces*. ACM, 2018, pp. 211–223.

[75] V. Emamian, M. Kaveh, and A. Tewfik. "Robust Clustering of Acoustic Emission Signals Using the Kohonen Network". In: *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2000, pp. 3891–3894.

[76] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks". In: *Nature* 542.7639 (2017), pp. 115–118.

[77] European Parliament and Council of the European Union. *Regulation (EU) 2016/679 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*. 2016.

[78] European Parliament and Council of the European Union. *Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts*. 2021.

[79] L. P. G. Evans, N. M. Adams, and C. Anagnostopoulos. "When Does Active Learning Work?" In: *Advances in Intelligent Data Analysis XII*. Vol. 8207. Springer Berlin Heidelberg, 2013, pp. 174–185.

[80] T. Falk et al. "U-Net: Deep Learning for Cell Counting, Detection, and Morphometry". In: *Nature Methods* 16.1 (2019), pp. 67–70.

[81] F.-L. Fan, J. Xiong, M. Li, and G. Wang. "On Interpretability of Artificial Neural Networks: A Survey". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 5.6 (2021), pp. 741–760.

[82] I. Faye, B. B. Samir, and M. M. Eltoukhy. "Digital Mammograms Classification Using a Wavelet Based Feature Extraction Method". In: *2009 2nd International Conference on Computer and Electrical Engineering*. IEEE, 2009, pp. 318–322.

[83] D. S. Ferreira, G. L. B. Ramalho, F. N. S. Medeiros, A. G. C. Bianchi, C. M. Carneiro, and D. M. Ushizima. "Saliency-Driven System With Deep Learning for Cell Image Classification". In: *2019 IEEE 16th International Symposium on Biomedical Imaging*. IEEE, 2019, pp. 1284–1287.

[84] A. Filby. "Sample Preparation for Flow Cytometry Benefits from Some Lateral Thinking". In: *Cytometry Part A* 89.12 (2016), pp. 1054–1056.

[85] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane. "Adversarial Attacks on Medical Machine Learning". In: *Science* 363.6433 (2019), pp. 1287–1289.

[86] M. Finsterbusch, W. C. Schrottmaier, J. B. Kral-Pointner, M. Salzmann, and A. Assinger. "Measuring and Interpreting Platelet-Leukocyte Aggregates". In: *Platelets* 29.7 (2018), pp. 677–685.

[87] E. Fleisig, R. Abebe, and D. Klein. "When the Majority Is Wrong: Modeling Annotator Disagreement for Subjective Tasks". In: *arXiv* Preprint (2023), p. 2305.06626.

[88] A. W. Flores, K. Bechtel, and C. Lowenkamp. “False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks.” In: *Federal Probation* 80.2 (2016).

[89] A.-K. Frison, P. Wintersberger, A. Riener, C. Schartmüller, L. N. Boyle, E. Miller, and K. Weigl. “In UX We Trust: Investigation of Aesthetics and Usability of Driver-Vehicle Interfaces and Their Impact on the Perception of Automated Driving”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, pp. 1–13.

[90] V. Gamper, A. Butz, and K. Diepold. “Sooner or Later?: Immediate Feedback as a Source of Inspiration in Electronic Brainstorming”. In: *Proceedings of the 29th Australian Conference on Computer-Human Interaction*. ACM, 2017, pp. 182–190.

[91] Y. Gao, M. Zhou, and D. N. Metaxas. “UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation”. In: *Medical Image Computing and Computer Assisted Intervention*. Vol. 12903. Springer International Publishing, 2021, pp. 61–71.

[92] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. “Shortcut Learning in Deep Neural Networks”. In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673.

[93] P. Getreuer. “Chan-Vese Segmentation”. In: *Image Processing On Line* 2 (2012), pp. 214–224.

[94] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam. “The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care”. In: *The Lancet Digital Health* 3.11 (2021), e745–e750.

[95] T. G. Gill. “Expert Systems Usage: Task Change and Intrinsic Motivation”. In: *Management Information Systems Quarterly* 20.3 (1996), p. 301.

[96] N. Gillespie, S. Lockey, and C. Curtis. *Trust in Artificial Intelligence: A Five Country Study*. Tech. rep. The University of Queensland and KPMG, 2021.

[97] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. “Explaining Explanations: An Overview of Interpretability of Machine Learning”. In: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics*. IEEE, 2018, pp. 80–89.

[98] C. Giordano, M. Brennan, B. Mohamed, P. Rashidi, F. Modave, and P. Tighe. “Accessing Artificial Intelligence for Clinical Decision-Making”. In: *Frontiers in Digital Health* 3 (2021), p. 645232.

[99] S. Glüge, S. Balabanov, V. H. Koelzer, and T. Ott. “Evaluation of Deep Learning Training Strategies for the Classification of Bone Marrow Cell Images”. In: *Computer Methods and Programs in Biomedicine* 243 (2024), p. 107924.

[100] T. Go, J. H. Kim, H. Byeon, and S. J. Lee. “Machine Learning-based In-line Holographic Sensing of Unstained Malaria-infected Red Blood Cells”. In: *Journal of Biophotonics* 11.9 (2018), e201800101.

[101] F. C. Godoi, S. Prakash, and B. R. Bhandari. “3D Printing Technologies Applied for Food Design: Status and Prospects”. In: *Journal of Food Engineering* 179 (2016), pp. 44–54.

[102] P. Goktas and R. S. Carbajo. “PPSW–SHAP: Towards Interpretable Cell Classification Using Tree-Based SHAP Image Decomposition and Restoration for High-Throughput Bright-Field Imaging”. In: *Cells* 12.10 (2023), p. 1384.

[103] J. P. Golden, G. A. Justin, M. Nasir, and F. S. Ligler. “Hydrodynamic Focusing - a Versatile Tool”. In: *Analytical and Bioanalytical Chemistry* 402.1 (2012), pp. 325–335.

[104] A. Gomolin, E. Netchiporouk, R. Gniadecki, and I. V. Litvinov. “Artificial Intelligence Applications in Dermatology: Where Do We Stand?” In: *Frontiers in Medicine* 7 (2020), p. 100.

[105] C. González-Gonzalo, E. F. Thee, C. C. Klaver, A. Y. Lee, R. O. Schlingemann, A. Tufail, F. Verbraak, and C. I. Sánchez. “Trustworthy AI: Closing the Gap between Development and Integration of AI Systems in Ophthalmic Practice”. In: *Progress in Retinal and Eye Research* 90 (2022), p. 101034.

[106] B. Goodman and S. Flaxman. “European Union Regulations on Algorithmic Decision Making and a “Right to Explanation””. In: *AI Magazine* 38.3 (2017), pp. 50–57.

[107] D. A. Gorog et al. “Current and Novel Biomarkers of Thrombotic Risk in COVID-19: A Consensus Statement from the International COVID-19 Thrombosis Biomarkers Colloquium”. In: *Nature Reviews Cardiology* 19.7 (2022), pp. 475–495.

[108] N. Goswami, Y. R. He, Y.-H. Deng, C. Oh, N. Sobh, E. Valera, R. Bashir, N. Ismail, H. Kong, T. H. Nguyen, C. Best-Popescu, and G. Popescu. “Label-Free SARS-CoV-2 Detection and Classification Using Phase Imaging with Computational Specificity”. In: *Light: Science & Applications* 10.1 (2021), p. 176.

[109] Y. Gou, Y. Jia, P. Wang, and C. Sun. “Progress of Inertial Microfluidics in Principle and Application”. In: *Sensors* 18.6 (2018), p. 1762.

[110] T. Grote. “Trustworthy Medical AI Systems Need to Know When They Don’t Know”. In: *Journal of Medical Ethics* (2021), p. 107463.

[111] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. “A Survey of Methods for Explaining Black Box Models”. In: *ACM Computing Surveys* 51.5 (2019), pp. 1–42.

[112] G. Gulati, J. Song, A. D. Florea, and J. Gong. “Purpose and Criteria for Blood Smear Scan, Blood Smear Examination, and Blood Smear Review”. In: *Annals of Laboratory Medicine* 33.1 (2013), pp. 1–7.

[113] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. “On Calibration of Modern Neural Networks”. In: *Proceedings of the 34th International Conference on Machine Learning.* 2017, pp. 1321–1330.

[114] M. Habaza, B. Gilboa, Y. Roichman, and N. T. Shaked. “Tomographic Phase Microscopy with 180° Rotation of Live Cells in Suspension by Holographic Optical Tweezers”. In: *Optics Letters* 40.8 (2015), p. 1881.

[115] A. M. Hafiz and G. M. Bhat. “A Survey on Instance Segmentation: State of the Art”. In: *International Journal of Multimedia Information Retrieval* 9.3 (2020), pp. 171–189.

[116] M. Haifler, P. Girshovitz, G. Band, G. Dardikman, I. Madjar, and N. T. Shaked. “Interferometric Phase Microscopy for Label-Free Morphological Evaluation of Sperm Cells”. In: *Fertility and Sterility* 104.1 (2015), pp. 43–47.

[117] R. M. Haralick, I. Dinstein, and K. Shanmugam. “Textural Features for Image Classification”. In: *IEEE Transactions on Systems, Man and Cybernetics* 3.6 (1973), pp. 610–621.

[118] H. R. Hartson, T. S. Andre, and R. C. Williges. "Criteria For Evaluating Usability Evaluation Methods". In: *International Journal of Human-Computer Interaction* 15.1 (2003), pp. 145–181.

[119] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd Edition. Springer New York, 2009.

[120] N. Hatipoglu and G. Bilgin. "Cell Segmentation in Histopathological Images with Deep Learning Algorithms by Utilizing Spatial Relationships". In: *Medical & Biological Engineering & Computing* 55.10 (2017), pp. 1829–1848.

[121] O. Hayden and C. Klenk. "Morphology – Here I Come Again". In: *Cytometry Part A* 99.5 (2021), pp. 472–475.

[122] O. Hayden, K. Peschke, M. Reichert, C. Klenk, P. Knolle, and B. Höchst. "Analysis of Tissue Samples Using Quantitative Phase-Contrast Microscopy". US20240011888A1. 2024.

[123] K. He, G. Gkioxari, P. Dollar, and R. Girshick. "Mask R-CNN". In: *2017 IEEE International Conference on Computer Vision.* IEEE, 2017, pp. 2980–2988.

[124] D. Hendrycks, M. Mazeika, and T. Dietterich. "Deep Anomaly Detection with Outlier Exposure". In: *Proceedings of the International Conference on Learning Representations.* Poster, 2019.

[125] L.-V. Herm, K. Heinrich, J. Wanner, and C. Janiesch. "Stop Ordering Machine Learning Algorithms by Their Explainability! A User-Centered Investigation of Performance and Explainability". In: *International Journal of Information Management* 69 (2023), p. 102538.

[126] S. Hermawati and G. Lawson. "Establishing Usability Heuristics for Heuristics Evaluation in a Specific Domain: Is There a Consensus?" In: *Applied Ergonomics* 56 (2016), pp. 34–51.

[127] L. A. Herzenberg, S. C. De Rosa, and L. A. Herzenberg. "Monoclonal Antibodies and the FACS: Complementary Tools for Immunobiology and Medicine". In: *Immunology Today* 21.8 (2000), pp. 383–390.

[128] High-Level Expert Group on Artificial Intelligence of the European Commission. *Ethics Guidelines for Trustworthy AI.* Publications Office of the European Union, 2019.

[129] R. R. Hoffman, S. T. Mueller, G. Klein, M. Jalaeian, and C. Tate. "Explainable AI: Roles and Stakeholders, Desirements and Challenges". In: *Frontiers in Computer Science* 5 (2023), p. 1117848.

[130] H. Hoffmann. "Kernel PCA for Novelty Detection". In: *Pattern Recognition* 40.3 (2007), pp. 863–874.

[131] R. Hollandi et al. "nucleAIzer: A Parameter-free Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer". In: *Cell Systems* 10.5 (2020), pp. 453–458.

[132] A. Holzinger. "Interactive Machine Learning for Health Informatics: When Do We Need the Human-in-the-Loop?" In: *Brain Informatics* 3.2 (2016), pp. 119–131.

[133] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell. "What Do We Need to Build Explainable AI Systems for the Medical Domain?" In: *arXiv* Preprint (2017), p. 1712.09923.

[134] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller. “Causability and Explainability of Artificial Intelligence in Medicine”. In: *WIREs Data Mining and Knowledge Discovery* 9.4 (2019), e1312.

[135] S. Honavar. “Electronic Medical Records – The Good, the Bad and the Ugly”. In: *Indian Journal of Ophthalmology* 68.3 (2020), p. 417.

[136] R. Horstmeyer, R. Y. Chen, B. Kappes, and B. Judkewitz. “Convolutional Neural Networks That Teach Microscopes How to Image”. In: *arXiv* Preprint (2017), p. 1709.07223.

[137] S. Horton, K. A. Fleming, M. Kuti, L.-M. Looi, S. A. Pai, S. Sayed, and M. L. Wilson. “The Top 25 Laboratory Tests by Volume and Revenue in Five Different Countries”. In: *American Journal of Clinical Pathology* 151.5 (2019), pp. 446–451.

[138] E. Horvitz. “Principles of Mixed-Initiative User Interfaces”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 1999, pp. 159–166.

[139] E. Hüllermeier and W. Waegeman. “Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods”. In: *Machine Learning* 110.3 (2021), pp. 457–506.

[140] A. Jacovi. “Trends in Explainable AI (XAI) Literature”. In: *arXiv* Preprint (2023), p. 2301.05433.

[141] A. Jameson. “Understanding and Dealing With Usability Side Effects of Intelligent Processing”. In: *AI Magazine* 30.4 (2009), pp. 23–40.

[142] B. Javidi, I. Moon, S. Yeom, and E. Carapezza. “Three-Dimensional Imaging and Recognition of Microorganism Using Single-Exposure on-Line (SEOL) Digital Holography”. In: *Optics Express* 13.12 (2005), p. 4492.

[143] C. Jia-Xin and L. Sen. “A Medical Image Segmentation Method Based on Watershed Transform”. In: *The 5th International Conference on Computer and Information Technology.* 2005, pp. 634–638.

[144] J. Jiang, P. Trundle, and J. Ren. “Medical Image Analysis with Artificial Neural Networks”. In: *Computerized Medical Imaging and Graphics* 34.8 (2010), pp. 617–631.

[145] Y. Jo, H. Cho, S. Y. Lee, G. Choi, G. Kim, H.-S. Min, and Y. Park. “Quantitative Phase Imaging and Artificial Intelligence: A Review”. In: *IEEE Journal of Selected Topics in Quantum Electronics* 25.1 (2019), pp. 1–14.

[146] Y. Jo, J. Jung, M.-H. Kim, H. Park, S.-J. Kang, and Y. Park. “Label-Free Identification of Individual Bacteria Using Fourier Transform Light Scattering”. In: *Optics Express* 23.12 (2015), p. 15792.

[147] J. W. Johnson. “Adapting Mask-RCNN for Automatic Nucleus Segmentation”. In: *arXiv* Preprint (2018), p. 1805.00500.

[148] M. D. Joshi, A. H. Karode, and Suralkar. “White Blood Cells Segmentation and Classification to Detect Acute Leukemia Ms”. In: *International Journal of Emerging Trends & Technology in Computer Science* 2.3 (2013), pp. 147–151.

[149] P. Jourdain, D. Boss, B. Rappaz, C. Moratal, M.-C. Hernandez, C. Depeursinge, P. J. Magistretti, and P. Marquet. “Simultaneous Optical Recording in Multiple Cells by Digital Holographic Microscopy of Chloride Current Associated to Activation of the Ligand-Gated Chloride Channel GABAA Receptor”. In: *PLoS ONE* 7.12 (2012), e51041.

[150] T. Kanade, Z. Yin, R. Bise, S. Huh, S. Eom, M. F. Sandbothe, and M. Chen. “Cell Image Analysis: Algorithms, System and Applications”. In: *2011 IEEE Workshop on Applications of Computer Vision*. IEEE, 2011, pp. 374–381.

[151] M. Kaneko, K. Tsuji, K. Masuda, K. Ueno, K. Henmi, S. Nakagawa, R. Fujita, K. Suzuki, Y. Inoue, S. Teramukai, E. Konishi, T. Takamatsu, and O. Ukimura. “Urine Cell Image Recognition Using a Deep-learning Model for an Automated Slide Evaluation System”. In: *BJU International* 130.2 (2022), pp. 235–243.

[152] D. Katehakis, A. Kouroubali, and I. Fundulaki. “Towards the Development of a National eHealth Interoperability Framework to Address Public Health Challenges in Greece”. In: *1st International Workshop on Semantic Web Technologies for Health Data Management*. 2018.

[153] A. Katzmann, O. Taubmann, S. Ahmad, A. Mühlberg, M. Sühling, and H.-M. Groß. “Explaining Clinical Decision Support Systems in Medical Imaging Using Cycle-Consistent Activation Maximization”. In: *Neurocomputing* 458 (2021), pp. 141–156.

[154] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. “Key Challenges for Delivering Clinical Impact with Artificial Intelligence”. In: *BMC Medicine* 17.1 (2019), p. 195.

[155] B. Kemper, A. Bauwens, D. Bettenworth, M. Götte, B. Greve, L. Kastl, S. Ketelhut, P. Lenz, S. Mues, J. Schnekenburger, and A. Vollmer. “Label-Free Quantitative In Vitro Live Cell Imaging with Digital Holographic Microscopy”. In: *Bioanalytical Reviews*. Vol. 2. Springer Berlin Heidelberg, 2019.

[156] S. Khan, A. Jesacher, W. Nussbaumer, S. Bernet, and M. Ritsch-Marte. “Quantitative Analysis of Shape and Volume Changes in Activated Thrombocytes in Real Time by Single-shot Spatial Light Modulator-based Differential Interference Contrast Imaging”. In: *Journal of Biophotonics* 4.9 (2011), pp. 600–609.

[157] G. Kim, Y. Jo, H. Cho, H.-s. Min, and Y. Park. “Learning-Based Screening of Hematologic Disorders Using Quantitative Phase Imaging of Individual Red Blood Cells”. In: *Biosensors and Bioelectronics* 123 (2019), pp. 69–76.

[158] K. Kim, K. S. Kim, H. Park, J. C. Ye, and Y. Park. “Real-Time Visualization of 3-D Dynamic Microscopic Objects Using Optical Diffraction Tomography”. In: *Optics Express* 21.26 (2013), p. 32269.

[159] K. Kim, H. Yoon, M. Diez-Silva, M. Dao, R. R. Dasari, and Y. Park. “High-Resolution Three-Dimensional Imaging of Red Blood Cells Parasitized by Plasmodium Falciparum and in Situ Hemozoin Crystals Using Optical Diffraction Tomography”. In: *Journal of Biomedical Optics* 19.01 (2013), p. 1.

[160] M. K. Kim. *Digital Holographic Microscopy: Principles, Techniques, and Applications*. 1st Edition. Vol. 162. Springer Series in Optical Sciences. Springer New York, 2011.

[161] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. "The (Un)Reliability of Saliency Methods". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning.* Vol. 11700. Springer International Publishing, 2019, pp. 267–280.

[162] C. Klenk, D. Heim, M. Ugele, and O. Hayden. "Impact of Sample Preparation on Holographic Imaging of Leukocytes". In: *Optical Engineering* 59.10 (2019), p. 1.

[163] E. M. Knorr, R. T. Ng, and V. Tucakov. "Distance-Based Outliers: Algorithms and Applications". In: *The VLDB Journal The International Journal on Very Large Data Bases* 8.3-4 (2000), pp. 237–253.

[164] P. Kocsis, I. Shevkunov, V. Katkovnik, and K. Egiazarian. "Single Exposure Lensless Subpixel Phase Imaging: Optical System Design, Modelling, and Experimental Study". In: *Optics Express* 28.4 (2020), p. 4625.

[165] Y. Al-Kofahi, A. Zaltsman, R. Graves, W. Marshall, and M. Rusu. "A Deep Learning-Based Algorithm for 2-D Cell Segmentation in Microscopy Images". In: *BMC Bioinformatics* 19.1 (2018), p. 365.

[166] T. Kohonen. "Self-Organized Formation of Topologically Correct Feature Maps". In: *Biological Cybernetics* 43.1 (1982), pp. 59–69.

[167] B. Kompa, J. Snoek, and A. L. Beam. "Second Opinion Needed: Communicating Uncertainty in Medical Machine Learning". In: *npj Digital Medicine* 4.1 (2021), p. 4.

[168] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor. "User Preferences for Hybrid Explanations". In: *Proceedings of the Eleventh ACM Conference on Recommender Systems.* ACM, 2017, pp. 84–88.

[169] P. Koutsabasis, T. Spyrou, and J. Darzentas. "Evaluating Usability Evaluation Methods: Criteria, Method and a Case Study". In: *Human-Computer Interaction. Interaction Design and Usability.* Vol. 4550. Springer Berlin Heidelberg, 2007, pp. 569–578.

[170] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet Classification with Deep Convolutional Neural Networks". In: *Advances in neural information processing systems* 25 (2012).

[171] F. Kulwa, C. Li, X. Zhao, B. Cai, N. Xu, S. Qi, S. Chen, and Y. Teng. "A State-of-the-Art Survey for Microorganism Image Segmentation Methods and Future Potential". In: *IEEE Access* 7 (2019), pp. 100243–100269.

[172] R. Kurzban, M. L. Rigdon, and B. J. Wilson. "Incremental Approaches to Establishing Trust". In: *Experimental Economics* 11.4 (2008), pp. 370–389.

[173] S. Kusunose, Y. Shinomiya, T. Ushiwaka, N. Maeda, and Y. Hoshino. "Improving Individually Selectness for Immune Cells Using GradCAM". In: *2021 IEEE 5ht International Conference on Cybernetics.* IEEE, 2021, pp. 034–038.

[174] T. Kutscher, K. Eder, A. Marzi, Á. Barroso, J. Schnekenburger, and B. Kemper. "Cell Detection and Segmentation in Quantitative Digital Holographic Phase Contrast Images Utilizing a Mask Region-based Convolutional Neural Network". In: *OSA Optical Sensors and Sensing Congress.* OSA, 2021, JTu5A.23.

[175] E. Kwee, A. Peterson, M. Halter, and J. Elliott. "Practical Application of Microsphere Samples for Benchmarking a Quantitative Phase Imaging System". In: *Cytometry Part A* 99.10 (2021), pp. 1022–1032.

[176] A. B. Labrique and W. K.-Y. Pan. “Diagnostic Tests: Understanding Results, Assessing Utility, and Predicting Performance”. In: *American Journal of Ophthalmology* 149.6 (2010), pp. 878–881.

[177] V. K. Lam, T. C. Nguyen, V. Bui, B. M. Chung, L.-C. Chang, G. Nehmetallah, and C. B. Raub. “Quantitative Scoring of Epithelial and Mesenchymal Qualities of Cancer Cells Using Machine Learning and Quantitative Phase Imaging”. In: *Journal of Biomedical Optics* 25.02 (2020), p. 1.

[178] V. K. Lam, T. C. Nguyen, B. M. Chung, G. Nehmetallah, and C. B. Raub. “Quantitative Assessment of Cancer Cell Morphology and Motility Using Telecentric Digital Holographic Microscopy and Machine Learning”. In: *Cytometry Part A* 93.3 (2018), pp. 334–345.

[179] F. Lateef and Y. Ruichek. “Survey on Semantic Segmentation Using Deep Learning Techniques”. In: *Neurocomputing* 338 (2019), pp. 321–348.

[180] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. “Gradient-Based Learning Applied to Document Recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[181] Y. LeCun, Y. Bengio, and G. Hinton. “Deep Learning”. In: *Nature* 521.7553 (2015), pp. 436–444.

[182] S. Lee, H. Park, K. Kim, Y. Sohn, S. Jang, and Y. Park. “Refractive Index Tomograms and Dynamic Membrane Fluctuations of Red Blood Cells from Patients with Diabetes Mellitus”. In: *Scientific Reports* 7.1 (2017), p. 1039.

[183] L. S. Lehmann, A. L. Puopolo, S. Shaykevich, and T. A. Brennan. “Iatrogenic Events Resulting in Intensive Care Admission: Frequency, Cause, and Disclosure to Patients and Institutions”. In: *The American Journal of Medicine* 118.4 (2005), pp. 409–413.

[184] J. R. Lewis. “Sample Sizes for Usability Studies: Additional Considerations”. In: *The Journal of the Human Factors and Ergonomics Society* 36.2 (1994), pp. 368–378.

[185] H. Lieberman. “User Interface Goals, AI Opportunities”. In: *AI Magazine* 30.4 (2009), pp. 16–22.

[186] Z. C. Lipton. “The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery.” In: *Queue* 16.3 (2018), pp. 31–57.

[187] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. “A Survey on Deep Learning in Medical Image Analysis”. In: *Medical Image Analysis* 42 (2017), pp. 60–88.

[188] G. Litjens et al. “1399 H&E-stained Sentinel Lymph Node Sections of Breast Cancer Patients: The CAMELYON Dataset”. In: *GigaScience* 7.6 (2018).

[189] P. Y. Liu, L. K. Chin, W. Ser, H. F. Chen, C.-M. Hsieh, C.-H. Lee, K.-B. Sung, T. C. Ayi, P. H. Yap, B. Liedberg, K. Wang, T. Bourouina, and Y. Leprince-Wang. “Cell Refractive Index for Cell Biology and Disease Diagnosis: Past, Present and Future”. In: *Lab on a Chip* 16.4 (2016), pp. 634–644.

[190] X. Liu et al. “A Comparison of Deep Learning Performance against Health-Care Professionals in Detecting Diseases from Medical Imaging: A Systematic Review and Meta-Analysis”. In: *The Lancet Digital Health* 1.6 (2019), e271–e297.

[191] X. Liu et al. “Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence: The CONSORT-AI Extension”. In: *Nature Medicine* 26.9 (2020), pp. 1364–1374.

[192] P. Liznerski, L. Ruff, R. A. Vandermeulen, B. J. Franks, M. Kloft, and K.-R. Müller. “Explainable Deep One-Class Classification”. In: *Proceedings of the International Conference on Learning Representations*. Poster, 2021.

[193] A. J. London. “Artificial Intelligence and Black-Box Medical Decisions: *Accuracy versus Explainability*”. In: *Hastings Center Report* 49.1 (2019), pp. 15–21.

[194] F. Long. “Microscopy Cell Nuclei Segmentation with Enhanced U-Net”. In: *BMC Bioinformatics* 21.1 (2020), p. 8.

[195] A. Lopatina, S. Ropele, R. Sibgatulin, J. R. Reichenbach, and D. Güllmar. “Investigation of Deep-Learning-Driven Identification of Multiple Sclerosis Patients Based on Susceptibility-Weighted Images Using Relevance Analysis”. In: *Frontiers in Neuroscience* 14 (2020), p. 609468.

[196] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. “Accurate Intelligible Models with Pairwise Interactions”. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2013, pp. 623–631.

[197] D. Lowell, Z. C. Lipton, and B. C. Wallace. “Practical Obstacles to Deploying Active Learning”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2019, pp. 21–30.

[198] S. M. Lundberg and S.-I. Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 4768–4777.

[199] J. H. Luong, K. B. Male, and J. D. Glennon. “Biosensor Technology: Technology Push versus Market Pull”. In: *Biotechnology Advances* 26.5 (2008), pp. 492–500.

[200] Q. Ma, H. Ma, F. Xu, X. Wang, and W. Sun. “Microfluidics in Cardiovascular Disease Research: State of the Art and Future Outlook”. In: *Microsystems & Nanoengineering* 7.1 (2021), pp. 1–19.

[201] A. Madabhushi and G. Lee. “Image Analysis and Machine Learning in Digital Pathology: Challenges and Opportunities”. In: *Medical Image Analysis* 33 (2016), pp. 170–175.

[202] P. R. Magesh, R. D. Myloth, and R. J. Tom. “An Explainable Machine Learning Model for Early Detection of Parkinson’s Disease Using LIME on DaTSCAN Imagery”. In: *Computers in Biology and Medicine* 126 (2020), p. 104041.

[203] K. Marquet, N. Claes, E. De Troy, G. Kox, M. Droogmans, W. Schrooten, F. Weekers, A. Vlayen, M. Vandersteen, and A. Vleugels. “One Fourth of Unplanned Transfers to a Higher Level of Care Are Associated With a Highly Preventable Adverse Event: A Patient Record Review in Six Belgian Hospitals”. In: *Critical Care Medicine* 43.5 (2015), pp. 1053–1061.

[204] P. Marquet, C. Depeursinge, and P. J. Magistretti. “Review of Quantitative Phase-Digital Holographic Microscopy: Promising Novel Imaging Technique to Resolve Neuronal Network Activity and Identify Cellular Biomarkers of Psychiatric Disorders”. In: *Neurophotonics* 1.2 (2014), p. 020901.

[205] R. C. Mayer, J. H. Davis, and F. D. Schoorman. “An Integrative Model of Organizational Trust”. In: *The Academy of Management Review* 20.3 (1995), p. 709.

[206] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* 1st Edition. Wiley Series in Probability and Statistics. 2004.

[207] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro. “Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images”. In: *Medical Image Computing and Computer Assisted Intervention.* Vol. 11071. 2018, pp. 893–901.

[208] E. Meijering. “Cell Segmentation: 50 Years Down the Road [Life Sciences]”. In: *IEEE Signal Processing Magazine* 29.5 (2012), pp. 140–145.

[209] C. L. Meinert and S. Tonascia. *Clinical Trials: Design, Conduct, and Analysis.* 2nd Edition. Oxford University Press, 1986.

[210] L. Meintker, J. Ringwald, M. Rauh, and S. W. Krause. “Comparison of Automated Differential Blood Cell Counts From Abbott Sapphire, Siemens Advia 120, Beckman Coulter DxH 800, and Sysmex XE-2100 in Normal and Pathologic Samples”. In: *American Journal of Clinical Pathology* 139.5 (2013), pp. 641–650.

[211] F. Merola, P. Memmolo, L. Miccio, R. Savoia, M. Mugnano, A. Fontana, G. D’Ippolito, A. Sardo, A. Iolascon, A. Gambale, and P. Ferraro. “Tomographic Flow Cytometry by Digital Holography”. In: *Light: Science & Applications* 6.4 (2016), e16241–e16241.

[212] B. Midtvedt, S. Helgadottir, A. Argun, J. Pineda, D. Midtvedt, and G. Volpe. “Quantitative Digital Microscopy with Deep Learning”. In: *Applied Physics Reviews* 8.1 (2021), p. 011310.

[213] T. Miller. “Are We Measuring Trust Correctly in Explainability, Interpretability, and Transparency Research?” In: *arXiv* Preprint (2022), p. 2209.00651.

[214] T. Miller, P. Howe, and L. Sonenberg. “Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences”. In: *arXiv* Preprint (2017), p. 1712.00547.

[215] M. Mir, T. Kim, A. Majumder, M. Xiang, R. Wang, S. C. Liu, M. U. Gillette, S. Stice, and G. Popescu. “Label-Free Characterization of Emerging Human Neuronal Networks”. In: *Scientific Reports* 4.1 (2014), p. 4434.

[216] M. Mir, Z. Wang, Z. Shen, M. Bednarz, R. Bashir, I. Golding, S. G. Prasanth, and G. Popescu. “Optical Measurement of Cycle-Dependent Cell Growth”. In: *Proceedings of the National Academy of Sciences* 108.32 (2011), pp. 13124–13129.

[217] V. Mnih et al. “Human-Level Control through Deep Reinforcement Learning”. In: *Nature* 518.7540 (2015), pp. 529–533.

[218] E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, and D. Van Valen. “Deep Learning for Cellular Image Analysis”. In: *Nature Methods* 16.12 (2019), pp. 1233–1246.

[219] B. Mohanta, P. Das, and S. Patnaik. “Healthcare 5.0: A Paradigm Shift in Digital Healthcare System Using Artificial Intelligence, IOT and 5G Communication”. In: *2019 International Conference on Applied Machine Learning.* IEEE, 2019, pp. 191–196.

[220] H. Motherby, B. Nadjari, P. Friegel, J. Kohaus, U. Ramp, and A. Böcking. “Diagnostic Accuracy of Effusion Cytology”. In: *Diagnostic Cytopathology* 20.6 (1999), pp. 350–357.

[221] M. M. Moya and D. R. Hush. “Network Constraints and Multi-Objective Optimization for One-Class Classification”. In: *Neural Networks* 9.3 (1996), pp. 463–474.

[222] U. J. Muehlematter, P. Daniore, and K. N. Vokinger. “Approval of Artificial Intelligence and Machine Learning-Based Medical Devices in the USA and Europe (2015–20): A Comparative Analysis”. In: *The Lancet Digital Health* 3.3 (2021), e195–e203.

[223] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. “Definitions, Methods, and Applications in Interpretable Machine Learning”. In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22071–22080.

[224] Y. Nagao, M. Sakamoto, T. Chinen, Y. Okada, and D. Takao. “Robust Classification of Cell Cycle Phase and Biological Feature Extraction by Image-Based Deep Learning”. In: *Molecular Biology of the Cell* 31.13 (2020), pp. 1346–1354.

[225] A. Narla, B. Kuprel, K. Sarin, R. Novoa, and J. Ko. “Automated Classification of Skin Lesions: From Pixels to Practice”. In: *Journal of Investigative Dermatology* 138.10 (2018), pp. 2108–2110.

[226] M. Nassar, M. Doan, A. Filby, O. Wolkenhauer, D. K. Fogg, J. Piasecka, C. A. Thornton, A. E. Carpenter, H. D. Summers, P. Rees, and H. Hennig. “Label-Free Identification of White Blood Cells Using Machine Learning”. In: *Cytometry Part A* 95.8 (2019), pp. 836–842.

[227] H. Ng, S. Ong, K. Foong, P. Goh, and W. Nowinski. “Medical Image Segmentation Using K-Means Clustering and Improved Watershed Algorithm”. In: *2006 IEEE Southwest Symposium on Image Analysis and Interpretation*. IEEE, 2006, pp. 61–65.

[228] V.-L. Nguyen, M. H. Shaker, and E. Hüllermeier. “How to Measure Uncertainty in Uncertainty Sampling for Active Learning”. In: *Machine Learning* 111.1 (2022), pp. 89–122.

[229] T. H. Nguyen, S. Sridharan, V. Macias, A. Kajdacsy-Balla, J. Melamed, M. N. Do, and G. Popescu. “Automatic Gleason Grading of Prostate Cancer Using Quantitative Phase Imaging and Machine Learning”. In: *Journal of Biomedical Optics* 22.3 (2017), p. 036015.

[230] T. L. Nguyen, S. Pradeep, R. L. Judson-Torres, J. Reed, M. A. Teitell, and T. A. Zangle. “Quantitative Phase Imaging: Recent Advances and Expanding Potential in Biomedicine”. In: *ACS Nano* 16.8 (2022), pp. 11516–11544.

[231] J. Nielsen. “Finding Usability Problems through Heuristic Evaluation”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1992, pp. 373–380.

[232] J. Nielsen. *Usability Engineering*. 1st Edition. Academic Press, 1993.

[233] J. Nielsen. *10 Usability Heuristics for User Interface Design*. https://www.nngroup.com/articles/ten-usability-heuristics/. 2020.

[234] J. Nielsen and R. Molich. “Heuristic Evaluation of User Interfaces”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Empowering People*. ACM, 1990, pp. 249–256.

[235] M. Nishikawa et al. “Massive Image-Based Single-Cell Profiling Reveals High Levels of Circulating Platelet Aggregates in Patients with COVID-19”. In: *Nature Communications* 12.1 (2021), p. 7135.

[236] K. Nishimura, D. F. E. Ker, and R. Bise. “Weakly Supervised Cell Instance Segmentation by Propagating from Detection Response”. In: *Medical Image Computing and Computer Assisted Intervention*. Vol. 11764. Springer International Publishing, 2019, pp. 649–657.

[237] N. Nissim, M. Dudaie, I. Barnea, and N. T. Shaked. “Real-Time Stain-Free Classification of Cancer Cells and Blood Cells Using Interferometric Phase Microscopy and Machine Learning”. In: *Cytometry Part A* 99.5 (2021), pp. 511–523.

[238] J. Noyes and C. Baber. *User-Centred Design of Systems*. 1st Edition. Applied Computing. Springer London, 1999.

[239] J. Noyes and C. Baber. “Who Will Use the System?” In: *User-Centred Design of Systems*. Springer London, 1999, pp. 17–36.

[240] K. O’Neill, N. Aghaeepour, J. Špidlen, and R. Brinkman. “Flow Cytometry Bioinformatics”. In: *PLoS Computational Biology* 9.12 (2013), e1003365.

[241] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. “Attention U-Net: Learning Where to Look for the Pancreas”. In: *Medical Imaging with Deep Learning*. Poster, 2018.

[242] M. A. Onari, I. Grau, M. S. Nobile, and Y. Zhang. “Trustworthy Artificial Intelligence in Medical Applications: A Mini Survey”. In: *2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*. IEEE, 2023, pp. 1–8.

[243] M. Z. Othman, T. S. Mohammed, and A. B. Ali. “Neural Network Classification of White Blood Cell Using Microscopic Images”. In: *International Journal of Advanced Computer Science and Applications* 8.5 (2017).

[244] C. Ounkomol, S. Seshamani, M. M. Maleckar, F. Collman, and G. R. Johnson. “Label-Free Prediction of Three-Dimensional Fluorescence Images from Transmitted-Light Microscopy”. In: *Nature Methods* 15.11 (2018), pp. 917–920.

[245] Y. Ozaki et al. “Label-Free Classification of Cells Based on Supervised Machine Learning of Subcellular Structures”. In: *PLoS ONE* 14.1 (2019), e0211347.

[246] A. Páez. “The Pragmatic Turn in Explainable Artificial Intelligence (XAI)”. In: *Minds and Machines* 29.3 (2019), pp. 441–459.

[247] S. K. Paidi, P. Raj, R. Bordett, C. Zhang, S. H. Karandikar, R. Pandey, and I. Barman. “Raman and Quantitative Phase Imaging Allow Morpho-Molecular Recognition of Malignancy and Stages of B-cell Acute Lymphoblastic Leukemia”. In: *Biosensors and Bioelectronics* 190 (2021), p. 113403.

[248] I. Palatnik De Sousa, M. Maria Bernardes Rebuzzi Vellasco, and E. Costa Da Silva. “Local Interpretable Model-Agnostic Explanations for Classification of Lymph Node Metastases”. In: *Sensors* 19.13 (2019), p. 2969.

[249] D. P. Panagoulias, M. Virvou, and G. A. Tsihrintzis. “Tailored Explainability in Medical Artificial Intelligence-empowered Applications: Personalisation via the Technology Acceptance Model”. In: *2023 IEEE 35th International Conference on Tools with Artificial Intelligence*. IEEE, 2023, pp. 486–490.

[250] H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh. "A Deep Learning and Grad-CAM Based Color Visualization Approach for Fast Detection of COVID-19 Cases Using Chest X-ray and CT-Scan Images". In: *Chaos, Solitons & Fractals* 140 (2020), p. 110190.

[251] J. Park et al. "Artificial Intelligence-Enabled Quantitative Phase Imaging Methods for Life Sciences". In: *Nature Methods* 20.11 (2023), pp. 1645–1660.

[252] S. H. Park and K. Han. "Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction". In: *Radiology* 286.3 (2018), pp. 800–809.

[253] Y. Park, C. A. Best, K. Badizadegan, R. R. Dasari, M. S. Feld, T. Kuriabova, M. L. Henle, A. J. Levine, and G. Popescu. "Measurement of Red Blood Cell Mechanics during Morphological Changes". In: *Proceedings of the National Academy of Sciences* 107.15 (2010), pp. 6731–6736.

[254] Y. Park, C. Depeursinge, and G. Popescu. "Quantitative Phase Imaging in Biomedicine". In: *Nature Photonics* 12.10 (2018), pp. 578–589.

[255] U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly. "Explainable AI in Healthcare". In: *International Conference on Cyber Situational Awareness, Data Analytics and Assessment.* IEEE, 2020, pp. 1–2.

[256] N. Peiffer-Smadja, T. Rawson, R. Ahmad, A. Buchard, P. Georgiou, F.-X. Lescure, G. Birgand, and A. Holmes. "Machine Learning for Clinical Decision Support in Infectious Diseases: A Narrative Review of Current Applications". In: *Clinical Microbiology and Infection* 26.5 (2020), pp. 584–595.

[257] J. M. Perkel. "Cell Signaling: In Vivo Veritas". In: *Science* 316.5832 (2007), pp. 1763–1768.

[258] M. Poostchi, K. Silamut, R. J. Maude, S. Jaeger, and G. Thoma. "Image Analysis and Machine Learning for Detecting Malaria". In: *Translational Research* 194 (2018), pp. 36–55.

[259] G. Popescu. *Quantitative Phase Imaging of Cells and Tissues.* 1st Edition. McGraw-Hill, 2011.

[260] G. Popescu, Y. Park, N. Lue, C. Best-Popescu, L. Deflores, R. R. Dasari, M. S. Feld, and K. Badizadegan. "Optical Imaging of Cell Mass and Growth Dynamics". In: *American Journal of Physiology-Cell Physiology* 295.2 (2008), pp. C538–C544.

[261] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty. "Stakeholders in Explainable AI". In: *Artificial Intelligence in Government and Public Sector.* Poster, 2018.

[262] P. Pu and L. Chen. "Trust Building with Explanation Interfaces". In: *Proceedings of the 11th International Conference on Intelligent User Interfaces.* ACM, 2006, pp. 93–100.

[263] H. Qiao and J. Wu. "GPU-based Deep Convolutional Neural Network for Tomographic Phase Microscopy with L1 Fitting and Regularization". In: *Journal of Biomedical Optics* 23.6 (2018), p. 066003.

[264] K. Quig, E. G. Wheatley, and M. O'Hara. "Perspectives On Blood-Based Point-Of-Care Diagnostics". In: *Open Access Emergency Medicine* 11 (2019), pp. 291–296.

[265] R. F. Rahmat, F. S. Wulandari, S. Faza, M. A. Muchtar, and I. Siregar. “The Morphological Classification of Normal and Abnormal Red Blood Cell Using Self Organizing Map”. In: *IOP Conference Series: Materials Science and Engineering* 308 (2018), p. 012015.

[266] M. J. Raihan and A.-A. Nahid. “Malaria Cell Image Classification by Explainable Artificial Intelligence”. In: *Health and Technology* 12.1 (2022), pp. 47–58.

[267] E. R. Ranschaert, S. Morozov, and P. R. Algra. *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks.* 1st Edition. Springer International Publishing, 2019.

[268] B. Rappaz, A. Barbul, Y. Emery, R. Korenstein, C. Depeursinge, P. J. Magistretti, and P. Marquet. “Comparative Study of Human Erythrocytes by Digital Holographic Microscopy, Confocal Microscopy, and Impedance Volume Analyzer”. In: *Cytometry Part A* 73A.10 (2008), pp. 895–903.

[269] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang. “Deep Learning for Health Informatics”. In: *IEEE Journal of Biomedical and Health Informatics* 21.1 (2017), pp. 4–21.

[270] H. Rezaeilouyeh, A. Mollahosseini, and M. H. Mahoor. “Microscopic Medical Image Classification Framework via Deep Learning and Shearlet Transform”. In: *Journal of Medical Imaging* 3.4 (2016), p. 044501.

[271] M. Ribeiro, S. Singh, and C. Guestrin. ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations.* Association for Computational Linguistics, 2016, pp. 97–101.

[272] L. Righetti, R. Madhavan, and R. Chatila. “Unintended Consequences of Biased Robotic and Artificial Intelligence Systems [Ethical, Legal, and Societal Issues]”. In: *IEEE Robotics & Automation Magazine* 26.3 (2019), pp. 11–13.

[273] Y. Rivenson, Y. Zhang, H. Günaydın, D. Teng, and A. Ozcan. “Phase Recovery and Holographic Image Reconstruction Using Deep Learning in Neural Networks”. In: *Light: Science & Applications* 7.2 (2017), pp. 17141–17141.

[274] U.-P. Rohr, C. Binder, T. Dieterle, F. Giusti, C. G. M. Messina, E. Toerien, H. Moch, and H. H. Schäfer. “The Value of In Vitro Diagnostic Testing in Medical Practice: A Status Report”. In: *PLoS ONE* 11.3 (2016), e0149856.

[275] D. Roitshtain, L. Wolbromsky, E. Bal, H. Greenspan, L. L. Satterwhite, and N. T. Shaked. “Quantitative Phase Microscopy Spatial Signatures of Cancer Cells”. In: *Cytometry Part A* 91.5 (2017), pp. 482–493.

[276] O. Ronneberger, P. Fischer, and T. Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention.* Vol. 9351. 2015, pp. 234–241.

[277] A. S. Ross, M. C. Hughes, and F. Doshi-Velez. “Right for the Right Reasons: Training Differentiable Models by Constraining Their Explanations”. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence.* 2017, pp. 2662–2670.

[278] M. Rubin, O. Stein, N. A. Turko, Y. Nygate, D. Roitshtain, L. Karako, I. Barnea, R. Giryes, and N. T. Shaked. “TOP-GAN: Stain-free Cancer Cell Classification Using Deep Learning with a Small Training Set”. In: *Medical Image Analysis* 57 (2019), pp. 176–185.

[279] C. Rudin. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

[280] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft. “Deep One-Class Classification”. In: *Proceedings of the 35th International Conference on Machine Learning.* Proceedings of Machine Learning Research. PMLR, 2018, pp. 4393–4402.

[281] C. Rzepka and B. Berger. “User Interaction with AI-enabled Systems: A Systematic Review of IS Research”. In: *International Conference on Information Systems* 39 (2018).

[282] S. K. Sadanandan, P. Ranefall, S. Le Guyader, and C. Wählby. “Automated Training of Deep Convolutional Neural Networks for Cell Segmentation”. In: *Scientific Reports* 7.1 (2017), p. 7860.

[283] H. Sahoo. “Fluorescent Labeling Techniques in Biomolecules: A Flashback”. In: *RSC Advances* 2.18 (2012), p. 7017.

[284] W. Samek and K.-R. Müller. “Towards Explainable Artificial Intelligence”. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning.* Vol. 11700. Springer International Publishing, 2019, pp. 5–22.

[285] D. Saraswat, P. Bhattacharya, A. Verma, V. K. Prasad, S. Tanwar, G. Sharma, P. N. Bokoro, and R. Sharma. “Explainable AI for Healthcare 5.0: Opportunities and Challenges”. In: *IEEE Access* 10 (2022), pp. 84486–84517.

[286] A. Sauer, M. Dmitrieva, H. Han, and J. Rittscher. “Leveraging Inter-Annotator Disagreement for Semi-Supervised Segmentation”. In: *2023 IEEE 20th International Symposium on Biomedical Imaging.* IEEE, 2023, pp. 1–5.

[287] Z. El-Schich, S. Kamlund, B. Janicke, K. Alm, and A. G. Wingren. “Holography: The Usefulness of Digital Holographic Microscopy for Clinical Diagnostics”. In: *Holographic Materials and Optical Systems.* InTech, 2017.

[288] U. Schnars and W. P. O. Jüptner. “Digital Recording and Numerical Reconstruction of Holograms”. In: *Measurement Science and Technology* 13.9 (2002), R85–R101.

[289] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. “Support Vector Method for Novelty Detection”. In: *Proceedings of the 12th International Conference on Neural Information Processing Systems.* MIT Press, 1999, pp. 582–588.

[290] T. A. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, and K. Van Den Bosch. “Human-Centered XAI: Developing Design Patterns for Explanations of Clinical Decision Support Systems”. In: *International Journal of Human-Computer Studies* 154 (2021), p. 102684.

[291] V. Sehwag, M. Chiang, and P. Mittal. “SSD: A Unified Framework for Self-Supervised Outlier Detection”. In: *International Conference on Learning Representations.* Poster, 2021.

[292] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *2017 IEEE International Conference on Computer Vision.* IEEE, 2017, pp. 618–626.

[293] M. Sensoy, L. Kaplan, and M. Kandemir. "Evidential Deep Learning to Quantify Classification Uncertainty". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems.* 2018, pp. 3183–3193.

[294] A. Serban, K. Van Der Blom, H. Hoos, and J. Visser. "Practices for Engineering Trustworthy Machine Learning Applications". In: *2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI.* IEEE, 2021, pp. 97–100.

[295] S. Serte, A. Serener, and F. Al-Turjman. "Deep Learning in Medical Imaging: A Brief Review". In: *Transactions on Emerging Telecommunications Technologies* 33.10 (2022), e4080.

[296] B. Settles. *Active Learning.* 1st Edition. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer International Publishing, 2012.

[297] H. M. Shapiro. *Practical Flow Cytometry.* 4th Edition. John Wiley & Sons, 2005.

[298] R. T. Shinohara, E. M. Sweeney, J. Goldsmith, N. Shiee, F. J. Mateen, P. A. Calabresi, S. Jarso, D. L. Pham, D. S. Reich, and C. M. Crainiceanu. "Statistical Normalization Techniques for Magnetic Resonance Imaging". In: *NeuroImage: Clinical* 6 (2014), pp. 9–19.

[299] B. Shneiderman. "Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems". In: *ACM Transactions on Interactive Intelligent Systems* 10.4 (2020), pp. 1–31.

[300] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, and N. Elmqvist. *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* Sixth edition, global edition. Pearson, 2018.

[301] A. Shrikumar, P. Greenside, and A. Kundaje. "Learning Important Features through Propagating Activation Differences". In: *Proceedings of the 34th International Conference on Machine Learning.* 2017, pp. 3145–3153.

[302] K. Simonyan, A. Vedaldi, and A. Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *International Conference on Learning Representations.* Poster, 2014.

[303] A. Singh, S. Sengupta, and V. Lakshminarayanan. "Explainable Deep Learning Models in Medical Image Analysis". In: *Journal of Imaging* 6.6 (2020), p. 52.

[304] J. Sistermanns, Emken, Ellen, Hayden, Oliver, Weirich, Gregor, and Utschick, Wolfgang. "Unsupervised High-Throughput Segmentation of Cells and Cell Nuclei in Quantitative Phase Images". In: *IEEE 21th International Symposium on Biomedical Imaging.* IEEE, 2024.

[305] D. F. Sittig. "Grand Challenges in Medical Informatics?" In: *Journal of the American Medical Informatics Association* 1.5 (1994), pp. 412–413.

[306] B. Smith and W. Ceusters. "HL7 RIM: An Incoherent Standard". In: *Studies in Health Technology and Informatics* 124 (2006), pp. 133–138.

[307] K. Sokol and P. Flach. "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* ACM, 2020, pp. 56–67.

[308] B. Song et al. "Bayesian Deep Learning for Reliable Oral Cancer Image Classification". In: *Biomedical Optics Express* 12.10 (2021), p. 6422.

[309] Y. Song, E.-L. Tan, X. Jiang, J.-Z. Cheng, D. Ni, S. Chen, B. Lei, and T. Wang. “Accurate Cervical Cell Segmentation from Overlapping Clumps in Pap Smear Images”. In: *IEEE Transactions on Medical Imaging* 36.1 (2017), pp. 288–300.

[310] V. Sounderajah et al. “Developing Specific Reporting Guidelines for Diagnostic Accuracy Studies Assessing AI Interventions: The STARD-AI Steering Group”. In: *Nature Medicine* 26.6 (2020), pp. 807–808.

[311] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. “Striving for Simplicity: The All Convolutional Net”. In: *International Conference on Learning Representations*. 2015.

[312] J. Staats, A. Divekar, J. P. McCoy, and H. T. Maecker. “Guidelines for Gating Flow Cytometry Data for Immunological Assays”. In: *Immunophenotyping*. Vol. 2032. Springer New York, 2019, pp. 81–104.

[313] K. Stark and S. Massberg. “Interplay between Inflammation and Thrombosis in Cardiovascular Pathology”. In: *Nature Reviews Cardiology* 18.9 (2021), pp. 666–682.

[314] E. Štrumbelj and I. Kononenko. “Explaining Prediction Models and Individual Predictions with Feature Contributions”. In: *Knowledge and Information Systems* 41.3 (2014), pp. 647–665.

[315] H. Su, Z. Yin, S. Huh, and T. Kanade. “Cell Segmentation in Phase Contrast Microscopy Images via Semi-Supervised Classification over Optics-Related Features”. In: *Medical Image Analysis* 17.7 (2013), pp. 746–765.

[316] D. Sui, W. Liu, J. Chen, C. Zhao, X. Ma, M. Guo, and Z. Tian. “A Pyramid Architecture-Based Deep Learning Framework for Breast Cancer Detection”. In: *BioMed Research International* 2021 (2021), pp. 1–10.

[317] A. Suissa-Peleg, D. Haehn, S. Knowles-Barley, V. Kaynig, T. R. Jones, A. Wilson, R. Schalek, J. W. Lichtman, and H. Pfister. “Automatic Neural Reconstruction from Petavoxel of Electron Microscopy Data”. In: *Microscopy and Microanalysis* 22.S3 (2016), pp. 536–537.

[318] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker. “An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success”. In: *npj Digital Medicine* 3.1 (2020), p. 17.

[319] S. Suzuki and K. Be. “Topological Structural Analysis of Digitized Binary Images by Border Following”. In: *Computer Vision, Graphics, and Image Processing* 30.1 (1985), pp. 32–46.

[320] J. Tack, S. Mo, J. Jeong, and J. Shin. “CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances”. In: *Advances in neural information processing systems* 33 (2020), pp. 11839–11852.

[321] F. Al-Tahhan, A. A. Sakr, D. A. Aladle, and M. Fares. “Improved Image Segmentation Algorithms for Detecting Types of Acute Lymphatic Leukaemia”. In: *IET Image Processing* 13.13 (2019), pp. 2595–2603.

[322] D. M. Tax and R. P. Duin. “Support Vector Data Description”. In: *Machine Learning* 54.1 (2004), pp. 45–66.

[323] M. Testi, M. Ballabio, E. Frontoni, G. Iannello, S. Moccia, P. Soda, and G. Vessio. “MLOps: A Taxonomy and a Methodology”. In: *IEEE Access* 10 (2022), pp. 63606–63618.

[324] The HDF Group. *Hierarchical Data Format, Version 5.* https://www.hdfgroup.org/HDF5/.

[325] E. Tjoa and C. Guan. “A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI”. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.11 (2021), pp. 4793–4813.

[326] A. Tomczak, S. Ilic, G. Marquardt, T. Engel, F. Forster, N. Navab, and S. Albarqouni. “Multi-Task Multi-Domain Learning for Digital Staining and Classification of Leukocytes”. In: *IEEE Transactions on Medical Imaging* 40.10 (2021), pp. 2897–2910.

[327] A. Tomczak, S. Ilic, G. Marquardt, T. Engel, N. Navab, and S. Albarqouni. “Digital Staining of White Blood Cells With Confidence Estimation”. In: *IEEE Transactions on Medical Imaging* 42.12 (2023), pp. 3895–3906.

[328] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty. “Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems”. In: *arXiv* Preprint (2018), p. 1806.07552.

[329] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg. “What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use”. In: *Machine Learning for Healthcare Conference.* PMLR, 2019, pp. 359–380.

[330] E. J. Topol. “High-Performance Medicine: The Convergence of Human and Artificial Intelligence”. In: *Nature Medicine* 25.1 (2019), pp. 44–56.

[331] J. D. Trolinger and M. M. Mansoor. “History and Metrology Applications of a Game-Changing Technology: Digital Holography”. In: *Journal of the Optical Society of America A* 39.2 (2022), A29.

[332] M. Ugele, M. Weniger, M. Leidenberger, Y. Huang, M. Bassler, O. Friedrich, B. Kappes, O. Hayden, and L. Richter. “Label-Free, High-Throughput Detection of P. Falciparum Infection in Sphered Erythrocytes with Digital Holographic Microscopy”. In: *Lab on a Chip* 18.12 (2018), pp. 1704–1712.

[333] M. Ugele, M. Weniger, M. Stanzel, M. Bassler, S. W. Krause, O. Friedrich, O. Hayden, and L. Richter. “Label-Free High-Throughput Leukemia Detection by Holographic Microscopy”. In: *Advanced Science* 5.12 (2018), p. 1800761.

[334] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel. “Medical Transformer: Gated Axial-Attention for Medical Image Segmentation”. In: *Medical Image Computing and Computer Assisted Intervention.* Vol. 12901. Springer International Publishing, 2021, pp. 36–46.

[335] S. Varma and J. Voldman. “A Cell-Based Sensor of Fluid Shear Stress for Microfluidics”. In: *Lab on a Chip* 15.6 (2015), pp. 1563–1573.

[336] A. Vellido. “The Importance of Interpretability and Visualization in Machine Learning for Applications in Medicine and Health Care”. In: *Neural Computing and Applications* 32.24 (2020), pp. 18069–18083.

[337] V. Venkatesh and H. Bala. “Technology Acceptance Model 3 and a Research Agenda on Interventions”. In: *Decision Sciences* 39.2 (2008), pp. 273–315.

[338] Vidhya Kamakshi and N. C. Krishnan. “Explainable Image Classification: The Journey So Far and the Road Ahead”. In: *AI* 4.3 (2023), pp. 620–651.

[339] J. Vuorte, S.-E. Jansson, and H. Repo. “Evaluation of Red Blood Cell Lysing Solutions in the Study of Neutrophil Oxidative Burst by the DCFH Assay”. In: *Cytometry* 43.4 (2001), pp. 290–296.

[340] S. Wachter, B. Mittelstadt, and C. Russell. “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR”. In: *Harvard Journal of Law & Technology* 31.1 (2017), pp. 841–887.

[341] H. Walach, T. Falkenberg, V. Fønnebø, G. Lewith, and W. B. Jonas. “Circular Instead of Hierarchical: Methodological Principles for the Evaluation of Complex Interventions”. In: *BMC Medical Research Methodology* 6.1 (2006), p. 29.

[342] J. Wang and T. Lukasiewicz. “Rethinking Bayesian Deep Learning Methods for Semi-Supervised Volumetric Medical Image Segmentation”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2022, pp. 182–190.

[343] K. Wang, Y. Li, Q. Kemao, J. Di, and J. Zhao. “One-Step Robust Deep Learning Phase Unwrapping”. In: *Optics Express* 27.10 (2019), p. 15100.

[344] Z. J. Wang, A. J. Walsh, M. C. Skala, and A. Gitter. “Classifying T Cell Activity in Autofluorescence Intensity Images with Convolutional Neural Networks”. In: *Journal of Biophotonics* 13.3 (2020), e201960050.

[345] H. Waters. “New $10 Million X Prize Launched for Tricorder-Style Medical Device”. In: *Nature Medicine* 17.7 (2011), pp. 754–754.

[346] M. Wazid, A. K. Das, N. Mohd, and Y. Park. “Healthcare 5.0 Security Framework: Applications, Issues and Future Research Directions”. In: *IEEE Access* 10 (2022), pp. 129429–129442.

[347] D. S. Weld and G. Bansal. “The Challenge of Crafting Intelligible Intelligence”. In: *Communications of the ACM* 62.6 (2019), pp. 70–79.

[348] A. M. Wendelboe and G. E. Raskob. “Global Burden of Thrombosis: Epidemiologic Aspects”. In: *Circulation Research* 118.9 (2016), pp. 1340–1347.

[349] C. S. Wickramasinghe, D. L. Marino, J. Grandio, and M. Manic. “Trustworthy AI Development Guidelines for Human System Interaction”. In: *13th International Conference on Human System Interaction*. IEEE, 2020, pp. 130–136.

[350] B. L. C. Wright, J. T. F. Lai, and A. J. Sinclair. “Cerebrospinal Fluid and Lumbar Puncture: A Practical Review”. In: *Journal of Neurology* 259.8 (2012), pp. 1530–1545.

[351] Y. Wu, Y. Rivenson, Y. Zhang, Z. Wei, H. Günaydin, X. Lin, and A. Ozcan. “Extended Depth-of-Field in Holographic Imaging Using Deep-Learning-Based Autofocusing and Phase Recovery”. In: *Optica* 5.6 (2018), p. 704.

[352] L. Wynants et al. “Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal”. In: *BMJ* (2020), p. m1328.

[353] F. Xing, Y. Xie, H. Su, F. Liu, and L. Yang. “Deep Learning in Microscopy Image Analysis: A Survey”. In: *IEEE Transactions on Neural Networks and Learning Systems* 29.10 (2018), pp. 4550–4568.

[354] F. Xing and L. Yang. “Robust Nucleus/Cell Detection and Segmentation in Digital Pathology and Microscopy Images: A Comprehensive Review”. In: *IEEE Reviews in Biomedical Engineering* 9 (2016), pp. 234–263.

[355] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu. "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges". In: *Natural Language Processing and Chinese Computing*. Vol. 11839. Springer International Publishing, 2019, pp. 563–574.

[356] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: *Proceedings of the 32nd International Conference on Machine Learning*. 2015, pp. 2048–2057.

[357] G. Yang, Q. Ye, and J. Xia. "Unbox the Black-Box for the Medical Explainable AI via Multi-Modal and Multi-Centre Data Fusion: A Mini-Review, Two Showcases and Beyond". In: *Information Fusion* 77 (2022), pp. 29–52.

[358] F. Yi, I. Moon, and B. Javidi. "Automated Red Blood Cells Extraction from Holographic Images Using Fully Convolutional Neural Networks". In: *Biomedical Optics Express* 8.10 (2017), p. 4466.

[359] J. Yoon, Y. Jo, M.-h. Kim, K. Kim, S. Lee, S.-J. Kang, and Y. Park. "Identification of Non-Activated Lymphocytes Using Three-Dimensional Refractive Index Tomography and Machine Learning". In: *Scientific Reports* 7.1 (2017), p. 6654.

[360] K. Yu, S. Berkovsky, D. Conway, R. Taib, J. Zhou, and F. Chen. "Do I Trust a Machine? Differences in User Trust Based on System Performance". In: *Human and Machine Learning*. Springer International Publishing, 2018, pp. 245–264.

[361] E. Yuan, M. Matusiak, K. Sirinukunwattana, S. Varma, Ł. Kidziński, and R. West. "Self-Organizing Maps for Cellular In Silico Staining and Cell Substate Classification". In: *Frontiers in Immunology* 12 (2021), p. 765923.

[362] M. D. Zeiler and R. Fergus. "Visualizing and Understanding Convolutional Networks". In: *European Conference on Computer Vision*. Vol. 8689. Springer International Publishing, 2014, pp. 818–833.

[363] M. D. Zeiler, G. W. Taylor, and R. Fergus. "Adaptive Deconvolutional Networks for Mid and High Level Feature Learning". In: *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2018–2025.

[364] Q.-S. Zhang and S.-C. Zhu. "Visual Interpretability for Deep Learning: A Survey". In: *Frontiers of Information Technology & Electronic Engineering* 19.1 (2018), pp. 27–39.

[365] Y.-Y. Zhang, J.-C. Wu, R. Hao, S.-Z. Jin, and L.-C. Cao. "Digital Holographic Microscopy for Red Blood Cell Imaging". In: *Acta Physica Sinica* 69.16 (2020), p. 164201.

[366] Y. Zhang, Y. Weng, and J. Lund. "Applications of Explainable Artificial Intelligence in Diagnosis and Surgery". In: *Diagnostics* 12.2 (2022), p. 237.

[367] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song. "The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2020, pp. 250–258.

[368] C. Zhou and R. C. Paffenroth. "Anomaly Detection with Robust Deep Autoencoders". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 665–674.

[369] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. “UNet++: A Nested U-Net Architecture for Medical Image Segmentation”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Vol. 11045. Springer International Publishing, 2018, pp. 3–11.

[370] H. Zola. “CD Molecules 2005: Human Cell Differentiation Molecules”. In: *Blood* 106.9 (2005), pp. 3123–3126.

Related Publications

[A] O. Hayden, J. Erber, S. Rasch, T. Lahmer, S. Röhrl, and C. Klenk. "Detection of Cell Aggregates Using Quantitative Phase-Contrast Microscopy". WO/2023/006996. 2023.

[B] O. Hayden, C. Klenk, and S. Röhrl. "Detection of Molecular Biological Objects, Cellular Biological Objects and Cell Aggregates Using Quantitative Phase-Contrast Microscopy". WO/2023/006372A1. 2023.

[C] A. Hein, S. Rohrl, T. Grobel, M. Lengl, N. Hafez, M. Knopp, C. Klenk, D. Heim, O. Hayden, and K. Diepold. "A Comparison of Uncertainty Quantification Methods for Active Learning in Image Classification". In: *2022 International Joint Conference on Neural Networks*. IEEE, 2022, pp. 1–8.

[D] C. Klenk, J. Erber, D. Fresacher, S. Röhrl, M. Lengl, D. Heim, H. Irl, M. Schlegel, B. Haller, T. Lahmer, K. Diepold, S. Rasch, and O. Hayden. "Platelet Aggregates Detected Using Quantitative Phase Imaging Associate with COVID-19 Severity". In: *Nature Communications Medicine* 3.1 (2023), p. 161.

[E] C. Klenk, D. Fresacher, S. Röhrl, D. Heim, M. Lengl, S. Schumann, M. Knopp, K. Diepold, S. Holdenrieder, and O. Hayden. "Measurement of Platelet Aggregation in Ageing Samples and After In-Vitro Activation". In: *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies*. SciTePress, 2023, pp. 57–65.

[F] C. Klenk, S. Röhrl, M. Cortina, D. Fresacher, K. Diepold, and O. Hayden. "Blutzellaggregate Als Neue Biomarkerklasse Für POC Hämatologie Automaten". In: *5. MÜNCHNER POCT SYMPOSIUM, Neue Perspektiven Für Querschnittstechnologien Und Erweiterte Anwendungsgebiete*. Klinikum rechts der Isar, Technische Universität München, 2022, p. 68.

[G] K. Peschke et al. "Label-Free Digital Holographic Microscopy to Characterize Inter- and Intratumoral Heterogeneity in Pancreatic Cancer". In: *Gastroenterology* 162.7 (2022), S–732.

[H] S. Röhrl, M. Ugele, C. Klenk, D. Heim, O. Hayden, and K. Diepold. "Autoencoder Features for Differentiation of Leukocytes Based on Digital Holographic Microscopy (DHM)". In: *Computer Aided Systems Theory – EUROCAST*. Vol. 12014. Springer International Publishing, 2020, pp. 281–288.

[I] M. Ugele, C. Klenk, D. Heim, S. Röhrl, F. Mehta, N. Vejzagic, K. Peschke, K. Diepold, C. Prazeres Da Costa, M. Reichert, M. Meissner, K. Götze, and O. Hayden. "Label-Free Biomarker Information by High-Throughput Holographic Microscopy to Support Detection of Cancer and Neglected Tropical Diseases". In: *Optical Methods for Inspection, Characterization, and Imaging of Biomaterials IV*. SPIE, 2019, p. 30.

Core Publications

[I] S. Röhrl, A. Hein, L. Huang, D. Heim, C. Klenk, M. Lengl, M. Knopp, N. Hafez, O. Hayden, and K. Diepold. "Outlier Detection Using Self-Organizing Maps for Automated Blood Cell Analysis". In: *International Conference on Machine Learning 2nd Workshop on Interpretable Machine Learning in Healthcare.* 2022, p. 2208.08834.

[II] S. Röhrl, L. Bernhard, M. Lengl, C. Klenk, D. Heim, M. Knopp, S. Schumann, O. Hayden, and K. Diepold. "Explainable Feature Learning with Variational Autoencoders for Holographic Image Analysis". In: *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies.* SciTePress, 2023, pp. 69–77.

[III] D. Fresacher, S. Röhrl, C. Klenk, J. Erber, H. Irl, D. Heim, M. Lengl, S. Schumann, M. Knopp, M. Schlegel, S. Rasch, O. Hayden, and K. Diepold. "Composition Counts: A Machine Learning View on Immunothrombosis Using Quantitative Phase Imaging". In: *Proceedings of the 8th Machine Learning for Healthcare Conference.* PMLR, 2023, pp. 208–229.

[IV] S. Röhrl, J. Groll, M. Lengl, S. Schumann, C. Klenk, D. Heim, M. Knopp, O. Hayden, and K. Diepold. "Towards Interpretable Classification of Leukocytes Based on Deep Learning". In: *International Conference on Machine Learning 3rd Workshop on Interpretable Machine Learning in Healthcare.* 2023, p. 2311.14485.

[V] S. Röhrl, H. Maier, M. Lengl, C. Klenk, D. Heim, M. Knopp, S. Schumann, O. Hayden, and K. Diepold. "Explainable Artificial Intelligence for Cytological Image Analysis". In: *Artificial Intelligence in Medicine.* Springer Nature Switzerland, 2023, pp. 75–85.

[VI] S. Röhrl, C. Janotte, C. Klenk, D. Heim, M. Lengl, A. Hein, M. Knopp, O. Hayden, and K. Diepold. "Rethinking Usability Heuristics for Modern Biomedical Interfaces". In: *Proceedings of the 16th International Conference on Advances in Computer-Human Interactions.* IARIA, 2023, pp. 77–84.

Core Publication I

Outlier Detection using Self-Organizing Maps for Automated Blood Cell Analysis

Stefan Röhrl, Alice Hein, Lucie Huang, Dominik Heim, Christian Klenk, Manuel Lengl, Martin Knopp, Nawal Hafez, Oliver Hayden, Klaus Diepold

Summary This paper presents a new approach to quality control in holographic blood cell imaging. It emphasizes the importance of implementing effective quality assurance and outlier detection methods to ensure stable data processing. The key contribution is the use of Self-Organizing Maps (SOMs) for outlier detection, which dynamically use morphological features of leukocytes in a data-driven manner. The evaluation shows that the SOM achieves an accuracy rate of 99.6% in identifying outliers, providing advantages in forming clusters corresponding to specific types of defects. This method offers automated clustering of leukocytes, aiding in early bias detection and facilitating the identification of relevant cell aggregates, such as those associated with COVID-19 evaluation. While not replacing specialized analysis, this approach demonstrates high accuracy and robustness, making it suitable for quality control in holographic blood cell imaging.

Own Contributions

- Initiating the research idea in cooperation with Alice Hein
- Gathering of related work and assessing the scientific context of this work
- Writing the complete manuscript and designing the figures
- Curating the dataset and planning the experiments
- Creating and engineering the feature extraction algorithms
- Revising code, validating experiments, and discussing the results
- Defending the work and integrating the reviewer feedback

S. Röhrl, A. Hein, L. Huang, D. Heim, C. Klenk, M. Lengl, M. Knopp, N. Hafez, O. Hayden, and K. Diepold. “Outlier Detection Using Self-Organizing Maps for Automated Blood Cell Analysis”. In: *International Conference on Machine Learning 2nd Workshop on Interpretable Machine Learning in Healthcare.* 2022, p. 2208.08834

Outlier Detection using Self-Organizing Maps for Automated Blood Cell Analysis

Stefan Röhrl [*1] Alice Hein [*1] Lucie Huang [*1] Dominik Heim [2] Christian Klenk [2] Manuel Lengl [1]
Martin Knopp [1 2] Nawal Hafez [1] Oliver Hayden [2] Klaus Diepold [1]

Abstract

The quality of datasets plays a crucial role in the successful training and deployment of deep learning models. Especially in the medical field, where system performance may impact the health of patients, clean datasets are a safety requirement for reliable predictions. Therefore, outlier detection is an essential process when building autonomous clinical decision systems. In this work, we assess the suitability of Self-Organizing Maps for outlier detection specifically on a medical dataset containing quantitative phase images of white blood cells. We detect and evaluate outliers based on quantization errors and distance maps. Our findings confirm the suitability of Self-Organizing Maps for unsupervised Out-Of-Distribution detection on the dataset at hand. Self-Organizing Maps perform on par with a manually specified filter based on expert domain knowledge. Additionally, they show promise as a tool in the exploration and cleaning of medical datasets. As a direction for future research, we suggest a combination of Self-Organizing Maps and feature extraction based on deep learning.

1. Introduction

Nowadays, many diseases like leukemia are diagnosed by analyzing blood samples and detecting unhealthy distributions of different types of blood cells (Mittal et al., 2022). Therefore, analysis of cellular structures make up a large part of medical laboratory tests. However, currently used gold standards of hematological analysis either have the disadvantage that they cannot classify certain cell types or are associated with a high manual effort (Meintker et al., 2013; Filby, 2016). Computer vision and machine learning (ML) in combination with contrast-rich digital holographic microscopy has the potential to perform such hematological analyses in a more cost effective, flexible and faster way (Jo et al., 2018).

Unfortunately, during the process of data collection, outliers such as defocused cells, duplets and debris may occur due to activation, apoptosis, and aggregation of cells or insufficient flow focusing. In the training stage, including these outliers in one's dataset may deteriorate model performance, since there is also no industrial grade calibrator for this holographic flow cytometry assay. In a production environment, outliers may even pose a safety issue if the model cannot reliably recognize them as such, potentially leading to a wrong classification of, say, debris as an interesting event. In this work, we examine the suitability of Self-Organizing Maps (SOMs) as a tool for the detection of outliers in a dataset of holographic microscopic images of white blood cells (WBCs). We first provide some background on SOMs in Section 2 and describe our dataset and experimental setup in Section 3. Section 4 presents our results. We end with a brief discussion of related work in Section 5 and ways our approach could be expanded upon in Section 6.

2. Background

The SOM is an unsupervised artificial neural network first proposed by Teuvo Kohonen (1990) in early 1981. This dimensionality reduction technique groups data points into clusters on a 2D lattice according to their mutual similarity. The lattice space of a SOM consists of a predefined number of neurons. Each neuron has its own weight vector, which is initialized through some initialization function (e.g. principal component analysis). The weight vector of a neuron j can be described as

$$w_j = [w_{j1}, w_{j2}, ..., w_{jd}]^T,\ j = 1, 2, ..., J,$$

where J is the number of neurons and d the number of input features.

The SOM is then trained for a set number of iterations by

[*] Equal contribution [1] Chair of Data Processing, Technical University of Munich, Germany [2] Heinz-Nixdorf Chair of Biomedical Electronics, Technical University of Munich, Germany. Correspondence to: Stefan Röhrl <stefan.roehrl@tum.de>.

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML). Copyright 2022 by the author(s).

choosing an input data point $x \in \mathbb{R}^d$ from the training dataset and computing its activation distance to all other neurons. The index $c = c(x)$ of the neuron with the closest Euclidean distance to x, also called the Best Matching Unit (BMU), is determined using

$$c(x) = \underset{j}{argmin} \left\| x - w_j \right\|, \ j = 1, 2, ..., J.$$

Based on a predefined spread (e.g. standard deviation σ), a neighborhood kernel $h_{j,c(x)}(n)$ controls the update influence on the surrounding of the BMU. The weight vectors of the BMU and its neighbors w_j are then updated according to a time-variant learning rate $\alpha(n)$ using

$$w_j(n+1) = w_j(n) + \alpha(n) \cdot h_{j,c(x)}(n) \cdot (x(n) - w_j(n)),$$

where n stands for the current iteration. Algorithm 1 summarizes this process.

After successful training, the weight vectors of the SOM have adjusted to reflect the distribution of the input data in a topology-preserving manner: data points which are similar to each other in the input space are matched onto neurons close to each other in the lattice space (Kaski, 1997). This is a useful property for the detection of outliers within a large dataset, as inliers are expected to form large and dense clusters of neurons in the lattice space. Outliers on the other hand are expected to be scattered across the lattice space with a large distance from the dense clusters.

Algorithm 1 SOM training algorithm

```
initialize weight vectors $w$ of all neurons
$N \leftarrow$ *number of iterations*
$J \leftarrow$ *number of neurons*
**for** $n \leftarrow 1$ to $N$ **do**
    $x \leftarrow$ random input data point from the input dataset
    **for** $j \leftarrow 1$ to $J$ **do**
        calculate distance $d_j(n) = \|x - w_j(n)\|$
    **end for**
    calculate index for BMU $c(x) = \underset{j}{argmin}\ d_j(n)$
    determine neighborhood function $h_{j,c(x)}(n)$ based on $\sigma$ and $c(x)$
    **for** $j \leftarrow 1$ to $J$ **do**
        update weights with $w_j(n+1)$
            $= w_j(n) + \alpha(n) \cdot h_{j,c(x)}(n) \cdot (x(n) - w_j(n))$
    **end for**
**end for**
```

3. Methods

3.1. Data

The dataset used in this work[1] consists of quantitative phase images of four types of WBCs (eosinophils, lymphocytes, monocytes, and neutrophils), taken by a digital holographic microscope. These images of size 512×384 pixels contain multiple cells per image and represent their optical density. Using threshold segmentation, the raw phase images are segmented to yield single cell image patches of size 50×50 pixels. Examples can be seen in Figure 2. For this work, we used three segmented datasets:

- **Unfiltered dataset**, 447,541 images
  This dataset contains images of 41,881 eosinophils, 77,672 lymphocytes, 58,760 monocytes and 269,228 neutrophils. Since it has not been manually cleaned, there are an unknown number of inliers and outliers.
- **Inlier dataset**, 82,056 images
  This dataset was created by filtering images based on predefined thresholds for the four morphological features *optical height max*, *circularity*, *area* and *equivalent diameter* of each cell. The four classes of WBCs are balanced to 20,514 images per class.
- **Outlier dataset**, 10,136 images
  The dataset contains 352 images captured with focus set 7.5 μm over the ideal focus and 803 images with focus set 15 μm over the ideal focus. 7,749 images contain high background noise and 1,232 images were captured at the border of the microfluidic channel, which leads to high interferences due to light scattering.

All segmented images were normalized to the range of the inlier dataset, and six ($d = 6$) morphological features were extracted, namely *area*, *circularity*, *equivalent diameter*, *optical height max*, *optical height variance*, and *energy* (Ugele et al., 2018; Röhrl et al., 2019).

3.2. Experiments

After preprocessing, we trained a SOM on the inlier dataset, then tested the model on the outlier and unfiltered dataset and evaluated the detected outliers and inliers. For evaluation, we used the *average quantization error*, which is the normed average of the quantization errors of all input samples, calculated using

$$E_{AQ} = \frac{1}{M} \sum_{i=1}^{M} \left\| x_i - w_c \right\| \quad \text{with } c = \underset{j}{argmin} \left\| x_i - w_j \right\|.$$

[1] All human samples were collected with informed consent and procedures approved by application 620/21 S-KK of the ethic committee of the Technical University of Munich.

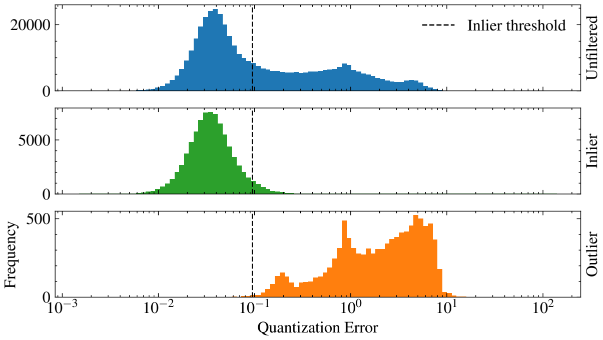


Figure 1. Quantization error distributions for the three datasets

Here, M is the size of the input dataset and j the index of the respective neuron. The smaller the average quantization error, the better a fixed-sized SOM reflects the input dataset. We defined samples with a quantization error greater than the 2σ deviation of all quantization errors as outliers.

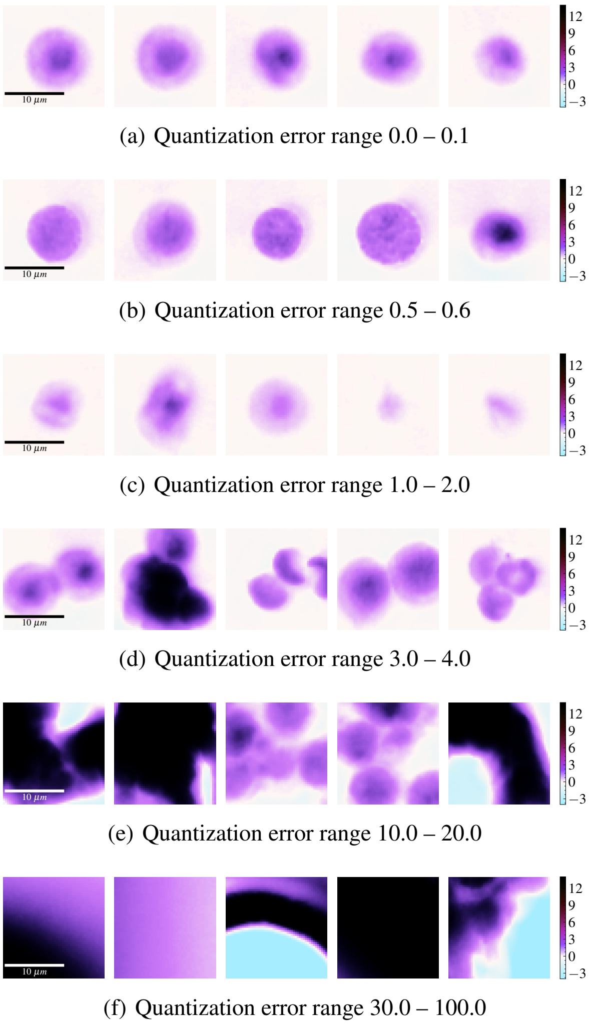
As per Kohonen's recommendation (1990), our SOM consisted of $5\sqrt{K}$ neurons, where K is the cardinality of our inlier dataset. Its shape was chosen such that its ratio of height to width equaled the ratio of the two largest eigenvalues of its autocorrelation matrix (Ponmalai & Kamath, 2019), resulting in a 65×22 lattice. The SOM was trained with a sigma of 1, learning rate of 1, hexagonal topology, gaussian neighborhood function and euclidean activation distance, as this was found to be a suitable hyperparameter configuration in preliminary tests with a 5-fold cross-validation, leading to the lowest quantization error. All experiments were implemented in Python and made use of the Scikit-learn[2], OpenCV[3], TensorFlow[4], Keras[5], and MiniSOM[6] libraries.

4. Results

The middle graph of Figure 1 shows the distribution of all inlier quantization errors. As can be seen, most of the errors fall within a small range around 0.04, which indicates that the SOM was trained to fit the inlier dataset well. According examples for inliers are displayed in Figure 2(a). Next, we evaluate the quantization errors of the outlier dataset. If the SOM worked perfectly, all errors should be greater than the inlier threshold.

This is confirmed by Figure 1 (bottom), where 99.6% of all data are correctly detected as outliers. Finally, we tested the SOM on the unfiltered dataset consisting of an unknown number of unlabeled inliers and outliers. As expected, most of the quantization errors for the unfiltered dataset lay within the threshold of 0.095, while the rest stretches out to large quantization error ranges, yielding an outlier percentage of 43.26%. That is approximately the same amount as detected with the currently used filtering method, which relies on manually specified feature thresholds based on domain expertise.

Taking a look at the inliers and outliers detected in the unfiltered dataset, we observe that in error range [0.5, 0.6], the detected outliers start to take on irregular shapes, such as too small or unclean circles. Cells in error range [1.0, 2.0] often have blurred and irregular contours. Range [3.0, 4.0] covers the case of double cells, which where mistaken for single cells in the segmentation process. Larger error ranges contain completely irregular cells or edge cases like the border of the microfluidic channel or air bubbles.



(a) Quantization error range 0.0 – 0.1

(b) Quantization error range 0.5 – 0.6

(c) Quantization error range 1.0 – 2.0

(d) Quantization error range 3.0 – 4.0

(e) Quantization error range 10.0 – 20.0

(f) Quantization error range 30.0 – 100.0

Figure 2. Examples of inliers and outliers detected by the SOM in the unfiltered dataset

[2] https://scikit-learn.org

[3] https://opencv.org

[4] https://tensorflow.org

[5] https://keras.io

[6] https://github.com/JustGlowing/minisom

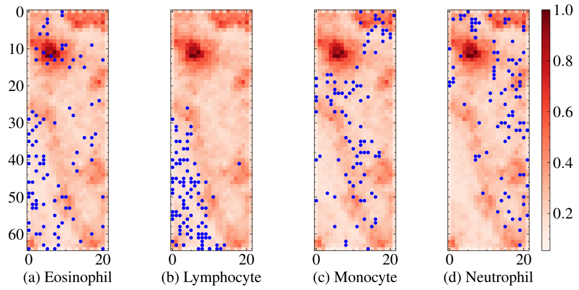


Figure 3. Positions of four inlier classes (a) eosinophil, (b) lymphocyte, (c) monocyte and (d) neutrophil on SOM distance map

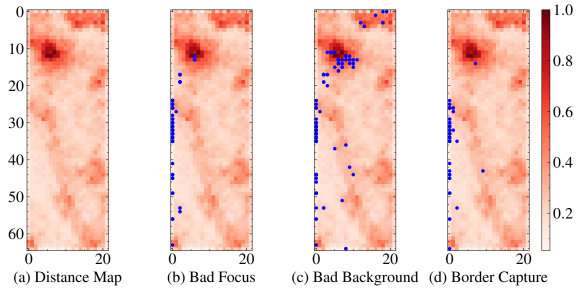


Figure 4. SOM distance map (a) and positions of outlier types (b) bad focus, (c) bad background and (d) border capture

A further evaluation technique we used was to inspect where on the SOM distance map the inliers and outliers were positioned. A distance map shows the distance of each neuron to its closest neighbors. The lighter the neuron, the smaller the distance to its neighbor neurons. Figure 3 displays the distance map as the aforementioned 65×22 lattice as a background pattern. Each sub-figure shows that almost all inliers were plotted in light regions of the distance map, confirming the assumption that clusters with many neurons close to each other represent dense inlier classes. Additionally, the winning neurons of input data points from the same white blood cell classes formed clusters, suggesting that the SOM had not only learned to distinguish inliers and outliers, but also to some extent the four different classes of inliers.

This pattern is also confirmed by Figure 4, which plots the positions of different types of outliers on the distance map. In contrast to the inlier data points, outliers tend to be positioned in darker, that is, less dense regions, or at the edge of the SOM.

5. Related Work

Out-Of-Distribution (OOD) detection methods provide important safety mechanisms to prevent real-world systems from failing when confronted with anomalous data and have thus been the focus of much research. The three main categories of OOD detection approaches are classification-based, nearest neighbor-based, and clustering-based techniques (Chandola et al., 2009), where SOMs can be said to belong to the latter category. Previous applications of SOMs for cluster-based OOD identification include intrusion detection (Labib & Vemuri, 2002), fault detection (Emamian et al., 2000) and fraud detection (Brockett et al., 1998). While this work uses a low-dimensional feature representation of the input objects, it is also possible to apply SOMs directly on pixel values as shown by Penn (2002) on hyperspectral imagery data. Xiao et al. (2018) extend this idea and combine the SOM with a deep neural network to obtain a *change graph* in synthetic aperture radar images used for environmental monitoring. In the domain of tissue cell analysis *in silico*, Yuan et al. (2021) presented a SOM for segmentation and classification. Rahmat et al. (2018) successfully demonstrate the morphological analysis of red blood cells which encourages the adaption of SOMs to WBCs in this work.

6. Discussion and Conclusion

In this work, we confirmed the suitability of a SOM-based OOD detection approach on a dataset of holographic blood cell images. The SOM reached an accuracy of 99.69% on a test set of outliers created through physical manipulations during the imaging or the sample preparation. When applied on a dataset with an unknown number of inliers and outliers, it performed similarly to a filter based on manually specified feature thresholds. Therefore, it spares the medical experts time consuming and expensive manual labor. The SOM-based method also enabled the observation of different types of outliers in different ranges of quantization errors, such as duplets and edge cases. This was not possible using the current filtering method. Hence, we achieved a more generalizable and robust approach to clean the vast holographic flow cytometry datasets. In addition, the optimized SOM could be used to distinguish between different classes of inliers, visible as separate clusters on the distance map.

However, the SOM still relies on extensive pre-processing to extract selected features. A next step would therefore be to take advantage of recent advances in deep learning by combining convolutional neural networks for feature extraction with SOMs for dimensionality reduction and OOD detection. Given the SOM's clustering abilities, we also envision further applications such as dataset exploration and efficient data annotation by labelling entire parts of the SOM rather than individual examples.

Acknowledgement The authors would like to especially honor the contributions of D. Heim for the sample preparation and recording of the measurements and L. Huang for the software implementation and experiments. This research was funded by the German Federal Ministry for Education and Research (BMBF) with the funding ID ZN 01 | S17049.

References

Brockett, P. L., Xia, X., and Derrig, R. A. Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud. *Journal of Risk and Insurance*, 65(2):245–274, 1998.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3): 1–58, 2009.

Emamian, V., Kaveh, M., and Tewfik, A. H. Robust Clustering of Acoustic Emission Signals Using the Kohonen Network. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pp. 3891–3894. IEEE Computer Society, 2000.

Filby, A. Sample preparation for flow cytometry benefits from some lateral thinking. *Cytometry Part A: the journal of the International Society for Analytical Cytology*, 89 (12):1054–1056, 2016.

Jo, Y., Cho, H., Yun Lee, S., Choi, G., Kim, G., Min, H.-s., and Park, Y. Quantitative Phase Imaging and Artificial Intelligence: A Review. *IEEE Journal of Selected Topics in Quantum Electronics*, 25(1):1–14, 2018.

Kaski, S. *Data Exploration Using Self-organizing Maps*. Acta polytechnica Scandinavica. Finnish Academy of Technology, 1997.

Kohonen, T. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

Labib, K. and Vemuri, R. NSOM: A real-time network-based intrusion detection system using self-organizing maps. *Networks and Security*, 21(1), 2002.

Meintker, L., Ringwald, J., Rauh, M., and Krause, S. W. Comparison of automated differential blood cell counts from Abbott Sapphire, Siemens Advia 120, Beckman Coulter DxH 800, and Sysmex XE-2100 in normal and pathologic samples. *American journal of clinical pathology*, 139(5):641–650, 2013.

Mittal, A., Dhalla, S., Gupta, S., and Gupta, A. Automated analysis of blood smear images for leukemia detection: a comprehensive review. *ACM Computing Surveys (CSUR)*, 2022.

Penn, B. Using self-organizing maps for anomaly detection in hyperspectral imagery. In *Proceedings, IEEE Aerospace Conference*, volume 3, pp. 1531–1535, 2002.

Ponmalai, R. and Kamath, C. Self-Organizing Maps and Their Applications to Data Analysis. Technical report, Lawrence Livermore National Lab. (LLNL), Livermore, CA (United States), 2019.

Rahmat, R. F., Wulandari, F. S., Faza, S., Muchtar, M. A., and Siregar, I. The morphological classification of normal and abnormal red blood cell using self organizing map. *IOP Conference Series: Materials Science and Engineering*, 308:012015, 2018.

Röhrl, S., Ugele, M., Klenk, C., Heim, D., Hayden, O., and Diepold, K. Autoencoder Features for Differentiation of Leukocytes based on Digital Holographic Microscopy (DHM). In *Computer Aided Systems Theory - EUROCAST*, pp. 281–288, 2019.

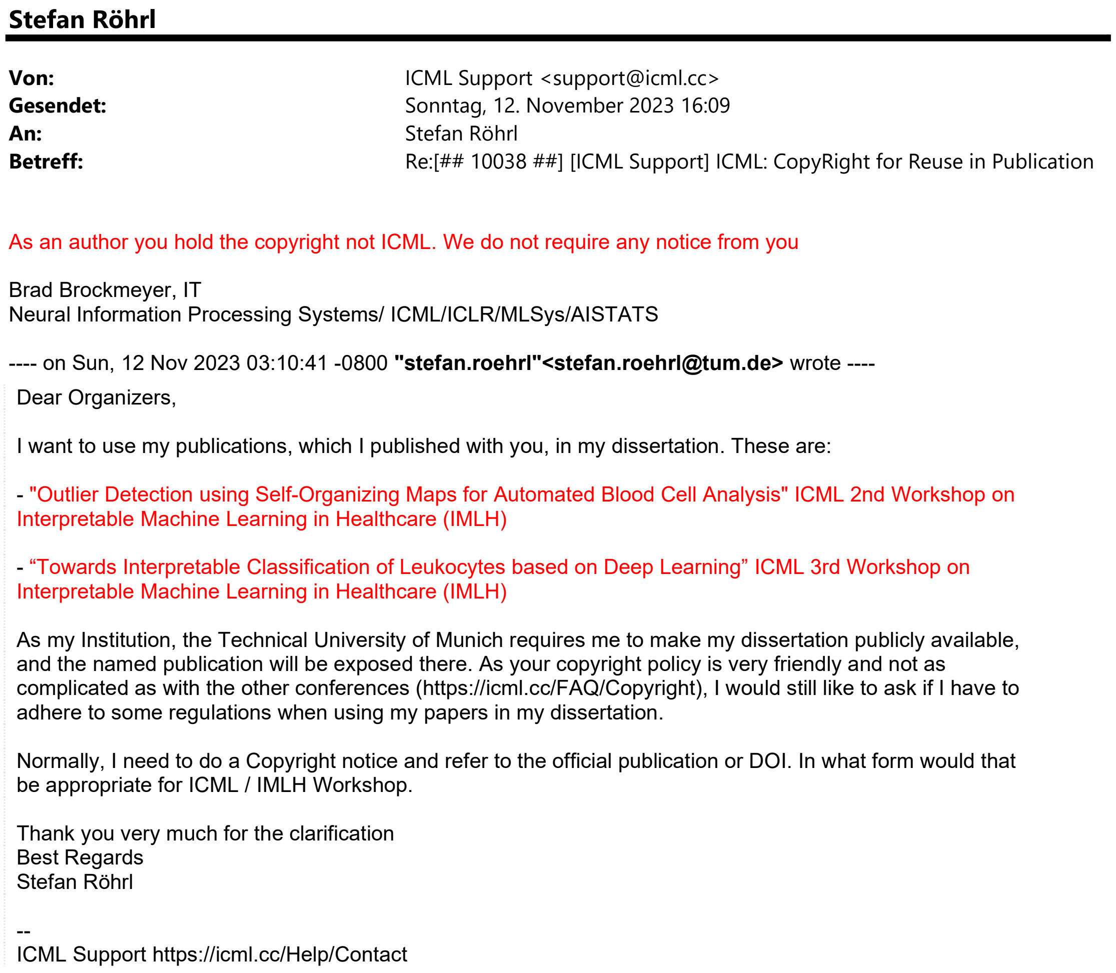
Ugele, M., Weniger, M., Stanzel, M., Bassler, M., Krause, S. W., Friedrich, O., Hayden, O., and Richter, L. Label-free high-throughput leukemia detection by holographic microscopy. *Advanced Science*, 5(12):1800761, 2018.

Xiao, R., Cui, R., Lin, M., Chen, L., Ni, Y., and Lin, X. SOMDNCD: Image Change Detection Based on Self-Organizing Maps and Deep Neural Networks. *IEEE Access*, 6:35915–35925, 2018.

Yuan, E., Matusiak, M., Sirinukunwattana, K., Varma, S., Kidziński, L., and West, R. Self-organizing maps for cellular in silico staining and cell substate classification. *Frontiers in Immunology*, 12:765923, 2021.

Copyright

No permission from the publisher is necessary as the authors hold the exclusive copyright. No rights were transferred to ICML.



Stefan Röhrl

Von:	ICML Support <support@icml.cc>
Gesendet:	Sonntag, 12. November 2023 16:09
An:	Stefan Röhrl
Betreff:	Re:[## 10038 ##] [ICML Support] ICML: CopyRight for Reuse in Publication

As an author you hold the copyright not ICML. We do not require any notice from you

Brad Brockmeyer, IT
Neural Information Processing Systems/ ICML/ICLR/MLSys/AISTATS

---- on Sun, 12 Nov 2023 03:10:41 -0800 **"stefan.roehrl"<stefan.roehrl@tum.de>** wrote ----

> Dear Organizers,
>
> I want to use my publications, which I published with you, in my dissertation. These are:
>
> - "Outlier Detection using Self-Organizing Maps for Automated Blood Cell Analysis" ICML 2nd Workshop on Interpretable Machine Learning in Healthcare (IMLH)
>
> - "Towards Interpretable Classification of Leukocytes based on Deep Learning" ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)
>
> As my Institution, the Technical University of Munich requires me to make my dissertation publicly available, and the named publication will be exposed there. As your copyright policy is very friendly and not as complicated as with the other conferences (https://icml.cc/FAQ/Copyright), I would still like to ask if I have to adhere to some regulations when using my papers in my dissertation.
>
> Normally, I need to do a Copyright notice and refer to the official publication or DOI. In what form would that be appropriate for ICML / IMLH Workshop.
>
> Thank you very much for the clarification
> Best Regards
> Stefan Röhrl
>
> --
> ICML Support https://icml.cc/Help/Contact

Core Publication II

Explainable Feature Learning with Variational Autoencoders for Holographic Image Analysis

Stefan Röhrl, Lukas Bernhard, Manuel Lengl, Christian Klenk, Dominik Heim, Martin Knopp, Simon Schumann, Oliver Hayden, Klaus Diepold

Summary This paper presents a modified Variational Autoencoder (VAE) tailored for integrating human comprehensible features in holographic image analysis to improve communication about cellular structures. The VAE architecture facilitates interpretable representations of quantitative phase data, aiding in leukocyte classification and outlier detection. The resulting latent space enables pre-filtering based on specific cell characteristics and achieves high accuracy in focus detection, erythrocyte-leukocyte discrimination, and leukocyte subtype analysis. Furthermore, the method provides a practical tool for streamlining quality assurance and data cleaning in large datasets. Despite the potential limitations of the two-dimensional feature space, the approach promotes interdisciplinary dialogue by providing visible and understandable insights into holographic cell images.

Own Contributions

- Initiating the research idea
- Gathering of related work and assessing the scientific context of this work
- Writing the complete manuscript and designing the figures
- Planning the experiments and curating the datasets
- Revising and completing the software framework
- Validating experiments and discussion of the results
- Defending the work and integrating the reviewer feedback

S. Röhrl, L. Bernhard, M. Lengl, C. Klenk, D. Heim, M. Knopp, S. Schumann, O. Hayden, and K. Diepold. “Explainable Feature Learning with Variational Autoencoders for Holographic Image Analysis”. In: *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies*. SciTePress, 2023, pp. 69–77 Online: https://www.scitepress.org/publishedPapers/2023/116328/pdf/index.html

Explainable Feature Learning with Variational Autoencoders for Holographic Image Analysis

Stefan Röhrl[1*][a], Lukas Bernhard[1*][b], Manuel Lengl[1][c], Christian Klenk[2][d], Dominik Heim[2][e],
Martin Knopp[1,2][f], Simon Schumann[1][g], Oliver Hayden[2][h] and Klaus Diepold[2][i]

[1] *Chair of Data Processing, Technical University of Munich, Germany*

[2] *Heinz-Nixdorf Chair of Biomedical Electronics, Technical University of Munich, Germany*

Keywords: Quantitative Phase Imaging, Blood Cell Analysis, Machine Learning, Variational Autoencoder, Digital Holographic Microscopy, Microfluidics, Flow Cytometry.

Abstract: Digital holographic microscopy (DHM) has a high potential to be a new platform technology for medical diagnostics on a cellular level. The resulting quantitative phase images of label-free cells, however, are widely unfamiliar to the bio-medical community and lack in their degree of detail compared to conventionally stained microscope images. Currently, this problem is addressed using machine learning with opaque end-to-end models or inadequate handcrafted morphological features of the cells. In this work we present a modified version of the variational Autoencoder (VAE) to provide a more transparent and interpretable access to the quantitative phase representation of cells, their distribution and their classification. We can show a satisfying performance in the presented hematological use cases compared to classical VAEs or morphological features.

1 INTRODUCTION

Quantitative Phase Imaging (QPI) in combination with microfluidics proves to be an extremely flexible method for the analysis of cellular samples (Nguyen et al., 2022). The resulting optical tool allows researchers to investigate kinetic and morphological anomalies of cells free of labeling costs while preserving a high amount of detail. The sample presentation via a microfluidics cartridge leverages the approach to high throughput comparable to modern *flow cytometry* devices and therefore a profound statistical validity. Hence, it is not surprising that the method offers great potential in the research, diagnosis and treatment of various diseases. Recent publications in the medical fields of oncology (Lam et al., 2019; Nguyen et al., 2017) and hematology (Paidi et al., 2021; Ugele et al., 2018) are only a small subset of its capabilities. Furthermore, advances in machine learning have also been applied to this discipline, enabling automated processing, segmentation, and differentiation for a wide variety of problems (Jo et al., 2019). Besides their usage for improving the phase reconstruction technique itself (Allier et al., 2022; Paine and Fienup, 2018), big convolutional neural networks (CNNs) surpassed many classical approaches for instance segmentation and object classification. These black boxes show great performances for the retrieval and analysis of blood as well as tissue cells (Midtvedt et al., 2021; Kutscher et al., 2021).

Besides all their advantages, holography and a microfluidics system for sample presentation entails some new challenges. Performing a classical blood smear, as the gold standard for hematological analysis, ensures a defined orientation of the cells and a precise alignment in the focal plane of a microscope (Barcia, 2007). A microfluidics cartridge holds some uncertainties here. In addition, there is the absence of the usual color information and the lack to selectively label individual cell components. Of course, it is still possible to catch sight of a misaligned red blood cell, but the differentiation of white blood cell (WBC) dif-

[a] https://orcid.org/0000-0001-6277-3816
[b] https://orcid.org/0000-0002-5694-0902
[c] https://orcid.org/0000-0001-8763-6201
[d] https://orcid.org/0000-0002-4664-8107
[e] https://orcid.org/0000-0001-8463-1544
[f] https://orcid.org/0000-0002-1136-2950
[g] https://orcid.org/0000-0002-7074-473X
[h] https://orcid.org/0000-0002-2678-8663
[i] https://orcid.org/0000-0003-0439-7511

[*] These authors contributed equally to this work

€This research was funded by BMBF ZN 01 | S17049.

Röhrl, S., Bernhard, L., Lengl, M., Klenk, C., Heim, D., Knopp, M., Schumann, S., Hayden, O. and Diepold, K.
Explainable Feature Learning with Variational Autoencoders for Holographic Image Analysis.
DOI: 10.5220/0011632800003414
In *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023) - Volume 2: BIOIMAGING*, pages 69-77
ISBN: 978-989-758-631-6; ISSN: 2184-4305

ferentiation becomes impossible for the human eye.

Here, we want to enable human researches to retake control of the quality assurance in their cell selection pipeline. Also, the classification itself should become more transparent as when using huge state of the art CNNs. We present a fused approach of a lightweight variational Autoencoder in combination with a small classifier, as this technique allows an assessment of the underlying data and the decision making process on a human like level of abstraction. Unintuitive low-level features are often incapable of describing the desired behavior of an analysis pipeline. The Autoencoder approach provides an easy visual interface and the ability to present an enormous data set in a compact way. We demonstrate this behavior in different experiments involving whole blood samples, purified white blood cells as well as defocused and misaligned cells.

2 MICROSCOPY AND DATA SET

2.1 Digital Holographic Microscopy

A digital holographic microscope is capable of obtaining high-quality phase images of samples by using the principle of interference between an object beam and a reference beam. This makes it very interesting for bio-medical applications (Jo et al., 2019) as holography solves the problem of low contrast associated with typical bright-field microscopy caused by the transparent nature of most biological cells. This problem is usually overcome by staining or molecular labeling of cells, which requires time-consuming preparation and analysis (Barcia, 2007; Sahoo, 2012; Klenk et al., 2019). Phase images, on the other hand, reveal much more detailed cell structures compared to intensity images.

We use a customized differential holographic microscope by *Ovizio Imaging Systems* as shown in Figure 1. It enables label-free cell imaging of untreated blood cells in suspension. Our approach is closely related to *off-axis diffraction phase microscopy* (Dubois and Yourassowsky, 2015), but allows us to use a low-coherence light source and does not rely on a reference beam. Precise focusing of cells is performed with a 50×500 µm PMMA (polymethyl methacrylate) microfluidics channel. We are using four sheath flows to center blood cells in the channel and avoid contact with the channel walls. More detailed information about the used holographic microscope can be found in (Dubois and Yourassowsky, 2008) and (Ugele et al., 2018).

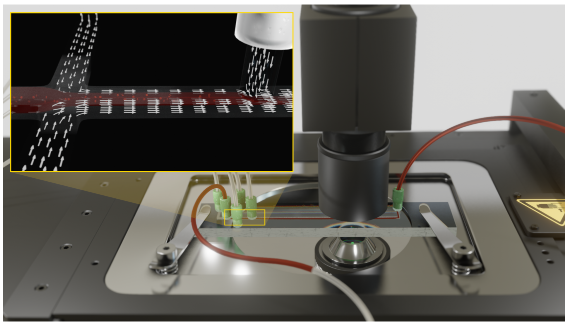
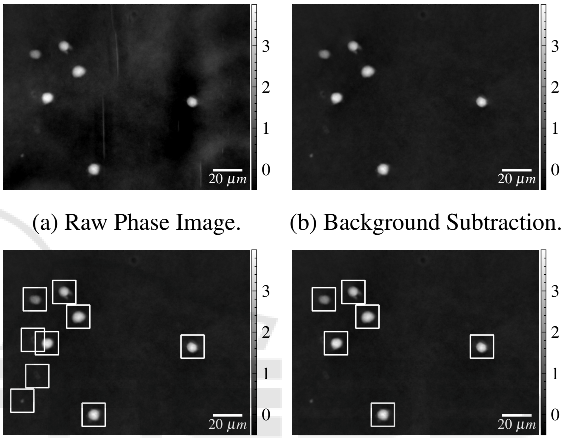


Figure 1: The PMMA chip uses hydrodynamic focusing to align the sample stream in the focal plane of the digital holographic microscope.



(a) Raw Phase Image. (b) Background Subtraction.

(c) Segmentation. (d) Filtering.

Figure 2: Several pre-processing steps are required to obtain clean image patches of individual cells.

2.2 Pre-Processing

The microscope setup provides quantitative phase images with a size of 512×384 pixels containing multiple cells. We apply several pre-processing steps to obtain isolated image patches, which contain the individual cells. Figure 2a shows an example of an unprocessed phase image of white blood cells in the microfluidics channel.

2.2.1 Background Subtraction

To remove background noise and artifacts of the microfluidics channel, background subtraction is required. The background is estimated using the median of 1,000 images, which gives much better results compared to using the mean. Due to the fixed orientation of the lens, camera, light, and microfluidics channel, the background is assumed to be static over the whole recording. As a result of background subtraction Figure 2b clearly shows a minimized expression of noise and artifacts compared to the raw image.

2.2.2 Segmentation

To find the important regions of the image that contain cells, we apply a binary thresholding to the phase images. Here, a phase shift threshold of 0.3 rad provides good results for filtering out small debris. From the resulting binary images, we extract the contours of each region of interest using the OpenCV `findContours` implementation of the algorithm proposed by (Suzuki and Abe, 1985). As Figure 2c shows, not only valid cells are identified by this rather simple method of object detection.

2.2.3 Filtering

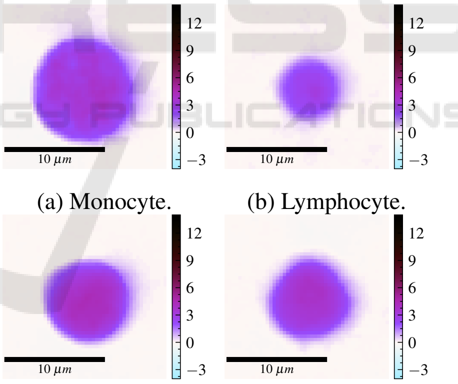
Debris and smaller cell fragments are likely to contain enough optical mass to be sensed by the thresholding procedure. Therefore, a first simple size filter is applied, so only contours covering more than 30 pixels are stored with the corresponding 48×48 pixel image area around their center. An exemplary result containing six valid cells can be seen in Figure 2d. Whereas this task could also be solved by the proposed approach, this filtering step restricts the variety of events and simplifies the convergence of the used machine learning models, allowing us to employ smaller neural networks.

2.3 Data Sets

All samples used in this work are provided by three healthy donors[1] while keeping the measurement protocols as consistent as possible. Since our microscopy approach illustrated in Section 2.1 works label-free and therefore does not require any sample preparation, the **whole blood** (2.3.2) and **defocused** (2.3.3) data set were measured within 15 minutes after blood collection. To minimize spatial coincidences of cells a 1:100 diluted blood sample is used for the robust microfluidics flow focusing. To distill single fractions of the five common types of white blood cells as a ground truth, we isolate the cells for the **leukocyte** (2.3.1) data sets. Therefore, these samples have an additional preparation time of maximum three hours. The measurement itself only takes less than two minutes for every sample resulting in more than 10,000 uncorrelated frames each. These frames are preprocessed as outlined in Section 2.2 yielding the desired phase image patches of single cells.

2.3.1 Leukocytes

Responsible for the immune defense, white blood cells represent the most interesting group for the diagnosis of diseases and the general state of human health. While making up only 1.5% of the total cell count, these cells are in focus of every modern hematology analysis device. These so-called leukocytes can be divided in five major groups. For healthy individuals, Neutrophils (62%) make up the biggest proportion, followed by Lymphocytes (30%), Monocytes (5.3%), Eosinophils (2.3%) and Basophils (0.4%) (Alberts, 2017; Young et al., 2013). We apply the isolation protocol according to (Ugele, 2019; Klenk et al., 2019): Starting from a whole blood sample[2], the leukocytes are separated from the red blood cells using *selective hypotonic water lysis* as proposed by (Vuorte et al., 2001). Remaining fragments are filtered out using an *Erythrocyte Depletion Kit*. Five different *Immunomagnetic Isolation Kits* from *Miltenyi Biotec* are then employed to obtain the individual fractions of WBCs. With this process we gathered single cell images of 77,672 Lymphocytes, 58,760 Monocytes, 41,881 Eosinophils and 269,228 Neutrophils. Note that a 100% purity of those fractions can not be ensured.



(a) Monocyte.

(b) Lymphocyte.

(c) Neutrophil.

(d) Eosinophil.

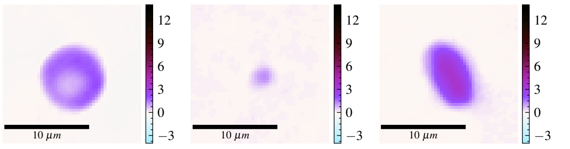
Figure 3: The quantitative phase shift is color mapped in a Giemsa stain (Barcia, 2007) fashion.

2.3.2 Whole Blood

Whole blood samples are of high value for many diagnostics as they do not require any sample preparation besides anticoagulants, which are already present in a blood tube, and are therefore very close to *in vivo* conditions. Omitting time consuming purification or staining steps facilitate insights to volatile effects in the sample. Mainly consisting of red blood

[1] All human samples were collected with informed consent and procedures approved by application 620/21 S-KK of the ethic committee of the Technical University of Munich.

[2] EDTA is used to prevent coagulation.



(a) RBC. (b) Thrombocyte. (c) Tilted RBC.

Figure 4: The whole blood samples contain besides white blood cells mainly (a) red blood cells and (b) platelets. For red blood cells the orientation (c) is crucial.

cells (erythrocytes), white blood cells (leukocytes) and platelets (thrombocytes) are only a minority in the human blood (Sender et al., 2016). Typical examples for red blood cells and platelets can be seen in Figure 4. For comparability with the white blood cells, we apply the same artificial Giemsa stain. The *viscoelastic focusing* in the channel cannot guarantee the alignment of the erythrocytes to the focal plane. E.g. a tilted red blood cell as displayed in Figure 4c cannot be used for malaria detection (Ugele, 2019).

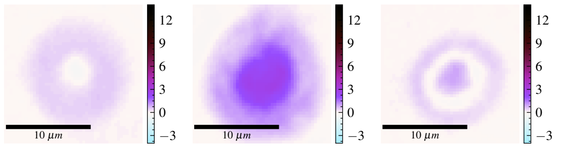
The only preparation step for all whole blood samples is a dilution of 1:100 to facilitate the segmentation of individual cells. With the current laboratory prototype and manual dilution step, results are obtained within 15 min after blood draw. (Advanced workflow integration could reduce the time-to-result even further.) The whole blood data set contains a total 126,480 images of single cells.

2.3.3 Defocused Cells

To simulate the behavior of unskilled measurement personal, a technical defect or challenges of the optical setup (Cao et al., 2022), we created different captures from whole blood with a obviously misaligned focus. We use the microscope stage to place the microfluidics channel and thereby the sample stream at different offsets above as well as below the focal plane of the objective. The misplacement ranges from -10μm to +10μm with respect to the ideal focus. Figure 5 shows these clearly defocused images which are again colored according to the previously introduced scheme. As it may happen that individual cells get out of focus even in a well calibrated setup, these images serve as training set to detect this effect. These cells are no longer usable for serious image analysis since refocusing is impossible with our optical setup. With this setting, we captured 7,269 examples of defocused cells.

3 METHODOLOGY

Dimensionality reduction is an important area of unsupervised learning. For high-dimensional data such as images, it is often necessary to reduce dimensionality as a pre-processing step. This provides deeper insight into the structure of the data and often improves the performance of classification or regression models. One of the most popular dimensionality reduction techniques is principal component analysis (PCA), which can provide deep insights into the most important features of a data set (Jolliffe and Cadima, 2016). The use of PCA implies an underlying linear system, which cannot always be guaranteed. In contrast, the *Autoencoder* approach used in this work represents an alternative, which, as a neural network, is not bound to these assumptions (Schmidhuber, 2015). As a deep-learning technique it utilizes non-linearly activated neurons which are organized in layers to encode data samples into a compressed latent space (similar to principal components) and decode this compact representation to recover the original data. The behavior and learned codes of an Autoencoder can be affected by the number of codes (size of the latent space) and hidden layers in use. It is important to note that compared to PCA, which maximizes the variance of the codes, the interpretation of the learned codes is highly dependent on the trained data set.



(a) -10 µm. (b) −5 µm. (c) +5 µm

Figure 5: This data set contains cell images which where captured with different focal offsets with respect to the ideal focal plane.

3.1 Variational Autoencoder

Variational Autoencoders. (VAEs) introduce an additional constraint to the latent space (Kingma and Welling, 2013). The encoding should not only represent the original data as well as possible, but should also follow a certain distribution (usually a Gaussian distribution). This makes the latent space continuous and allows sampling, which means we can generate artificial data by changing the value of the encodings. This generative behavior provides a deeper insight into the learned feature representation, especially when sliding over an encoding (see Figure 7), one can see the effect and its intensity of that feature on the data at the output layer (Larsen et al., 2016). The encoder is trained to encode the input data set X into a distribution $Q(z|X)$ represented by a mean vector μ_z and a standard deviation vector σ_z. This allows sampling from that distribution to obtain an encoding vector z which is fed into the decoder network to create a reconstruction $\hat{X} = P(X|z)$. Hereby, the encoder is forced to create codes following a prior distribution $P(z)$ by including the Kullback-Leibler Divergence D_{KL} of the learned distribution $Q(z|X)$ and the desired prior distribution $P(z)$ in the loss function of the VAE (Perez-Cruz, 2008). Hence, the loss

$$\mathcal{L}(X,\hat{X},z) = MSE(X,\hat{X}) + D_{KL}[Q(z|X)||P(z)] \quad (1)$$

optimizes the reconstruction error under the constraint of a Gaussian distribution.

As we work directly on image data, the use of **convolutional** layers instead of dense layers is obvious, since these proved to be state of the art in all sorts of image classification and object detection tasks over the last decade (Ciregan et al., 2012; Krizhevsky et al., 2012). This leads to an improved representation of spatial information in the VAE.

A well-known problem of VAEs are entangled codes, which means that the codes are correlated and a learned characteristic of the data is represented in more than one encoding, leading to a reduced interpretability of the latent space. Employing **β-VAEs** addresses this problem (Burgess et al., 2018) by driving the network to disentangle its encodings using an updated loss function

$$\mathcal{L}_{\beta}(X,\hat{X},z) = MSE(X,\hat{X}) + \beta D_{KL}[Q(z|X)||P(z)]. \quad (2)$$

Choosing $\beta > 1$ emphasizes the Kullback-Leibler Divergence which forces z to be even more multivariate Gaussian and consequently $\mu_z \rightarrow 0$ and $\sigma_z \rightarrow 1$. This reduces the correlation between the encodings z_i leading to three important properties (Higgins et al., 2017):

- z approximates a basis for the latent space Z
- The network is encouraged to use as few dimensions of z as possible
- The latent space is smoothed out, improving the generative behavior and allowing clearer interpretations of the information stored in the encodings.

3.2 Classifying Variational Autoencoder

The aforementioned approaches do not incorporate prior knowledge about the data samples and can learn in an unsupervised way. Therefore, the trained encoder provides not necessarily clear and distinct clusters in an interpretable manner. **Conditional Variational Autoencoders** allow to add another condition c to the encoder $Q(z|X,c)$ and decoder $P(X|z,c)$ of the VAE. This changes the latent space from a normal distribution $P(z)$ to a conditional distribution $P(z|c)$, yielding to some kind of class awareness of the encoder. Several publications showed the advantages of this architecture as a generative model (Mishra et al., 2018; Yan et al., 2016; Maaløe et al., 2016; Kingma et al., 2014). Nevertheless, this turns into chicken-egg problem for new samples, as a class label must be assigned to the unknown data point in order to be encoded correctly.

To overcome this problem we came up with a new architecture to provide the VAE additional information about labels during training while preserving the encoding and generative nature of the VAE. The **classifying VAE** (claVAE) is equipped with an additional fully connected classifier network[3] which is connected to the μ_z from the latent space as shown in Figure 6. This provides the encoder and the latent space with information about the ground truth labels of the data, so that the encoder can optimally place the data in the latent space by grouping samples of one class together (path a) while maintaining a continuous space from which we can sample (path b). The decoder is responsible for reconstructing the original image from the latent space (path c). Combining the back-propagated errors along the three paths yields a loss function

$$\mathcal{L}_{claVAE}(X,\hat{X},z,y,\hat{y}) = \mathcal{L}_{\beta}(X,\hat{X},z) + \theta\mathcal{L}_{BCE}(Y,\hat{Y}), \quad (3)$$

where θ controls the influence of the Binary Cross-Entropy loss $\mathcal{L}_{BCE}$ between the ground truth label Y and the prediction $\hat{Y}$.

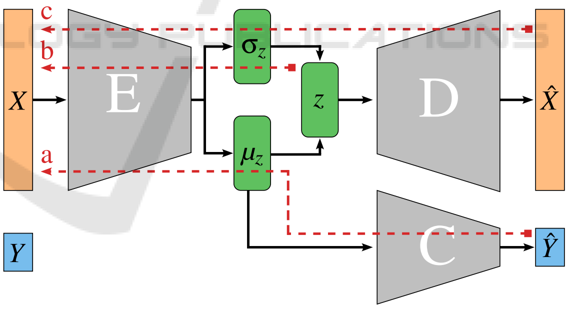


Figure 6: The claVAE architecture combines the classification error (a) the Kullback-Leibler Divergence (b) and the reconstruction error (c).

3.3 Experimental Setup

For the training of the individual models, the described data sets are divided into 60% training set (of which 20% for validation) and 40% test set for the evaluations shown later. Depending on the combination of data sets, the samples are balanced using random undersampling according to their class label.

[3] Inspired by https://www.datacamp.com/tutorial/autoencoder-classifier-python accessed Jan 16, 2020

The neural network architecture is kept constant between all models: The encoder (E) consists of four convolutional layers with max pooling, two dropout layers with a dropout rate of 0.25 and two dense layers connected to z. The decoder (D) is implemented with three dense layers to increase the dimensionality of the bottleneck z and adapt it to five subsequent *transpose convolutional* layers. The claVAE is additionally equipped with a small dense classifier network (C) consisting of three hidden layers attached to z and a *softmax* layer as output. The parameters $\beta = 0.1$ and $\theta = 1$ to weight the components of the loss function are chosen via a grid search and visual inspection of the latent space. We could chose to encode more information by increasing the dimensions of the latent space, striving for a better classification accuracy. Accordingly, Figure 7 shows different kinds of characteristics stored in each additional dimension of $z \in \mathbb{R}^3$. Though we keep the latent space two-dimensional to preserve its easy visualization and clarity.

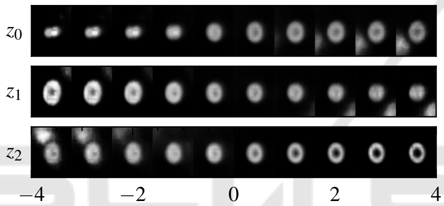


Figure 7: A three-dimensional latent space can encode more details of the input data. Here, one component of the latent vector z_i is varied, while the others z_j and z_k are kept at zero.

4 RESULTS

4.1 Overview

A typical workflow in a new project starts with getting an overview. Therefore, we train the claVAE with all data and labels introduced in Section 2.3. The resulting latent space in Figure 8 resembles a map for all components of the presented blood samples, which can be easily interpreted by the human observer. Region a) contains all defocused cells or cell aggregates. These cells cannot easily be processed further in a meaningful way and can be considered as outliers for the scenarios presented in this work. Since they come in a wide variety in shape and size, it is not surprising that they occupy a large share in latent space. Regular shaped white blood cells and well-aligned red blood cells can be found in the smaller areas b) and d), respectively. The claVAE places red blood cells, which might be unusable for further analysis as they are tilted vertically, in sector c). It is visible how the approach also tries to map the concept of orientation. The last division e) contains only the smaller cells like platelets or fragments. This arrangement is quite stable over repeated iterations of training with random initialization and randomly sub-sampled data sets. The individual placement of the groups may vary or the latent space might be rotated, but it can always be used as an intuitive map to filter the cells of interest for subsequent and more detailed analysis. We see this way of pre-filtering cells as a distinct advantage over selection by morphological metrics, as it is more similar to the established gating workflow. Furthermore, it allows a discussion of this processing step on a higher level, which is more in line with human nature to make decisions, especially in this interdisciplinary context.

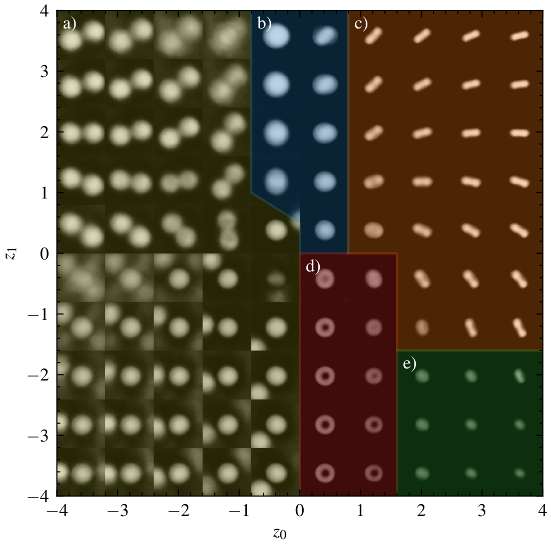


Figure 8: The spatial representation of the cells in the latent space of the claVAE can be partitioned in five groups: a) defocused and doublets cells; b) WBCs; c) tilted RBCs; d) RBCs; e) Platelets.

4.2 Focus Detection

To make sure that no defocused cells get into the data set, it is possible to sensitize claVAE to this application case. We take well-focus WBCs b) and defocused cells a) using the filters from before and provide the according labels from our training sets. Figure 9 shows the resulting distributions of the test set in the latent space. We can see the well focused cells mapped to the left whereas the defocused cells dominate the right half plane. Aggregates of two or more cells tend to be rather blurred, due to their size and the limited optical depth of the microscope, and are therefore mapped more to the right. This can be seen by

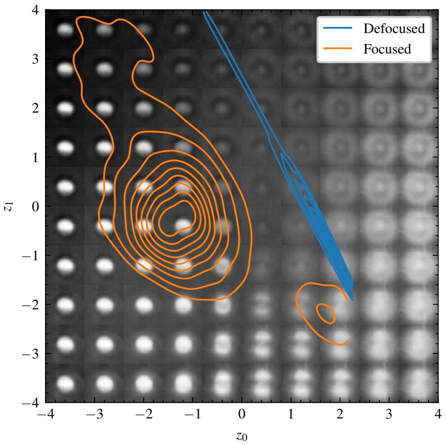


Figure 9: The density estimation of the test samples is easily separable due to the practical arrangement of the latent space.

the smaller right-bound population originating from the focused data set. The trained classifier reaches an accuracy around 96% when deciding if a cell is well-focused or not. However, with this conveniently arranged latent space, it would also be possible to use simple logistic regression or a threshold as a decision unit. Without the additional loss on the classification error, the training results from a β-VAE show a more unstable behavior and consequently support the use of the claVAE instead of a conventional variational Autoencoder.

4.3 Whole Blood Components

Considering only whole blood samples and purified white blood cells for training, we aim to achieve more detailed insights in the discrimination of RBCs and WBCs. Both classes show a rather easy separability in the latent space of this specialized claVAE. Drawn in Figure 10 the RBCs populate the top part and the WBCs are rather at the bottom.

Under the assumption that whole blood is practically RBCs, we neglected the other blood components in our labeling. Looking at the apostate group of RBCs, we hoped the claVAE would also find WBCs hiding under an incorrect ground truth label. Unfortunately, the lower orange population consists of doublet RBCs which where misplaced due to their bigger appearance. The prolonged sample preparation time and the special treatment of the purified WBCs might have changed their appearance compared to the ones in the untreated whole blood samples. However, the classification task in this space turns out to be rather simple again, as the populations are basically linearly separable. The employed classifier can differentiate both classes with an accuracy of around 97% based on their encoded representation.

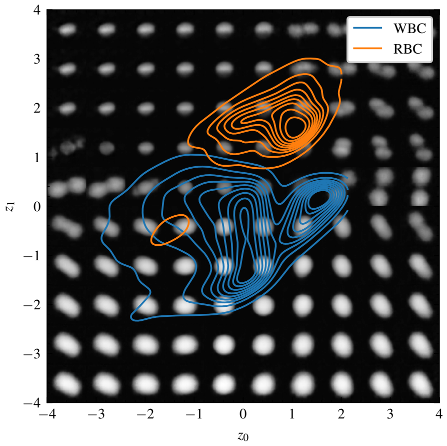


Figure 10: WBCs and RBCs mostly populate different regions of the latent space and are suitably distinguishable.

4.4 Four-Part Differential

Getting more and more into the details of hematology we now select only the available four single fractions of WBCs as a training set. The rendering of the latent space in the background of Figure 11 first suggests the distribution according to the size ratios of the individual groups. As expected, the arrangements of Neutrophils and Eosinophils overlap more clearly, while the distributions for Lymphocytes and Monocytes are better differentiated. Considering the classification performance already while training, the four groups get pulled in different directions with respect to the origin of the latent space. Using only the β-VAE the mapping looks even worse. In general, the overlapping regions lead to problems in classifi-

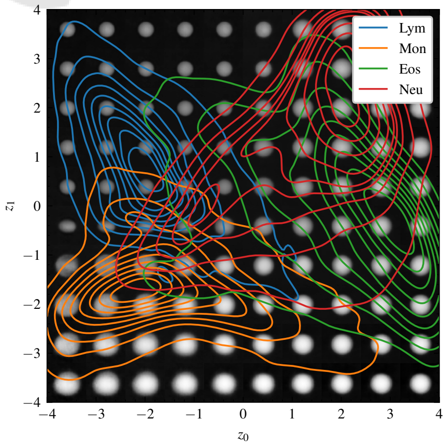
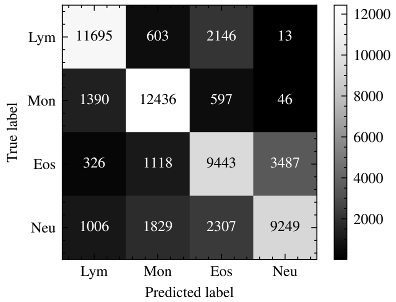


Figure 11: The four leukocyte sub-populations are drawn apart in the latent space of the claVAE but still overlap in many areas.

cation. With this latent representation, the classifier network only reaches an accuracy of 74% performing the four-part differential. Having a closer look at the confusion matrix in Figure 12, it is evident that Neutrophils and Eosinophils get mixed up. Also Lymphocytes get partly confused with Eosinophils. Note that a possible origin of this classification error might be the initial impurity of the ground truth labels themselves. The classification performance could be improved by allowing more dimensions for the latent space, since two dimensions seem to be insufficient to preserve the precise details of the rater similar leukocytes. Though, we choose not to do this as a high-dimensional space would loose its intuitiveness and would need a more complex interface for humans to access it.



True label \ Predicted label	Lym	Mon	Eos	Neu
Lym	11695	603	2146	13
Mon	1390	12436	597	46
Eos	326	1118	9443	3487
Neu	1006	1829	2307	9249

Figure 12: The confusion matrix for the four-part differential reveals the respective classification mistakes between the cell types.

5 CONCLUSION

In summary, we can say that the developed approach is well suited to obtain a compact overview of a large data set. Researchers can use it to perform robust and illustrative quality assurance as well as data cleaning, as it is more intuitive and visual than nitpicking rules of morphological features. In most of the demonstrated use cases the claVAE generates clear and separable embeddings in its latent space, which can be easily selected or classified. Its continuity and transparency gives the method the potential to be more robust against outliers and unknown data compared with large and opaque black-box approaches. In our interdisciplinary research, claVAE provides us with a basis for “eye-level” exchange, even with people from outside the domain.

Yet, the method will never be totally accurate since it would be necessary to sample the latent space at an infinitesimal level to prove its continuity. Even if the latent space appears linearly separable and easy to overlook, the employed encoder still uses a convolutional neural network, which cannot be fully explained and may hide some incontinuities. As we chose the latent space two-dimensional, we fostered its accessibility for human observers, but also limited the encoding power of the claVAE. This prevents us from resolving the subtle differences in the white blood cells needed for a classical five-part differential with sufficient accuracy.

Nevertheless, we plan to employ this non-linear method for dimensionality reduction in a zoomable user interface. Eventually, even novice users can get an intuitive overview and perform gating in visual and comprehensible manner. With further improvements of DHM in the field of label-free cell imaging, it is to be expected that phase imaging flow cytometry and will be able to reach the high accuracy required for automated hematology analysis.

ACKNOWLEDGMENTS

The authors would like to especially honor the contributions of L. Bernhard for the software implementation and experiments as well as D. Heim and C. Klenk for the sample preparation and measurements.

REFERENCES

Alberts, B. (2017). *Molecular biology of the cell*. WW Norton & Company.

Allier, C., Hervé, L., Paviolo, C., Mandula, O., Cioni, O., Pierré, W., Andriani, F., Padmanabhan, K., and Morales, S. (2022). CNN-Based Cell Analysis: From Image to Quantitative Representation. *Frontiers in Physics*, 9:848.

Barcia, J. J. (2007). The Giemsa stain: Its History and Applications. *International Journal of Surgical Pathology*, 15(3):292–296.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. (2018). Understanding Disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*.

Cao, R., Kellman, M., Ren, D., Eckert, R., and Waller, L. (2022). Self-calibrated 3D differential phase contrast microscopy with optimized illumination. *Biomedical Optics Express*, 13(3):1671–1684.

Ciregan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3642–3649. IEEE.

Dubois, F. and Yourassowsky, C. (2008). Digital holographic microscope for 3D imaging and process using it.

Dubois, F. and Yourassowsky, C. (2015). Off-axis interferometer.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*.

Jo, Y., Cho, H., Lee, S. Y., Choi, G., Kim, G., Min, H. S., and Park, Y. K. (2019). Quantitative Phase Imaging and Artificial Intelligence: A Review. *IEEE Journal of Selected Topics in Quantum Electronics*, 25(1):1–14.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Kingma, D. P., Mohamed, S., Jimenez Rezende, D., and Welling, M. (2014). Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Klenk, C., Heim, D., Ugele, M., and Hayden, O. (2019). Impact of sample preparation on holographic imaging of leukocytes. *Optical Engineering*, 59(10):102403.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kutscher, T., Eder, K., Marzi, A., Barroso, A., Schnekenburger, J., and Kemper, B. (2021). Cell Detection and Segmentation in Quantitative Digital Holographic Phase Contrast Images Utilizing a Mask Region-based Convolutional Neural Network. In *OSA Optical Sensors and Sensing Congress 2021 (AIS, FTS, HISE, SENSORS, ES)*, page JTu5A.23. Optica Publishing Group.

Lam, V. K., Nguyen, T., Phan, T., Chung, B.-M., Nehmetallah, G., and Raub, C. B. (2019). Machine learning with optical phase signatures for phenotypic profiling of cell lines. *Cytometry Part A*, 95(7):757–768.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1558–1566. PMLR.

Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. (2016). Auxiliary deep generative models. In *International Conference on Machine Learning*, pages 1445–1453. PMLR.

Midtvedt, B., Helgadottir, S., Argun, A., Pineda, J., Midtvedt, D., and Volpe, G. (2021). Quantitative digital microscopy with deep learning. *Applied Physics Reviews*, 8(1):011310.

Mishra, A., Krishna Reddy, S., Mittal, A., and Murthy, H. A. (2018). A generative model for zero shot learning using conditional variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2188–2196.

Nguyen, T. H., Sridharan, S., Macias, V., Kajdacsy-Balla, A., Melamed, J., Do, M. N., and Popescu, G. (2017). Automatic Gleason grading of prostate cancer using quantitative phase imaging and machine learning. *Journal of Biomedical Optics*, 22(3):036015.

Nguyen, T. L., Pradeep, S., Judson-Torres, R. L., Reed, J., Teitell, M. A., and Zangle, T. A. (2022). Quantitative phase imaging: Recent advances and expanding potential in biomedicine. *American Chemical Society Nano*, 16(8):11516–11544.

Paidi, S. K., Raj, P., Bordett, R., Zhang, C., Karandikar, S. H., Pandey, R., and Barman, I. (2021). Raman and quantitative phase imaging allow morpho-molecular recognition of malignancy and stages of B-cell acute lymphoblastic leukemia. *Biosensors and Bioelectronics*, 190:113403.

Paine, S. W. and Fienup, J. R. (2018). Machine learning for improved image-based wavefront sensing. *Optics Letters*, 43(6):1235–1238.

Perez-Cruz, F. (2008). Kullback-Leibler divergence estimation of continuous distributions. In *2008 IEEE International Symposium on Information Theory*, pages 1666–1670.

Sahoo, H. (2012). Fluorescent labeling techniques in biomolecules: A flashback. *Royal Society of Chemistry Advances*, 2(18):7017–7029.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS biology*, 14(8):e1002533.

Suzuki, S. and Abe, K. (1985). Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics and Image Processing*, 30(1):32–46.

Ugele, M. (2019). *High-throughput hematology analysis with digital holographic microscopy*. PhD thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU).

Ugele, M., Weniger, M., Stanzel, M., Bassler, M., Krause, S. W., Friedrich, O., Hayden, O., and Richter, L. (2018). Label-Free High-Throughput Leukemia Detection by Holographic Microscopy. *Advanced Science*, 5(12).

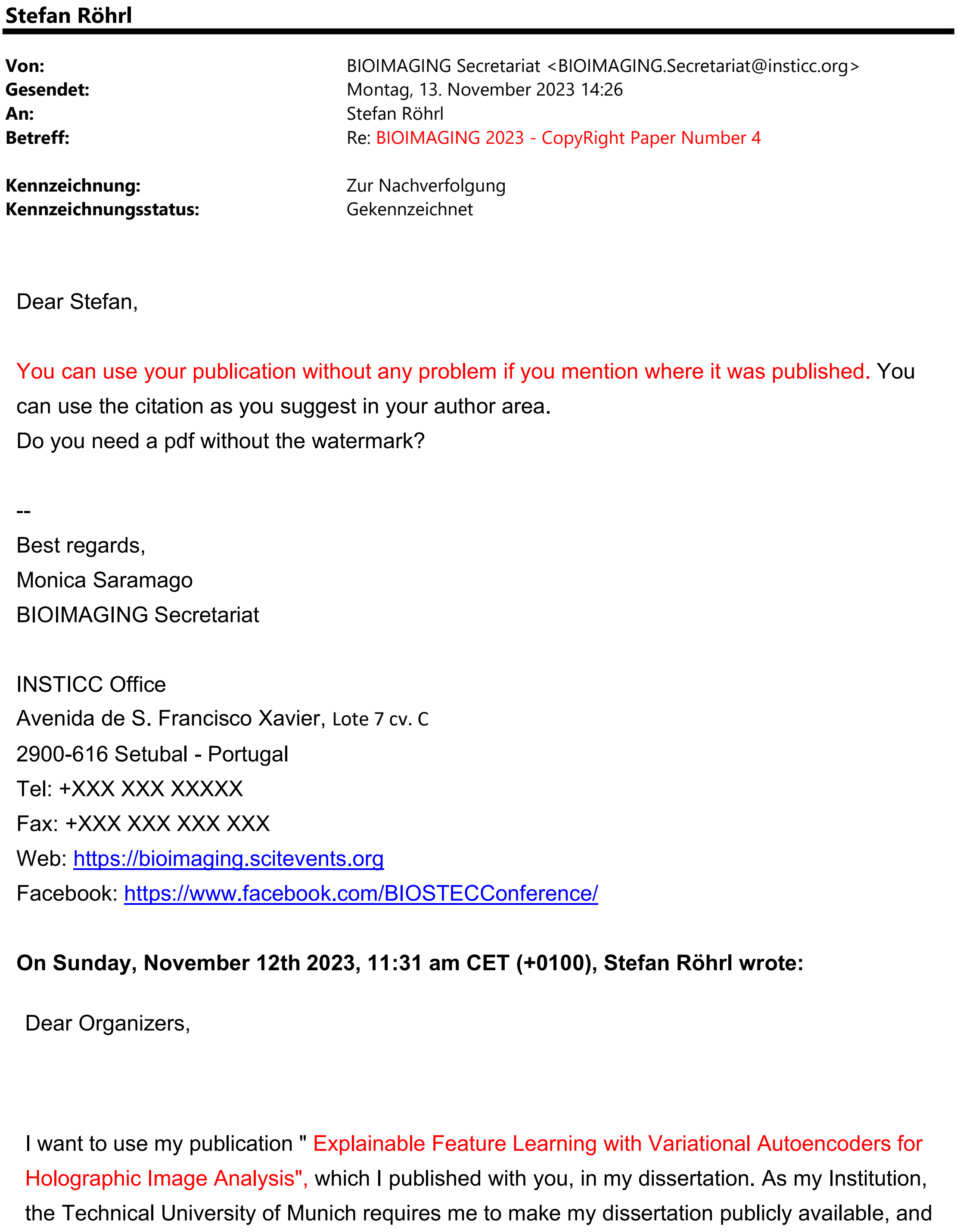
Vuorte, J., Jansson, S.-E., and Repo, H. (2001). Evaluation of red blood cell lysing solutions in the study of neutrophil oxidative burst by the DCFH assay. *Cytometry*, 43(4):290–296.

Yan, X., Yang, J., Sohn, K., and Lee, H. (2016). Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer.

Young, B., Woodford, P., and O'Dowd, G. (2013). *Wheater's functional histology E-Book: a text and colour atlas*. Elsevier Health Sciences.

Copyright

The publisher granted permission to reprint the publication in this document on November 13th, 2023.



Stefan Röhrl

Von:	BIOIMAGING Secretariat <BIOIMAGING.Secretariat@insticc.org>
Gesendet:	Montag, 13. November 2023 14:26
An:	Stefan Röhrl
Betreff:	Re: BIOIMAGING 2023 - CopyRight Paper Number 4
Kennzeichnung:	Zur Nachverfolgung
Kennzeichnungsstatus:	Gekennzeichnet

Dear Stefan,

You can use your publication without any problem if you mention where it was published. You can use the citation as you suggest in your author area.
Do you need a pdf without the watermark?

--
Best regards,
Monica Saramago
BIOIMAGING Secretariat

INSTICC Office
Avenida de S. Francisco Xavier, Lote 7 cv. C
2900-616 Setubal - Portugal
Tel: +XXX XXX XXXXX
Fax: +XXX XXX XXX XXX
Web: https://bioimaging.scitevents.org
Facebook: https://www.facebook.com/BIOSTECConference/

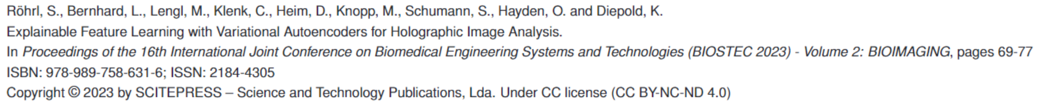
On Sunday, November 12th 2023, 11:31 am CET (+0100), Stefan Röhrl wrote:

Dear Organizers,

I want to use my publication " Explainable Feature Learning with Variational Autoencoders for Holographic Image Analysis", which I published with you, in my dissertation. As my Institution, the Technical University of Munich requires me to make my dissertation publicly available, and the named publication will be exposed there. Hence, I have the following questions, as the Copyright Transfer PDF does not specify this in detail.

- In which form am I allowed to include my publication? Is it the form you host on your server (https://www.scitepress.org/Papers/2023/116328/116328.pdf) or my Camera Ready version? Is it also possible to use your server version with your header and footer but without the**watermark** as it interferes with the images?

- What does the Copyright Notice have to look like exactly? If I use the version from your server, there is already this notice included:



Röhrl, S., Bernhard, L., Lengl, M., Klenk, C., Heim, D., Knopp, M., Schumann, S., Hayden, O. and Diepold, K.
Explainable Feature Learning with Variational Autoencoders for Holographic Image Analysis.
In *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023) - Volume 2: BIOIMAGING*, pages 69-77
ISBN: 978-989-758-631-6; ISSN: 2184-4305
Copyright © 2023 by SCITEPRESS – Science and Technology Publications, Lda. Under CC license (CC BY-NC-ND 4.0)

If this is not possible and I have to use another PDF version, where and how should the notice be placed?

Thank you very much for the clarification.

Best Regards,

Stefan Röhrl

--

Stefan Röhrl, M.Sc.

Lehrstuhl für Datenverarbeitung

Core Publication III

Composition Counts: A Machine Learning View on Immunothrombosis using Quantitative Phase Imaging

David Fresacher, Stefan Röhrl, Christian Klenk, Johanna Erber, Hedwig Irl, Manuel Lengl, Simon Schumann, Dominik Heim, Martin Knopp, Martin Schlegel, Sebastian Rasch, Oliver Hayden, Klaus Diepold

Summary This paper presents a novel processing pipeline for detecting and quantitatively analyzing blood cell aggregates, focusing on platelet and leukocyte-platelet aggregates. It investigates these formations as potential predictive biomarkers for COVID-19 and sepsis risk assessment. The pipeline integrates classical computer vision and machine learning techniques, with Mask R-CNN proving most effective for detection, segmentation, and classification. Comparative analysis reveals a trade-off between predictive and descriptive accuracy, with opaque end-to-end approaches showing higher performance. Experimental results demonstrate the effectiveness of the pipeline on clinical samples, suggesting the potential utility of these aggregates as biomarkers based on immunothrombotic events. Finally, the paper calls for further research and broader clinical trials to validate and explore factors affecting aggregate composition and diagnostic potential.

Own Contributions

- Initiating the research idea together with Christian Klenk
- Embedding the work in the scientific context together with David Fresacher
- Writing an essential part of the manuscript and designing the figures and plots
- Planning & validating the experiments and discussing the results
- Curating and preprocessing the datasets
- Defending the work and integrating the reviewer feedback with David Fresacher

D. Fresacher, S. Röhrl, C. Klenk, J. Erber, H. Irl, D. Heim, M. Lengl, S. Schumann, M. Knopp, M. Schlegel, S. Rasch, O. Hayden, and K. Diepold. "Composition Counts: A Machine Learning View on Immunothrombosis Using Quantitative Phase Imaging". In: *Proceedings of the 8th Machine Learning for Healthcare Conference*. PMLR, 2023, pp. 208–229 Online: https://proceedings.mlr.press/v219/fresacher23a.html

Composition Counts: A Machine Learning View on Immunothrombosis using Quantitative Phase Imaging

David Fresacher[*1] DAVID.FRESACHER@TUM.DE
Stefan Röhrl[*1] STEFAN.ROEHRL@TUM.DE
Christian Klenk[2] CHRISTIAN.KLENK@TUM.DE
Johanna Erber[3] JOHANNA.ERBER@MRI.TUM.DE
Hedwig Irl[3] HEDWIG.IRL@MRI.TUM.DE
Dominik Heim[2] DOMINIK.HEIM@TUM.DE
Manuel Lengl[1] M.LENGL@TUM.DE
Simon Schumann[1] SIMON.SCHUMANN@TUM.DE
Martin Knopp[1,2] MARTIN.KNOPP@TUM.DE
Martin Schlegel[3] MARTIN.SCHLEGEL@TUM.DE
Sebastian Rasch[3] SEBASTIAN.RASCH@TUM.DE
Oliver Hayden[2] OLIVER.HAYDEN@TUM.DE
Klaus Diepold[1] KLDI@TUM.DE
[1] *Chair of Data Processing,*
[2] *Heinz-Nixdorf-Chair of Biomedical Electronics, and*
[3] *University Hospital rechts der Isar*
Technical University of Munich, Germany
[*] *These two authors contributed equally to this work*

Abstract

Thrombotic complications are a leading cause of death worldwide, often triggered by inflammatory conditions such as sepsis and COVID-19, due to a close relationship between inflammation and hemostasis known as immunothrombosis. Platelet activation and leukocyte-platelet aggregation play key roles in microthrombotic events, yet there are no routine diagnostic predictive biomarkers based on these factors. This work presents a novel processing pipeline using label-free Quantitative Phase Imaging (QPI) for the detection and quantitative analysis of blood cell aggregates without sample preparation. For evaluation, we use different test scenarios and measure performance at different stages of the pipeline to gain a better understanding of the critical points. We show that, among other classical and machine learning techniques, the Mask R-CNN approach achieves the best results for detection, segmentation, and classification of cell aggregates. The method successfully identifies aggregate levels in whole blood samples and shows elevated levels in >90% of patients with COVID-19 or sepsis compared to healthy reference samples, indicating the potential of platelet and leukocyte-platelet aggregates as biomarkers for thrombotic diseases.

1. Introduction

Motivation Thrombotic conditions are considered the leading cause of mortality worldwide and the number of patients is steadily increasing, especially in developing and first world countries (Wendelboe and Raskob, 2016). Different types of thrombosis include arterial thrombosis (e.g. in the form of coronary heart disease or ischemic stroke) and venous thrombosis (e.g. in the form of deep vein thrombosis or pulmonary embolism). As thrombotic events are closely related to coagulation (blood clotting) and hemostasis in general, one of the key players are platelets (thrombocytes). Their dysfunction can have serious consequences. Thrombocyte hyperreactivity can lead to venous or arterial thrombosis and subsequently to pulmonary embolism, myocardial infarction, and stroke (Engelmann and Massberg, 2013; Nicolai et al., 2020).

Until recently, hemostasis and inflammation were thought to be completely separate physiological processes. However, recent research has shown that these two processes are intimately linked. This close relationship between coagulation and inflammation is called **immunothrombosis** (Engelmann and Massberg, 2013), which is based on the interaction of immune cells and thrombosis-related molecules. Immunothrombosis is an important defense mechanism to prevent the systemic spread of pathogens through the bloodstream by facilitating the recognition, containment, and destruction of pathogens (Stark and Massberg, 2021). However, uncontrolled immunothrombosis leads to a general risk of blood clotting, promoting the formation of microthrombi and, in the worst case, organ failure (Engelmann and Massberg, 2013).

The most recent and prominent example of an uncontrolled inflammatory response associated with thrombotic risk is COVID-19. While in most cases this infection is asymptomatic or accompanied by mild flu-like symptoms, in severe cases pulmonary complications associated with a systemic inflammatory response can occur, with potentially fatal consequences. Many recent publications indicate the occurrence of immunothrombosis with micro- and macrovascular thrombi (Nicolai et al., 2020; Schulte-Schrepping et al., 2020; Nishikawa et al., 2021; Zuo et al., 2021). Another example of the emergence of immunothrombosis is sepsis, where an initially appropriate and targeted immune response becomes generalized and harmful hyperactivation leading to organ failure (Hotchkiss et al., 2016). The appearance of activated platelets and leukocyte-platelet aggregates plays an important role in this process (Assinger et al., 2019). Due to its acute pathology, an immediate medical response is required, showing the necessity of early diagnosis (Rhodes et al., 2017).

Despite the emerging demand, no diagnostic predictive biomarker is available for routine economic diagnosis due to the highly complex pre-analytics and sample preparation required (with typically expensive antibody-based activation markers) as well as the short lifetime of cell aggregates (Finsterbusch et al., 2018). However, with the use of QPI, label-free analysis of blood cells and their aggregates becomes feasible, possibly even in a point-of-care application (Nguyen et al., 2022).

Problem statement While cell detection and classification have already been demonstrated for phase images of blood cells obtained with QPI (Ugele et al., 2018b,a; Paidi et al., 2021), the analysis of cell aggregates has proven to be more difficult due to their complex morphology, small details and short lifetime (Finsterbusch et al., 2018). In addition, their rare occurrence usually requires extensive sample preparation (Nishikawa et al., 2021).

Therefore, in this work, we design and test a data processing system that allows for the analysis of phase images of whole blood samples (obtained by QPI) for the size, number and composition of platelet and leukocyte-platelet aggregates. In addition, we evaluate relationships and correlations of these aggregate data with disease and infection using clinical samples from patients with COVID-19 and sepsis. The concept is the implementation and evaluation of a three-step pipeline for the quantitative analysis of aggregates and their components. The first step is the **detection and separation** of aggregates, specifically platelet aggregates and leukocyte-platelet aggregates, in whole blood samples. The second step is to **evaluate** the detected aggregates. This includes assessing the number of cells in an aggregate and the specific type of each cell. The last step is the integration of all the previous results and the search for **correlations with immunothrombotic diseases**. Specifically, we are analyzing sepsis and COVID-19 in comparison to healthy individuals using multiple samples from 27 subjects.

Generalizable Insights about Machine Learning in the Context of Healthcare

In our work, we present improved approaches to better understand the effects of immunothrombosis and to generate detailed information about the composition of volatile microthrombotic events. This is done under more demanding conditions because, unlike previous methods, we work label-free and with whole blood, which minimizes sample preparation. Here, we can show that our proposed machine learning pipeline is more robust to these conditions and generalizes better than the state of the art. To this aim, we not only evaluate the end-to-end performance, but also measure meaningful metrics at different points within the pipeline to better assess the behavior of the algorithms. In addition, we are introducing test scenarios to incrementally approach real-world conditions and exemplary clinical use cases in order to identify the factors that cause problems for the algorithms. We hope to lay the groundwork for using the QPI platform technology to analyze blood cell aggregates as biomarkers for predictive and individual diagnostics in subsequent clinical studies. Finally, we provide best practices for expansion into new applications that have already been shown to be related to immunothrombosis, such as hemophilia (Riedl et al., 2017), anticoagulation therapies (Lazaridis et al., 2022) or cardiovascular diseases in general (Furman et al., 2001; Allen et al., 2019).

2. Background and Related Work

Before proceeding to our proposed approach, we will look at the state of the art in observing the biomedical effects we are interested in. We will also give insights into the QPI technology and its combination with machine learning.

2.1. Medical Relevance

While the coagulation process was first discovered more than 100 years ago, in recent years coagulopathy, thrombocytopathy, and immunothrombosis have attracted increasing interest in the scientific community due to the discovery of the important role that coagulation plays in the development of cardiovascular diseases (Bhatt and Topol, 2003). Since then, a great deal of research has been conducted in this area. The role of thrombosis as an independent process of innate immunity was investigated by Engelmann and Massberg (2013), leading to the introduction of the term immunothrombosis. Successively, several researchers have shown the intricate relationship between hemostasis and inflammation (Stark and Massberg, 2021; Reyes et al., 2020; van der Poll et al., 2017).

Due to the recent emergence of a new variant of coronaviruses causing COVID-19, which has evolved into a worldwide pandemic, a great deal of research has been initiated targeting the thrombotic features of this disease. Nicolai et al. (2020) provided evidence for the involvement of immunothrombosis, while Zuo et al. (2021) discussed the process behind the formation of microthrombi, and Schulte-Schrepping et al. (2020) provided detailed insights into the systemic immune response. Most notable is the work of Nishikawa et al. (2021), who were able to show a direct link between aggregates and disease severity. Unfortunately, the analysis of the aggregates remains superficial and is limited to estimating the area of the aggregates and a fixed conversion factor for the number of platelets. No individual analysis of the aggregate components is performed, as proposed by Klenk et al. (2023). Moreover, their method requires a laborious sample preparation of up to eight hours, which, as shown, denies access to most volatile microthrombotic events (Finsterbusch et al., 2018).

Although not a new topic, sepsis has recently gained importance in scientific research. Among others Levi et al. (2013) discussed thromboembolic disease, thrombophilia, and coagulopathy in septic patients, and Assinger et al. (2019) examined the contribution of platelets to sepsis severity and outcome.

2.2. Technical Background

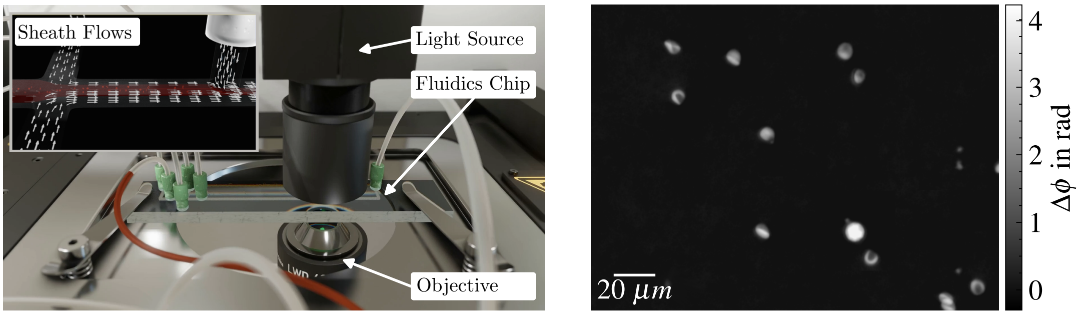
A QPI microscope uses the principle of interference to measure not only the transmission of light, but also its phase shift $\Delta\phi$, and thus to infer the optical density of cellular structures. Recently, QPI has gained relevance through its combination with machine learning, transforming cytometry into a computer vision problem (Jo et al., 2018). As a new platform technology it solves the problem of low contrast associated with typical brightfield microscopy, caused by the transparent nature of most cells. Traditionally, this would require time-consuming sample preparation, staining or genetic fluorescent labeling of cells, which can directly affect cell morphology (Barcia, 2007; Sahoo, 2012; Klenk et al., 2019).

For this project, we utilize an *off-axis diffraction phase microscope* by *Ovizio Imaging Systems* as shown in Figure 1(a). In combination with a microfluidics channel, it allows label-free cell imaging of unprocessed blood cells in suspension under near *in-vivo* conditions. A 528 nm *Super-LED Köhler illumination* provides the light source, shining on a 50 µm × 500 µm polymethyl methacrylate (PMMA) microfluidics channel that uses sheath flows to focus the sample stream within the depth of field of the 40× objective. The integrated optical setup then projects the interference patterns onto a camera sensor that captures the cells at 105 frames per second. For more information on the setup used, see Dubois and Yourassowsky (2008) and Ugele et al. (2018b).

The resulting interference patterns (hologram images) contain the intensity and phase information, that can be extracted by a reconstruction algorithm (Schnars and Jüptner, 1994). In this work, we use only phase images, as shown in Figure 1(b), because they contain most of the information about the internal structure and morphology of the observed cells.

2.3. Quantitative Phase Imaging and Machine Learning

Machine Learning methods recently entered the field of quantitative phase imaging. Besides their application in the phase reconstruction process itself (Jo et al., 2018; Paine and Fienup, 2018), the strength of these techniques, especially the Convolutional Neural



(a) Microscope components (b) A raw phase image

Figure 1: Optical setup and the resulting image

Network (CNN), lies in the segmentation and enhancement of the images as well as the differentiation of individual cells. The U-Net (Ronneberger et al., 2015; Zhang et al., 2018; Midtvedt et al., 2021) and the Mask R-CNN (He et al., 2017; Kutscher et al., 2021) show promising results for identifying and segmenting blood and tissue cells. The greatest opportunities for these technologies lie in the combination of the label-free holography and the beyond-human classification power of current neural network architectures. Bacteria (Jo et al., 2015) or leukocytes (Ozaki et al., 2019) are analyzed and classified based on their sub-cellular structures. In oncology, leukemia and the detection of its sub-types can be addressed using dimensionality reduction techniques and morphological features (Ugele et al., 2018b; Paidi et al., 2021). Nevertheless, the automated filtering and classification of cells in a high-throughput scenario like ours remains. As the phase representation of most cell types is unfamiliar to biological and medical experts, the generation of a ground truth needed for supervised learning is laborious, if not impossible (Filby, 2016; Ugele, 2019). Another obstacle is the work with whole blood samples which would provide the most convenient and simple clinical workflow without intensive and time consuming sample preparation. Reaching for rare events demands a reliable chain of filtering and outlier detection techniques, since otherwise feature extraction, dimensionality reduction, neural networks as well as classical discriminators are prone to failure (Röhrl et al., 2019).

3. Methods

In this section, we introduce the algorithms used in different steps of our proposed cell aggregate analysis pipeline. Also the metrics for their evaluation are presented.

3.1. Segmentation

The first step in the analysis of blood cell aggregates is segmentation, which allows for the identification of aggregates and their components. Numerous methods have been developed and evaluated for use in biomedical imaging, as segmentation is an essential practice. In this work, we focus on three methods, one classical segmentation method, namely watershed segmentation, and two machine learning methods, U-NET and Mask R-CNN, all of which have previously been successfully used in biomedical segmentation tasks.

Watershed Watershed segmentation is a classical method based on region growing (Beucher and Lantuéjoul, 1979; Vincent and Soille, 1991). It starts from a seed point and iteratively adds neighboring pixels in a similar way to how water floods a region. Watershed is relatively simple and fast, especially compared to machine learning based segmentation methods, yet provides good segmentation results for biomedical purposes (Ng et al., 2006).

U-NET The U-NET is a type of CNN specifically designed for semantic segmentation of biomedical images (Ronneberger et al., 2015). It is a derivative of the fully convolution network (FCN) (Shelhamer et al., 2017) designed to work with very few training images. The U-NET takes its name from its symmetrical U-shape, which consists of a contracting path and an expanding path. The contracting path uses a typical CNN architecture consisting of convolutional, rectified linear unit (ReLU), and max-pooling layers. Each step of the expansive path uses upsampling and convolution while concatenating higher resolution feature maps from the contractive path with the upsampled features. While the U-NET was originally designed for semantic segmentation only, the use of a *boundary loss* function allows its adoption for instance segmentation. In this work, an *Adam* optimizer (Kingma and Ba, 2015) was combined with a compound loss function of *cube loss* (Wang and Chung, 2018) and *boundary loss* (Kervadec et al., 2019).

Mask R-CNN The Mask R-CNN is designed for instance segmentation (He et al., 2017). It performs both object detection and object mask computation simultaneously. The Mask R-CNN is based on the Faster R-CNN (Ren et al., 2017), a region-based CNN. In this work, a ResNet50 (He et al., 2016) is used as the backbone. For the training process, a *stochastic gradient descent* optimizer with momentum was combined with a compound loss function of *classification loss*, *bounding-box loss*, and *mask loss* as defined by Ren et al. (2017) and He et al. (2017).

3.2. Classification

Unless classification has already been performed during the segmentation, the second step in our pipeline is to classify segmented cell images. This classification task considers three classes of cells relevant to blood analysis, the coagulation system, and possible diseases (erythrocytes, leukocytes, and platelets), as described in Section 2.1.

Gating Gating is a popular method in biology and medicine for manually dividing a set of cells into distinct clusters or populations (Staats et al., 2019). It typically relies on the use of software to apply a set of manually drawn gates that select regions in a 2D graphical representation of the data. This technique is most commonly used to analyze flow cytometry data. The advantages are its simplicity and explainability, since the gates are generally based on expert knowledge of the cell characteristics. This explains its widespread use in biology and medicine. However, gating shows limited suitability for high-dimensional data and is typically based on manual subjective decisions leading to high inter-observer variability (Staats et al., 2019).

Morphological features In order to successfully apply manual gating techniques to images, features must first be extracted from the images. As suggested by a Ugele et al. (2018b) or Paidi et al. (2021), a set of hand-crafted morphological features based on cell size, shape, and texture is calculated for each individual cell. The outer contour line forms the basis for features describing size and shape. Texture features are computed from the gray-level co-occurrence matrix (GLCM), which represents the distribution of co-occurring pixel values in an image and is commonly used for texture analysis in image processing (Haralick et al., 1973). These features are highly intuitive and explainable, providing excellent interpretability for expert gating. To classify the elements of cell aggregates, we only use features that are robust to changes in the shape and contour of the cell (like *homogeneity* or *optical height*), since others can experience shifting due to aggregation.

Random forest Random forest is an ensemble classification method based on decision trees proposed by Breiman (2001). A decision tree is a machine learning model that combines a series of decisions based on variable values. For random forest classification, a large number of decision trees are automatically constructed based on different fractions of the given data set. For classification of unknown samples, the average result of all trained decision trees is used. This concept reduces overfitting very effectively and works well for more complex classification tasks in high-dimensional feature spaces. Nguyen et al. (2017) successfully used a combination of morphological features and random forest classification for the grading of prostate cancer.

3.3. Evaluation Metrics

To assess the segmentation and classification quality of the proposed methods, evaluation metrics are needed. For this purpose, the Intersection over Union (IoU) is used, as usual for the evaluation of segmentation and object detection methods, and combined with the metrics precision, recall, and F_1-score.

Segmentation performance The **IoU**, or Jaccard index (Jaccard, 1912), is a popular evaluation metric used in instance segmentation and object detection. It is a measure of the similarity between two shapes, in the case of instance segmentation, the predicted region $\hat{A}$ and the ground-truth A

$$\text{IoU} = \frac{\text{area of overlap}}{\text{area of union}} = \frac{A \cap \hat{A}}{A \cup \hat{A}} \in [0, 1] \,. \tag{1}$$

The IoU is invariant to scale and therefore a very powerful metric for the evaluation of segmentation algorithms.

Detection and classification performance Since the IoU alone is only partially useful for evaluating a real-world application, a minimum IoU threshold is typically defined for an instance (or object) to be considered as correctly recognized, and evaluation metrics such as precision or recall are used. In this work, we use an IoU threshold of 0.4, since overly detailed localization and masking of the cells is not necessarily needed, while correct detection of the cell amounts and types is more important. We use the definitions given by Powers (2011) for **precision** and **recall**

$$\text{Precision} = \frac{T_p}{T_p + F_p}, \quad \text{Recall} = \frac{T_p}{T_p + F_n} \qquad \text{with} \quad \begin{array}{l} T_p : \text{true positives} \\ F_p : \text{false positives} \\ F_n : \text{false negatives} \end{array} \tag{2}$$

and the resulting harmonic mean as the so called $\mathbf{F_1}$**-Score**

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \,. \tag{3}$$

Aggregate analysis To better evaluate the segmentation performance for the specific task of detecting and counting single cells and cell aggregates, two custom metrics are used: The **aggregate composition** score evaluates whether all parts of the analyzed aggregate are correctly detected quantitatively as defined by

$$\text{AC} = \frac{1}{K}\sum_{i=1}^{K} ac_i \quad \text{with} \quad ac_i = \begin{cases} 1 & \text{if } \hat{n}_{class} = n_{class} \; \forall class \in \{\text{ery}, \text{leuko}, \text{thrombo}\} \\ 0 & \text{else} \end{cases} \tag{4}$$

where n_{class} is the number of elements of class *class* in the image patch and K is the number of images patches in the dataset. The **event type** score

$$\text{ET} = \frac{1}{K}\sum_{i=1}^{K} et_i \quad \text{with} \quad et_i = \begin{cases} 1 & \text{if } \hat{t}_i = t_i \\ 0 & \text{else} \end{cases} \tag{5}$$

and t as the type of aggregate or single cell $t \in \{$*single erythrocyte*, *multiple erythrocytes*, *single platelet*, *platelet aggregate*, *single leukocyte*, *leukocyte-platelet aggregate*$\}$ assesses if the type of aggregate or cell is correctly detected (qualitatively).

Regression analysis To evaluate possible correlations in our experiments concerning mixing ratios or activation, we employ the following models and methods: For **linear relations** we use a simple linear mapping of an independent variable x on a dependent variable y

$$\hat{y} = ax + b \qquad \text{minimizing} \quad L(x) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{6}$$

where y is the real world observation and $\hat{y}$ is the prediction of the model. For **nonlinear relations** we chose an exponential model

$$\hat{y} = a \log(x) + b \tag{7}$$

which can be fitted by an iterative estimation algorithm. Therefore, we use the *Levenberg–Marquardt algorithm* (Levenberg, 1944). To evaluate the fit of the regression models we apply the Normalized Root-Mean-Square Error (NRMSE)

$$\text{NRMSE} = \frac{\sqrt{\text{MSE}}}{y_{max} - y_{min}} \qquad \text{with} \quad \text{MSE} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{8}$$

where N is the number of observations, which allows us to compare data and models of different scales (James et al., 2013). The coefficient of determination

$$R^2 = \frac{\text{Var}(\hat{y})}{\text{Var}(y)} = 1 - \frac{\text{Var}(e)}{\text{Var}(y)} \qquad \text{with} \quad e = y - \hat{y} \tag{9}$$

provides us with a measure of the quality for the respective fit, by comparing the variance of the observed data $\text{Var}(y)$ to the variance explained by the model $\text{Var}(\hat{y})$ (Devore, 2015).

3.4. Experimental Setup

In order to achieve a quantitative evaluation of aggregates and their components and to enable pathological analysis of clinical samples, four processing pipelines have been developed. The first two approaches represent rather simple computer vision-based methods, consisting of a combination of watershed segmentation and different classification algorithms, namely expert gating and random forest. They require four processing steps: segmentation, feature extraction, classification, and a final analysis of the results. The third approach uses a U-NET and the fourth approach is based on a Mask R-CNN. These two approaches require only two processing steps because both U-NET and Mask R-CNN are capable of both segmentation and classification.

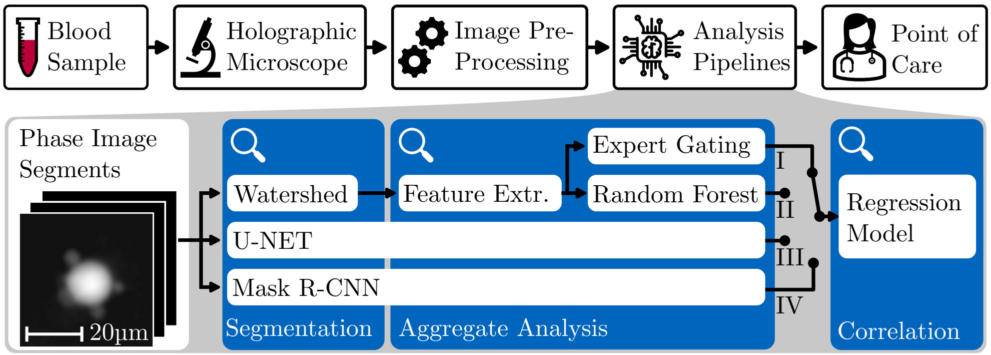


Figure 2: While the overall workflow is kept close to existing assays, we investigate the performances of the aggregate analysis pipelines I-IV.

4. Data Set

For our experiments and trainings, we use several data sets to better assess the strengths and weaknesses of the evaluated approaches and to gradually increase the level of difficulty.

4.1. BBBC038

To obtain a baseline evaluation of the segmentation methods, the publicly available BBBC038 data set (Caicedo et al., 2019) is used. It contains a variety of two-dimensional light microscopy images of stained nuclei as displayed in Figure 5 in the Appendix. For this experiment, the BBBC038 data set is divided into three parts, a training, test, and validation set (60:20:20). The training set is used to train both U-NET and Mask R-CNN, the validation set is used for hyperparameter tuning, and all three methods are evaluated on the test set.

4.2. Expert Labeled Data Set

For our particular use case we need a more accurate assessment of the segmentation. Therefore, we labeled a data set of 100 images of blood cells captured by our QPI microscope. It consists of 50% single and multiple erythrocytes and 50% platelets and leukocyte-platelet aggregates. The images were manually masked by biomedical personnel using a brush tool, resulting in a very accurate segmentation. Respective examples are shown in Figure 7 in the Appendix.

4.3. Synthetic Aggregate Data Set

The performance and reliability of a neural network is highly dependent on the quality, quantity, and selection of the training data. Acquiring ground-truth data using an unsupervised or self-supervised method, as is sometimes done when no ground-truth is available, does not solve this problem, because the trained network will never be able to outperform the quality of the training data. Alternatively, especially for the Mask R-CNN, training with only single cell images would produce adequate results, but the shape and contour of the cell images and masks will change slightly when the cells are part of aggregates, which will degrade the performance of the network, as analyzed in section 5.1. As labeling by experts is costly and time consuming we chose to generate a synthetic data set like Prastawa et al. (2005) or Gupta et al. (2016) by stitching together multiple single cell images to form cell aggregates. Its generation procedure and example images can be found in Appendix C.1.

4.4. Clinical Samples

Activated platelets For this data set, platelets are extracted from whole blood using two centrifugation steps to first extract platelet-rich plasma and then concentrate it to a pellet, which is then resuspended in a buffer solution (Bernlochner et al., 2021). After extraction, the platelets are artificially activated with the platelet activator thrombin receptor activating peptide (TRAP). Activation causes the platelets to form volatile aggregates that disintegrate over time. As shown experimentally by Michelson et al. (2001), based on measurements of platelet surface P-selectin and the occurrence of monocyte-platelet and neutrophil-platelet aggregates in whole blood, a peak of aggregation is expected after a few minutes, followed by a decline until normal levels are restored 60-120 minutes after the addition of TRAP (Michelson et al., 2001).

Five series of measurements are performed, to assess platelet aggregation levels at 7.5, 15, 30, 60, 90, and 120 minutes after application of 10 µM TRAP. Three measurements are taken at each time step, each containing approximately 5,000 platelets.

Activated platelets spiked in whole blood Aiming for data more closely related to whole blood, and to test the robust detection of aggregates as a tiny minority of events in the sample stream, we created another data set. As before, pure platelet samples are extracted from whole blood and then activated with TRAP. The activated platelets are then mixed with whole blood samples at various mixing ratios [0%, 10%, 30%, 50%, 70%, 90%, 100%]. Three samples, each containing approximately 40,000 cells, are measured for each mixing ratio. These mixing ratios should be clearly observable and the amount of aggregates detected should be dependent on the amount of activated platelets added.

Activated whole blood samples To match the conditions of Michelson et al. (2001), this data set consists of whole blood samples collected and activated by either adenosine diphosphate (ADP) or TRAP. While TRAP is a synthetic peptide, ADP is a nucleotide that binds to three specific platelet membrane receptors, triggering platelet aggregation and shape change (Murugappa and Kunapuli, 2006). For comparison, three types of samples were analyzed: untreated whole blood, whole blood activated by adding 10 µM ADP, and whole blood activated by adding 10 µM TRAP. Each sample was captured six times, with each capture containing approximately 40,000 cells.

Healthy reference Reference samples are collected from seven healthy donors, both male and female, between the ages of 29 and 67, with no history of disease. Blood samples are diluted 1:100 in a measurement buffer consisting of phosphate buffered saline (PBS) and polyethylene oxide analogous to Klenk et al. (2023) and analyzed immediately. From each sample, we record three measurements of approximately 40,000 cells each.

Sepsis For an immunothrombotic disease associated with blood cell aggregation, we collect samples from seven Intensive Care Unit (ICU) patients diagnosed with sepsis, both male and female, between the ages of 45 and 80, at multi-day intervals, typically three samples per patient. The time between blood draw and sample analysis is less than 30 minutes, which is critical for accurate assessment of aggregation. Samples are carefully transported to the nearby prototype to ensure minimal mechanical disturbance. Three measurements of 7,500 images each are taken from each sample for analysis. A single measurement typically contains approximately 30,000 cells.

COVID-19 For COVID-19, we collect samples from thirteen ICU patients (both male and female) diagnosed with PCR-confirmed wild-type SARS-CoV-2 infection between the ages of 51 and 91 at multi-day intervals, typically five samples per patient. As before, less than 30 minutes elapse between blood collection and sample analysis. From each sample, three measurements of 7,500 images each are recorded for analysis, with each measurement typically containing approximately 30,000 cells.

5. Results

The experimental results are organized in three sections. The first two sections will preselect the most appropriate aggregate analysis pipeline, which is then used to process the clinical samples in the last section. Therefore, we employ the proposed measurement points marked in blue in Figure 2 and the corresponding evaluation metrics from Section 3.3.

5.1. Segmentation

BBBC038 data set Testing the algorithms on the BBBC038 data set provides a first trend for their segmentation performance. Although this data set does not contain blood cell aggregates, it is very diverse and challenging due to its complex structures. Our observations are printed in Table 1(a). While watershed segmentation achieves only mediocre results, U-NET and Mask R-CNN achieve reasonably good results, with the Mask R-CNN showing the best overall recall and precision. Figure 8 in the Appendix shows the according visualizations.

Expert labeled data set To see if these trends continue, we switch to the expert labeled data set, which is closer related to our real-world applications. Table 1(b) demonstrates again the superiority of the Mask R-CNN (see also Appendix Figure 9). However, the watershed algorithm achieves very close results. Interestingly, when comparing the performance of the differently trained Mask R-CNNs, a Mask R-CNN trained on the BBBC038 data set already shows quite good results. As expected, segmentation using the synthetic data set for training outperforms training using only single cell images due to the changing morphology of cells that are part of aggregates. For this application, the U-NET achieves the worst results, which is most likely due to the fact that it is not perfectly suited for instance segmentation, especially of small, slightly overlapping cells.

Table 1: Segmentation quality of used methods on the test data

(*a*) BBBC038 data set

	∅IoU	Recall	Precision	F_1
Watershed	0.594	0.747	0.641	0.690
U-NET	0.584	0.760	0.831	0.794
Mask R-CNN	**0.758**	**0.909**	**0.861**	**0.884**

[1]trained on BBBC038, [2]trained on single cell images, [3]trained on synthetic aggregate data set

(*b*) Expert labeled data set

	∅IoU	Recall	Precision	F_1
Watershed	0.727	0.928	0.931	0.930
U-NET[3]	0.619	0.801	0.914	0.854
Mask R-CNN[1]	0.716	0.878	0.821	0.849
Mask R-CNN[2]	0.583	0.831	0.852	0.841
Mask R-CNN[3]	**0.741**	**0.931**	**0.956**	**0.943**

5.2. Classification

As before, the expert labeled data set of 100 images of blood cells, here including the cell labels, is used to evaluate the performance of aggregate detection combining segmentation and classification. Again, the segmentation quality is evaluated based on the metrics described in Section 3.3, adding the correctness of the predicted classes as a requirement for accepted detected instances. The evaluation results are shown in Table 2. Similar to the previous results, Mask R-CNN shows the best performance with a slight decrease in both precision and recall due to classification inaccuracies. For watershed-based methods, random forest classification shows significantly better results than expert gating. The U-NET achieves slightly better results, but worse than the Mask R-CNN. This also shows in the scores for aggregate composition (AC) and event type (ET). The Mask R-CNN qualifies as an excellent aggregate detector having an ET score of 0.970. The U-Net is also quite suitable, while the watershed-based pipelines are too coarse to detect all aggregates or mix up the contained classes. Therefore, we will use the Mask R-CNN for the following experiments.

Table 2: Segmentation and classification performance on the expert labeled dataset

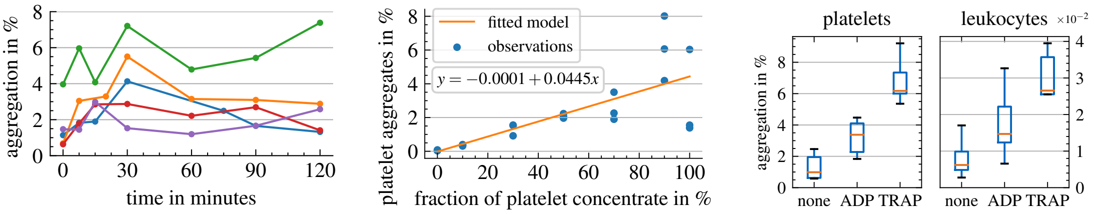
	∅IoU	Recall	Precision	F_1	AC	ET
Watershed + Gating	0.539	0.672	0.658	0.665	0.510	0.650
Watershed + Random Forest	0.630	0.790	0.773	0.782	0.580	0.730
U-NET	0.596	0.810	0.826	0.818	0.660	0.930
Mask R-CNN	**0.676**	**0.917**	**0.912**	**0.915**	**0.780**	**0.970**

5.3. Clinical Samples

Activated platelets Analysis of the five time-series measurements of activated platelets using our Mask R-CNN pipeline results in slightly different curves, as drawn in 3(*a*). However, there is a clear trend that shows a sharp increase in platelet aggregation from the beginning, peaking between 15 and 30 minutes. Thereafter, the aggregates begin to break down. These observations are roughly in line with expectations based on previous research. Only the activation seems to be a bit slower, reaching a maximum activation of only 3%-7%. This is probably caused by the fact that these experiments were performed with extracted platelets as opposed to whole blood in the case of Michelson et al. (2001). Platelet activation and aggregation is a complex process based on the coagulation cascade, and platelet-only samples lack many of the coagulants that normally promote platelet aggregation in whole blood.

Activated platelets spiked in whole blood To evaluate the detected platelet concentrations in the samples with different mixing ratios, we use linear regression. The fitted model $y = 0.0288 + 0.932x$ gives an almost perfect fit with a high coefficient of determination of $R^2 = 0.997$ and a NRMSE $= 0.018$, as shown in Figure 10 in the Appendix. These observations fit the expectation, as evidenced by the intercept, suggesting about 3% platelets in the whole blood sample, which is reasonable since it is in the typical range of 2.5%-8% (Bain, 2017). For the aggregation analysis, we also use linear regression. The fitted model can be seen in Figure 3(b) and features $R^2 = 0.552$ and NRMSE $= 0.18$.

Activated whole blood samples Analysis of the levels of platelet aggregates detected, as depicted in Figure 3(c), shows a clearly visible effect. In untreated whole blood samples almost no aggregates are observed, whereas in ADP activated samples 2 % to 4 % and TRAP activated samples 5.5 % to 8 % platelet aggregates are detected. Similarly, almost no leukocyte-platelet aggregates are observed in untreated blood samples. ADP activated samples showed 0.15 % to 0.25 % and TRAP activated samples 0.25 % to 0.35 % leukocyte-platelet aggregates. This difference between ADP- and TRAP-induced platelet aggregation is consistent with previous research by Olivier et al. (2016), where application of TRAP showed approximately 2.5 times higher aggregation than application of ADP.



(a) Activation time series (b) Aggregates in whole blood (c) Activated whole blood

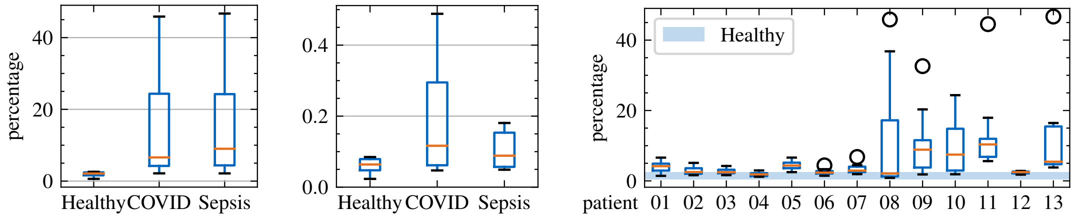
Figure 3: Detection of platelet aggregation induced by ADP or TRAP activation

Healthy reference For the reference samples, platelet aggregates are in the range 0.5 % to 2.5 % and leukocyte-platelet aggregates are in the range 0.01 % to 0.08 %, as shown in Figure 11 in the Appendix. These results are generally consistent with previous studies by Leytin et al. (2000) and Gerrits et al. (2016), which reported (1.02 ± 0.49) % and 0.001 % to 0.03 % respectively. For platelet aggregates, mostly 2-cell aggregates are observed and very few 3-cell or 4-cell aggregates, similar to the previous activation experiments. Almost all leukocyte-platelet aggregates contain only one leukocyte and mostly one (or sometimes two) platelets. This behavior is in line with expectations, as only minimal and very small aggregates are expected in healthy whole blood samples, since larger aggregates already pose a significant health risk, as described in Section 2.1.

Sepsis In our exemplary sepsis cohort, we observe both elevated platelet aggregation levels, as shown in Figure 4(a), and elevated leukocyte-platelet aggregation levels, as shown in Figure 4(*b*). Three out of seven patients (01, 02, and 04) have severely elevated levels of **platelet aggregates**, while only one is completely within the healthy reference range, as shown in Figure 12(*a*) in the Appendix. For **leukocyte-platelet aggregates**, all but one patient feature increased levels at least once, as plotted in Figure 12(*b*) in the Appendix. In addition, patients with higher aggregate levels also show a shift in aggregate size distribution with comparatively more larger aggregates, as shown in Appendix G.

COVID-19 For the COVID-19 patients, we record a similar picture with both elevated platelet aggregation levels and elevated leukocyte platelet aggregation levels, as shown in Figure 4(*a*) and 4(*b*). The effect is even more remarkable as 12 out of 13 patients have increased levels of **platelet aggregates** compared to the healthy donors. On closer inspection of Figure 4(*c*), some patients show extremely elevated levels, specifically patients 08, 09, 10, 11, and 13. Consistent with these observations, three of these four patients had a collapse of their clinical condition during observation (e.g., lung failure). In contrast, patient 04, who was transferred to the general ward during observation, shows very low platelet aggregation levels that remain within the reference range.

Looking at **leukocyte-platelet aggregates**, 11 of the 13 patients demonstrate elevated levels at least once, as shown in Figure 13(*b*) in the Appendix. For more severe cases (especially patients 08-11 and 13) this effect is clearly an indicator, but in milder cases leukocyte-platelet aggregate levels do not show a substantial shift. Similar to the sepsis experiments, samples with higher levels of aggregation also show comparatively more larger aggregates, both platelet aggregates and leukocyte-platelet aggregates, as shown in Appendix G.



(*a*) P aggregates (*b*) LP aggregates (*c*) P aggregates in COVID-19 patients

Figure 4: Platelet (P) and leukocyte-platelet (LP) aggregates for COVID-19 and sepsis

6. Discussion and Conclusion

In this work, we present a novel processing pipeline for the detection and quantitative analysis of blood cell aggregates and their components. Using QPI, this approach allows the assessment of platelet aggregation and microthrombus formation in label-free whole blood samples without the need for sample preparation. Specific detection of each component of an aggregate allows evaluation of the size and number of platelet and leukocyte-platelet aggregates.

In various test scenarios, we compared four different approaches using established and custom metrics at different stages of the pipelines. The first approach is a non-machine learning method consisting of a combination of watershed segmentation and two different classification methods. Watershed segmentation showed great potential for segmenting blood cell aggregates, but the evaluated classification method was not able to reliably distinguish leukocytes from erythrocytes. The extension with a data-driven random forest classifier as a second approach did not lead to the desired improvements. The third approach uses a U-NET, which we adapted for instance segmentation by using a boundary loss function, which showed decent results. The fourth approach is based on a Mask R-CNN trained on an artificially created synthetic aggregate data set. This approach showed the best results, with a precision of 0.956 and a recall of 0.931 on an expert-labeled test set, and most importantly, it yielded the correct category of cell or cell aggregate in 97% of the cases.

The Mask R-CNN processing pipeline was then evaluated on defined medical samples comprising activated platelets, activated platelets spiked in whole blood, and activated whole blood. These experiments demonstrated very reliable detection of platelet aggregates, but showed some limitations for leukocyte-platelet aggregates due to low statistical power.

Finally, we evaluated the quality of this method as a diagnostic predictive biomarker for immunothrombotic diseases by analyzing samples from patients with COVID-19 and sepsis. In both diseases, 90% of patients had aggregate levels above the healthy reference interval, with all severe patients having substantially higher aggregate levels (5-10 times higher than reference samples). In addition, these samples with particularly high aggregate levels also had consistently higher amounts of larger aggregates. In conclusion, the analysis of these clinical samples demonstrated the effectiveness of the proposed method and the potential of using the occurrence of platelet and leukocyte-platelet aggregates as biomarkers for the presence and severity of immunothrombotic diseases.

Limitations Due to the difficulty in obtaining blood samples from COVID-19 and sepsis patients during the pandemic, only a limited number of clinical patients were analyzed in this work, which does not allow a concrete diagnostic and therapeutic assessment of the occurrence of aggregates in the studied diseases. For a higher statistical power and a more precise assertion of the severity of both sepsis and COVID-19 (Rampotas and Pavord, 2021), a larger clinical study needs to be performed. We also need to collect more medical parameters as described by Poudel et al. (2021) or Gorog et al. (2022) to correlate our new biomarker with established biomarkers to prove advantages or discrepancies. In addition, we observed a significant decrease in aggregate levels between multiple acquisitions of the same sample, demonstrating the short lifespan of blood cell aggregates and confirming the need for immediate sample analysis, ideally in a point-of-care environment. The choice of anticoagulant also plays an important role and needs to be evaluated in future studies (Klenk et al., 2023). Finally, the new methodology needs to be proven in clinical applications before this platform technology can add value at the point of care. This is not least due to the acceptance of black box models that still needs to be built up, which can be achieved through long-term successful use. However, we hope that our methodology will open the way for further applications that can benefit from the detailed analysis of immunothrombosis (Engelmann and Massberg, 2013; Wendelboe and Raskob, 2016; Nicolai et al., 2020; Stark and Massberg, 2021).

References

Nicole Allen, Tessa J. Barrett, Yu Guo, Michael Nardi, Bhama Ramkhelawon, Caron B. Rockman, Judith S. Hochman, and Jeffrey S. Berger. Circulating monocyte-platelet aggregates are a robust marker of platelet activity in cardiovascular disease. *Atherosclerosis*, 282:11–18, 2019.

Alice Assinger, Waltraud C. Schrottmaier, Manuel Salzmann, and Julie Rayes. Platelets in sepsis: An update on experimental models and clinical data. *Frontiers in Immunology*, 10:1687, 2019.

Barbara J. Bain. *A Beginner's Guide to Blood Cells.* John Wiley & Sons Ltd, 3rd edition, 2017.

Juan José Barcia. The giemsa stain: Its history and applications. *International Journal of Surgical Pathology*, 15(3):292–296, 2007.

Isabell Bernlochner, Melissa Klug, Ditya Larasati, Moritz Von Scheidt, Donato Santovito, Michael Hristov, Christian Weber, Karl-Ludwig Laugwitz, and Dario Bongiovanni. Sorting and magnetic-based isolation of reticulated platelets from peripheral blood. *Platelets*, 32(1):113–119, 2021.

Serge Beucher and Christian Lantuéjoul. Use of watersheds in contour detection. In *International Workshop on Image Processing: Real-time Edge and Motion Detection/Estimation*, volume 132, 1979.

Deepak L. Bhatt and Eric J. Topol. Scientific and therapeutic advances in antiplatelet therapy. *Nature Reviews Drug Discovery*, 2(1):15–28, 2003.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Wilhelm Burger and Mark J. Burge. *Morphological Filters*, pages 181–208. Springer London, London, 2016.

Juan C. Caicedo, Allen Goodman, Kyle W. Karhohs, Beth A. Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, Mohammad Rohban, Shantanu Singh, and Anne E. Carpenter. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature Methods*, 16(12):1247–1253, 2019.

Jay L. Devore. *Probability and Statistics for Engineering and the Sciences.* Cengage Learning, 2015.

Frank Dubois and Catherine Yourassowsky. US patent 7,362.449 b2, 2008.

Bernd Engelmann and Steffen Massberg. Thrombosis as an intravascular effector of innate immunity. *Nature Reviews. Immunology*, 13(1):34–45, 2013.

Andrew Filby. Sample preparation for flow cytometry benefits from some lateral thinking. *Cytometry Part A*, 89(12):1054–1056, 2016.

Michaela Finsterbusch, Waltraud C. Schrottmaier, Julia B. Kral-Pointner, Manuel Salzmann, and Alice Assinger. Measuring and interpreting platelet-leukocyte aggregates. *Platelets*, 29(7):677–685, 2018.

Mark I. Furman, Marc R. Barnard, Lori A. Krueger, Marsha L. Fox, Elizabeth A. Shilale, Darleen M. Lessard, Peter Marchese, A.L Frelinger, Robert J. Goldberg, and Alan D. Michelson. Circulating monocyte-platelet aggregates are an early marker of acute myocardial infarction. *Journal of the American College of Cardiology*, 38(4):1002–1006, 2001.

Anja J. Gerrits, Andrew L. Frelinger, and Alan D. Michelson. Whole blood analysis of leukocyte-platelet aggregates. *Current Protocols in Cytometry*, 78:6.15.1–6.15.10, 2016.

Diana A Gorog, Robert F Storey, Paul A Gurbel, Udaya S Tantry, Jeffrey S Berger, Mark Y Chan, Daniel Duerschmied, Susan S Smyth, William A E Parker, Ramzi A Ajjan, Gemma Vilahur, Lina Badimon, Jurrien M Ten Berg, Hugo Ten Cate, Flora Peyvandi, Taia T Wang, and Richard C Becker. Current and novel biomarkers of thrombotic risk in covid-19: a consensus statement from the international covid-19 thrombosis biomarkers colloquium. *Nature reviews. Cardiology*, 19(7):475—495, 2022.

Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2315–2324, 2016.

Robert M. Haralick, Its'hak Dinstein, and K. Shanmugam. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(6):610–621, 1973.

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*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

Richard S. Hotchkiss, Lyle L. Moldawer, Steven M. Opal, Konrad Reinhart, Isaiah R. Turnbull, and Jean-Louis Vincent. Sepsis and septic shock. *Nature reviews. Disease primers*, 2:16045, 2016.

Paul Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer-Verlag, 2013.

Young Ju Jo, Jae Hwang Jung, Min hyeok Kim, Hyun Joo Park, Suk-Jo Kang, and Yong Keun Park. Label-free identification of individual bacteria using fourier transform light scattering. *Optics Express*, 23(12):15792–15805, 2015.

Young Ju Jo, Hyungjoo Cho, Sang Yun Lee, Gunho Choi, Geon Kim, Hyun Seok Min, and Yong Keun Park. Quantitative phase imaging and artificial intelligence: A review. *IEEE Journal of Selected Topics in Quantum Electronics*, 25(1):1–14, 2018.

Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *International Conference on Medical Imaging with Deep Learning*, pages 285–296. PMLR, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Christian Klenk, Dominik Heim, Matthias Ugele, and Oliver Hayden. Impact of sample preparation on holographic imaging of leukocytes. *Optical Engineering*, 59(10):102403, 2019.

Christian Klenk, David Fresacher, Stefan Röhrl, Dominik Heim, Manuel Lengl, Simon Schumann, Martin Knopp, Klaus Diepold, Stefan Holdenrieder, and Oliver Hayden. Measurement of platelet aggregation in ageing samples and after in-vitro activation. In *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 2: BIOIMAGING*, pages 57–65. INSTICC, SciTePress, 2023.

Sotiris B. Kotsiantis, Dimitris Kanellopoulos, and Panagiotis E. Pintelas. Data preprocessing for supervised leaning. *International Journal of Computer and Information Engineering*, 1(12):4104–4109, 2007.

Tobias Kutscher, Kai Eder, Anne Marzi, Álvaro Barroso, Jürgen Schnekenburger, and Björn Kemper. Cell Detection and Segmentation in Quantitative Digital Holographic Phase Contrast Images Utilizing a Mask Region-based Convolutional Neural Network. In *OSA Optical Sensors and Sensing Congress*, page JTu5A.23. Optica Publishing Group, 2021.

Dovena Lazaridis, Simon Leung, Lisa Kohler, Carla Hawkins Smith, Margaretta L Kearson, and Nathaniel Eraikhuemen. The impact of anticoagulation on covid-19 (sars cov-2) patient outcomes: a systematic review. *Journal of Pharmacy Practice*, 35(6):1000–1006, 2022.

Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164–168, 1944.

Marcel Levi, Marcus Schultz, and Tom van der Poll. Sepsis and thrombosis. *Seminars in Thrombosis and Hemostasis*, 39(5):559–566, 2013.

Valery Leytin, Meera Mody, John W Semple, Bernadette Garvey, and John Freedman. Flow cytometric parameters for characterizing platelet activation by measuring p-selectin (CD62) expression: theoretical consideration and evaluation in thrombin-treated platelet populations. *Biochemical and Biophysical Research Communications*, 269(1):85–90, 2000.

Alan D. Michelson, Marc R. Barnard, Lori A. Krueger, C. Robert Valeri, and Mark I. Furman. Circulating monocyte-platelet aggregates are a more sensitive marker of in vivo platelet activation than platelet surface p-selectin. *Circulation*, 104(13):1533–1537, 2001.

Benjamin Midtvedt, Saga Helgadottir, Aykut Argun, Jesús Pineda, Daniel Midtvedt, and Giovanni Volpe. Quantitative digital microscopy with deep learning. *Applied Physics Reviews*, 8(1):011310, 2021.

Swaminathan Murugappa and Satya P. Kunapuli. The role of ADP receptors in platelet function. *Frontiers in Bioscience: A Journal and Virtual Library*, 11:1977–1986, 2006.

H.P. Ng, S.H. Ong, K.W.C. Foong, P.S. Goh, and W.L. Nowinski. Medical image segmentation using k-means clustering and improved watershed algorithm. In *2006 IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 61–65, 2006.

Tan Huu Nguyen, Shamira Sridharan, Virgilia Macias, Andre Kajdacsy-Balla, Jonathan Melamed, Minh N. Do, and Gabriel Popescu. Automatic Gleason grading of prostate cancer using quantitative phase imaging and machine learning. *Journal of Biomedical Optics*, 22(3):036015, 2017.

Thang L. Nguyen, Soorya Pradeep, Robert L. Judson-Torres, Jason Reed, Michael A. Teitell, and Thomas A. Zangle. Quantitative phase imaging: Recent advances and expanding potential in biomedicine. *American Chemical Society Nano*, 16(8):11516–11544, 2022.

Leo Nicolai, Alexander Leunig, Sophia Brambs, Rainer Kaiser, Tobias Weinberger, Michael Weigand, Maximilian Muenchhoff, Johannes C. Hellmuth, Stephan Ledderose, Heiko Schulz, Clemens Scherer, Martina Rudelius, Michael Zoller, Dominik Höchter, Oliver Keppler, Daniel Teupser, Bernhard Zwißler, Michael von Bergwelt-Baildon, Stefan Kääb, Steffen Massberg, Kami Pekayvaz, and Konstantin Stark. Immunothrombotic dysregulation in COVID-19 pneumonia is associated with respiratory failure and coagulopathy. *Circulation*, 142(12):1176–1189, 2020.

Masako Nishikawa, Hiroshi Kanno, Yuqi Zhou, Ting-Hui Xiao, Takuma Suzuki, Yuma Ibayashi, Jeffrey Harmon, Shigekazu Takizawa, Kotaro Hiramatsu, Nao Nitta, et al. Massive image-based single-cell profiling reveals high levels of circulating platelet aggregates in patients with covid-19. *Nature Communications*, 12(1):1–12, 2021.

Christoph B. Olivier, Melanie Meyer, Hans Bauer, Katharina Schnabel, Patrick Weik, Qian Zhou, Christoph Bode, Martin Moser, and Philipp Diehl. The ratio of ADP- to TRAP-induced platelet aggregation quantifies p2y12-dependent platelet inhibition independently of the platelet count. *PloS One*, 11(2):e0149053, 2016.

Yusuke Ozaki, Hidenao Yamada, Hirotoshi Kikuchi, Amane Hirotsu, Tomohiro Murakami, Tomohiro Matsumoto, Toshiki Kawabata, Yoshihiro Hiramatsu, Kinji Kamiya, Toyohiko Yamauchi, Kentaro Goto, Yukio Ueda, Shigetoshi Okazaki, Masatoshi Kitagawa, Hiroya Takeuchi, and Hiroyuki Konno. Label-free classification of cells based on supervised machine learning of subcellular structures. *PLOS ONE*, 14:1–20, 2019.

Santosh Kumar Paidi, Piyush Raj, Rosalie Bordett, Chi Zhang, Sukrut H. Karandikar, Rishikesh Pandey, and Ishan Barman. Raman and quantitative phase imaging allow morpho-molecular recognition of malignancy and stages of B-cell acute lymphoblastic leukemia. *Biosensors and Bioelectronics*, 190:113403, 2021.

Scott W. Paine and James R. Fienup. Machine learning for improved image-based wavefront sensing. *Optics Letters*, 43(6):1235–1238, 2018.

Ayusha Poudel, Yashasa Poudel, Anurag Adhikari, Barun Babu Aryal, Debika Dangol, Tamanna Bajracharya, Anil Maharjan, and Rakshya Gautam. D-dimer as a biomarker for assessment of covid-19 prognosis: D-dimer levels on admission and its role in predicting disease outcome in hospitalized patients with covid-19. *PloS ONE*, 16(8):e0256744, 2021.

David M. W. Powers. Evaluation: from precision, recall and f-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

Marcel Prastawa, Elizabeth Bullitt, and Guido Gerig. Synthetic ground truth for validation of brain tumor MRI segmentation. In James S. Duncan and Guido Gerig, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005*, Lecture Notes in Computer Science, pages 26–33. Springer, 2005.

Alexandros Rampotas and Sue Pavord. Platelet aggregates, a marker of severe covid-19 disease. *Journal of Clinical Pathology*, 74(11):750–751, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

Miguel Reyes, Michael R. Filbin, Roby P. Bhattacharyya, Kianna Billman, Thomas Eisenhaure, Deborah T. Hung, Bruce D. Levy, Rebecca M. Baron, Paul C. Blainey, Marcia B. Goldberg, and Nir Hacohen. An immune-cell signature of bacterial sepsis. *Nature Medicine*, 26(3):333–340, 2020.

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, Bram Rochwerg, Gordon D. Rubenfeld, Derek C. Angus, Djillali Annane, Richard J. Beale, Geoffrey J. Bellinghan, Gordon R. Bernard, Jean-Daniel Chiche, Craig Coopersmith, Daniel P. De Backer, Craig J. French, Seitaro Fujishima, Herwig Gerlach, Jorge Luis Hidalgo, Steven M. Hollenberg, Alan E. Jones, Dilip R. Karnad, Ruth M. Kleinpell, Younsuk Koh, Thiago Costa Lisboa, Flavia R. Machado, John J. Marini, John C. Marshall, John E. Mazuski, Lauralyn A. McIntyre, Anthony S. McLean, Sangeeta Mehta, Rui P. Moreno, John Myburgh, Paolo Navalesi, Osamu Nishida, Tiffany M. Osborn, Anders Perner, Colleen M. Plunkett, Marco Ranieri, Christa A. Schorr, Maureen A. Seckel, Christopher W. Seymour, Lisa Shieh, Khalid A. Shukri, Steven Q. Simpson, Mervyn Singer, B. Taylor Thompson, Sean R. Townsend, Thomas Van der Poll, Jean-Louis Vincent, W. Joost Wiersinga, Janice L. Zimmerman, and R. Phillip Dellinger. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock: 2016. *Intensive Care Medicine*, 43(3):304–377, 2017.

Julia Riedl, Cihan Ay, and Ingrid Pabinger. Platelets and hemophilia: A review of the literature. *Thrombosis Research*, 155:131–139, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 234–241. Springer International Publishing, 2015.

Stefan Röhrl, Matthias Ugele, Christian Klenk, Dominik Heim, Oliver Hayden, and Klaus Diepold. Autoencoder features for differentiation of leukocytes based on digital holographic microscopy (DHM). In *Computer Aided Systems Theory - EUROCAST 2019*. Springer, 2019.

Harekrushna Sahoo. Fluorescent labeling techniques in biomolecules: A flashback. *RSC Advances*, 2(18):7017–7029, 2012.

Ulf Schnars and Werner Jüptner. Direct recording of holograms by a CCD target and numerical reconstruction. *Applied optics*, 33(2):179–181, 1994.

Jonas Schulte-Schrepping, Nico Reusch, Daniela Paclik, Kevin Baßler, Stephan Schlickeiser, Bowen Zhang, Benjamin Krämer, Tobias Krammer, Sophia Brumhard, Lorenzo Bonaguro, Elena De Domenico, Daniel Wendisch, Martin Grasshoff, Theodore S. Kapellos, Michael Beckstette, Tal Pecht, Adem Saglam, Oliver Dietrich, Henrik E. Mei, Axel R. Schulz, Claudia Conrad, Désirée Kunkel, Ehsan Vafadarnejad, Cheng-Jian Xu, Arik Horne, Miriam Herbert, Anna Drews, Charlotte Thibeault, Moritz Pfeiffer, Stefan Hippenstiel, Andreas Hocke, Holger Müller-Redetzky, Katrin-Moira Heim, Felix Machleidt, Alexander Uhrig, Laure Bosquillon de Jarcy, Linda Jürgens, Miriam Stegemann, Christoph R. Glösenkamp, Hans-Dieter Volk, Christine Goffinet, Markus Landthaler, Emanuel Wyler, Philipp Georg, Maria Schneider, Chantip Dang-Heine, Nick Neuwinger, Kai Kappert, Rudolf Tauber, Victor Corman, Jan Raabe, Kim Melanie Kaiser, Michael To Vinh, Gereon Rieke, Christian Meisel, Thomas Ulas, Matthias Becker, Robert Geffers, Martin Witzenrath, Christian Drosten, Norbert Suttorp, Christof von Kalle, Florian Kurth, Kristian Händler, Joachim L. Schultze, Anna C. Aschenbrenner, Yang Li, Jacob Nattermann, Birgit Sawitzki, Antoine-Emmanuel Saliba, Leif Erik Sander, and Deutsche COVID-19 OMICS Initiative (DeCOI). Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell*, 182(6):1419–1440.e23, 2020.

Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017.

Dalwinder Singh and Birmohan Singh. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97:105524, 2020.

Janet Staats, Anagha Divekar, J Philip McCoy, and Holden T Maecker. Guidelines for gating flow cytometry data for immunological assays. *Immunophenotyping: Methods and Protocols*, pages 81–104, 2019.

Konstantin Stark and Steffen Massberg. Interplay between inflammation and thrombosis in cardiovascular pathology. *Nature Reviews Cardiology*, 18(9):666–682, 2021.

Satoshi Suzuki and Keiichi Abe. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics and Image Processing*, 30(1):32–46, 1985.

Matthias Ugele. *High-throughput hematology analysis with digital holographic microscopy.* PhD thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 2019.

Matthias Ugele, Markus Weniger, Maria Leidenberger, Yiwei Huang, Michael Bassler, Oliver Friedrich, Barbara Kappes, Oliver Hayden, and Lukas Richter. Label-free, high-throughput detection of p. falciparum infection in sphered erythrocytes with digital holographic microscopy. *Lab on a Chip*, 18(12):1704–1712, 2018a.

Matthias Ugele, Markus Weniger, Manfred Stanzel, Michael Bassler, Stefan W. Krause, Oliver Friedrich, Oliver Hayden, and Lukas Richter. Label-free high-throughput leukemia detection by holographic microscopy. *Advanced Science*, 5(12), 2018b.

Tom van der Poll, Frank L. van de Veerdonk, Brendon P. Scicluna, and Mihai G. Netea. The immunopathology of sepsis and potential therapeutic targets. *Nature Reviews. Immunology*, 17(7):407–420, 2017.

Luc Vincent and Pierre Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):583–598, 1991.

Pei Wang and Albert C. S. Chung. Focal dice loss and image dilation for brain tumor segmentation. In Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Lecture Notes in Computer Science, pages 119–127. Springer International Publishing, 2018.

Aaron M. Wendelboe and Gary E. Raskob. Global burden of thrombosis: Epidemiologic aspects. *Circulation Research*, 118(9):1340–1347, 2016.

Gong Zhang, Tian Guan, Zhiyuan Shen, Xiangnan Wang, Tao Hu, Delai Wang, Yonghong He, and Ni Xie. Fast phase retrieval in off-axis digital holographic microscopy through deep learning. *Optics Express*, 26(15):19388–19405, 2018.

Yu Zuo, Yogendra Kanthi, Jason S. Knight, and Alfred H. J. Kim. The interplay between neutrophils, complement, and microthrombi in COVID-19. *Best Practice & Research. Clinical Rheumatology*, 35(1):101661, 2021.

Appendix A. Preprocessing

Preprocessing is an essential requirement in achieving good segmentation and classification results. The QPI setup provides 512 px by 384 px phase images containing multiple cells as displayed in Figure 1(*b*). These must be prepared to obtain usable images containing only a single cell or cell aggregate, while keeping cell aggregates intact. The operations required to achieve this are discussed below.

A.1. Background Subtraction

Disturbing artifacts and background noise can be removed by calculating the median of 100 images and subtracting it from each frame. This can be done as the imaging setup and the channel is assumed static.

A.2. Cell Detection

The detection of cells in the acquired images is done by thresholding and contour finding. First, binary thresholding is applied to the phase images. From the resulting binary images, contours are extracted based on the algorithm of Suzuki and Abe (1985). The extracted contours are filtered according to a minimum contour area and each cell or cell aggregate (represented by a contour) was then saved as an image snippet of 100 px by 100 px for further processing.

A.3. Masking

Individual cell masking is used to remove unwanted noise from fluid, particles, and other cells. This is done by first thresholding to remove any residual noise caused by the microfluidics channel. To improve the resulting mask, other cells or particles in the image are removed from the mask using the previously calculated contour, and any holes in the mask are filled using morphological dilation and erosion (Burger and Burge, 2016).

A.4. Normalization

Normalization is an essential preprocessing step in any machine learning application, especially when using neural networks. It transforms the feature or image values into a common range. Typical methods are either mean and standard deviation based (like z-score normalization) or minimum-maximum based (Singh and Singh, 2020) (Kotsiantis et al., 2007). For this work, the images were first clipped to limit the value range, as the images resulting from the holographic microscope (theoretically) have an unlimited value range. A minimum clipping value of 0.2 (due to the background) and a maximum clipping value of 4 were used, which showed good results for a combination of platelets, erythrocytes, and leukocytes (with a fully used value range and minimal clipping of cells). *Min-Max normalization* was then applied to transform the image values into the $[0, 1]$ interval suitable for neural networks.

Appendix B. Hyperparameter Optimization and Training

Hyperparameters are all configuration parameters of a neural network that can be set by the user. They directly control the behavior of the network during training and have a dominant impact on model performance. Hyperparameters control the network's architecture, regularization, and most importantly optimization. Tuning of hyperparameters, called hyperparameter optimization, is therefore needed to be able to exploit the full potential of a neural network. Since manual tuning is tedious and inefficient, automatic optimization is widely used.

The simplest optimization methods are grid search, which traverses the search space on a grid in an ordered fashion, and random search, which tries random combinations of hyperparameters. More advanced methods use Gaussian processes and early stopping.

In this work, a combination of the tree-structured parzen estimator (TPE) and the asynchronous successive halving approach (ASHA) is used. TPE is a sequential model based optimization approach. By describing the search space with a graph-structured generative process, a model of the relation between hyperparameters and measured performance of the neural network can be created. This model is then successively optimized by sequentially constructing models to approximate the performance of hyperparameters based on previous measurements and subsequently proposing new hyperparameter combinations. ASHA uses aggressive early stopping of bad performing training steps to allocate more time and computing power to more promising configurations. The combination of these two methods makes very efficient hyperparameter optimization possible.

U-Net For the U-NET, the parameters α, β_1, β_2 and ϵ were optimized based on the search space defined in Table 3. The best choices were $\alpha = 3 \times 10^{-4}$, $\beta_1 = 0.81$, $\beta_2 = 0.994$, and $\epsilon = 4 \times 10^{-8}$.

Parameter	Search space	Choice
learning rate α	loguniform$(1 \times 10^{-5}, 1 \times 10^{-2})$	3×10^{-4}
exp decay rate β_1	uniform$(0, 0.9)$	0.81
exp decay rate β_2	uniform$(0.9, 0.999)$	0.994
numerical stability parameter ϵ	loguniform$(1 \times 10^{-8}, 1)$	4×10^{-8}

Table 3: Search space and chosen value of the hyperparameter optimization of the U-NET

Mask R-CNN For the Mask R-CNN, the parameters learning rate, momentum and weight decay were optimized using the search space defined in Table 4. A learning rate of 1×10^{-3}, a momentum of 0.97, and a weight decay of 5×10^{-4} yielded the best results.

Parameter	Search space	Choice
learning rate lr	$\mathrm{loguniform}(1 \times 10^{-5}, 1 \times 10^{-2})$	1×10^{-3}
momentum γ	$\mathrm{uniform}(0.9, 0.999)$	0.97
weight decay wd	$\mathrm{loguniform}(1 \times 10^{-4}, 1 \times 10^{-1})$	5×10^{-4}

Table 4: Search space and chosen value of the hyperparameter optimization of the Mask R-CNN

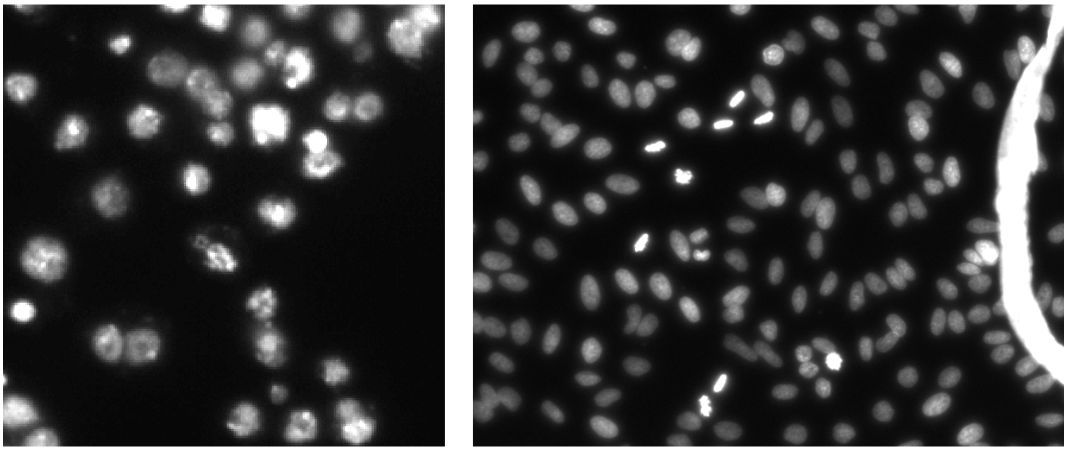
Appendix C. Visualizations

Since our work is very visual, we do not want to deprive readers of the corresponding images and segmentations.

C.1. Data Sets

This section contains exemplary images for the employed data sets.

BBBC038 The BBBC038 data set contains a variety of two-dimensional light microscopy images of stained nuclei. Two examplariy images are displayed in Figure 5.

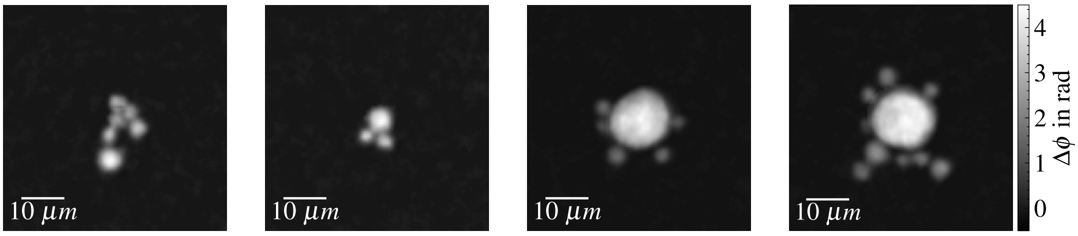


(a) Example image 1

(b) Example image 2

Figure 5: BBBC038 by Caicedo et al. (2019)

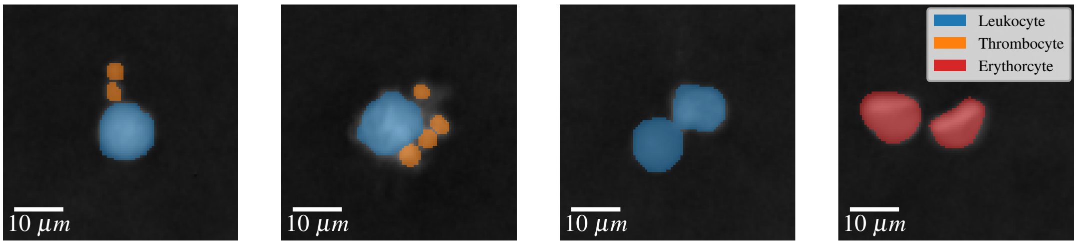
Synthetic data set To control the composition of Aggregates, we created a synthetic data set of aggregates, by stitching together multiple single cell images to form cell aggregates. This is based on pure blood cell populations (platelets, erythrocytes, leukocytes) extracted from whole blood by differential centrifugation and density gradient centrifugation. Single cell images and corresponding masks were extracted using threshold segmentation and manual filtering to remove cell duplicates, out-of-focus cells, and flow channel artifacts. The synthetic aggregate images are then iteratively assembled by randomly placing them side by side based on their contour. Since this is done simultaneously for the mask of the cell images, the ground truth label mask needed for training is created. The results are extremely close to real blood cell aggregates, as visualized in Figure 6.



(a) P Aggregate (b) P Aggregate (c) LP Aggregate (d) LP Aggregate

Figure 6: Synthetic platelet (P) and leukocyte-platelet (LP) aggregates

Expert labeled data set We asked a team of biomedical researchers and experts on QPI blood cell analysis to label a data set of 100 images. It consists of 50% single and multiple erythrocytes and 50% platelets and leukocyte-platelet aggregates. The images were manually masked by using a brush tool, resulting in a very accurate segmentation. The according examples are shown in Figure 7.



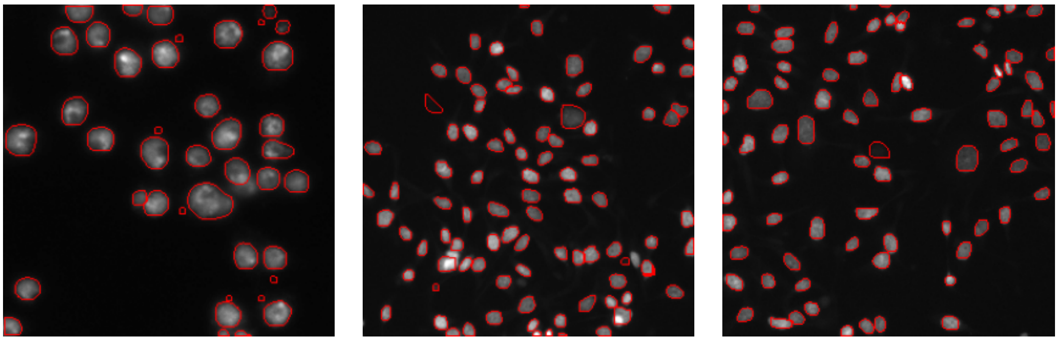
(a) LP Aggregate (b) LP Aggregate (c) Leukocytes (d) Erythrocytes

Figure 7: Expert labeled data set

C.2. Segmentation

This section contains sample images generated by our analysis pipeline. Here, the segmentation performance of the Mask R-CNN approach can be observed on the retrieved images patches.

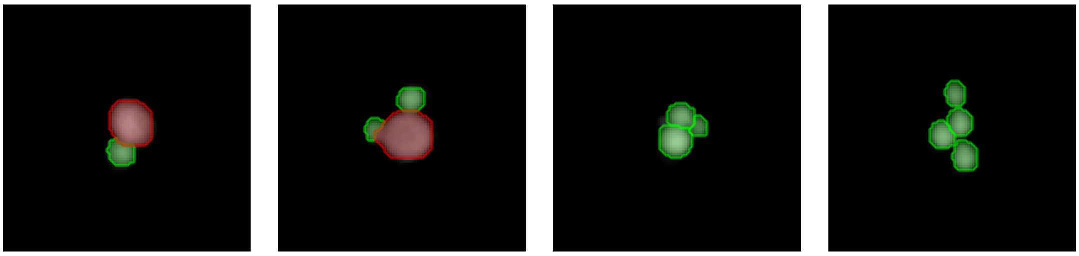
BBBC038 The complexity of the BBBC038 data set is a good benchmark to test the segmentation capabilities of the examined approaches. The Mask R-CNN does an excellent job of detecting a wide range of small and large cells while achieving a high IoU. The contours of the recognized cells are drawn red in Figure 8.



(*a*) Example 1 (*b*) Example 2 (*c*) Example 3

Figure 8: Segmentation examples of the Mask R-CNN on the BBBC038 data set

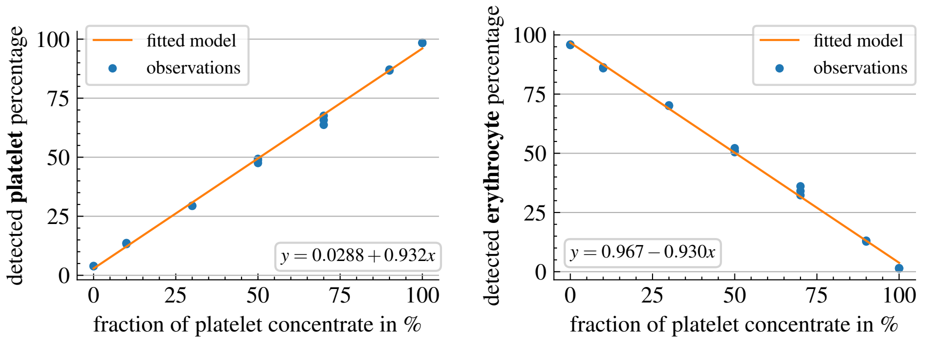
Expert labeled data set Closer to our actual use case, the segmentation of the expert labeled data set puts the focus on the detection of the individual aggregate components. Also for this challenging task, the Mask R-CNN shows a good performance in all quality measures (see Section 5.1). The according visualizations using red and green contours for each detected component can be seen in Figure 9.



(*a*) LP Aggregate (*b*) LP Aggregate (*c*) P Aggregate (*d*) P Aggregate

Figure 9: Segmentation examples of Mask R-CNN on the expert labeled data set: The color of the contour represents the predicted type. Leukocytes (L) are drawn in red and platelets (P) in green.

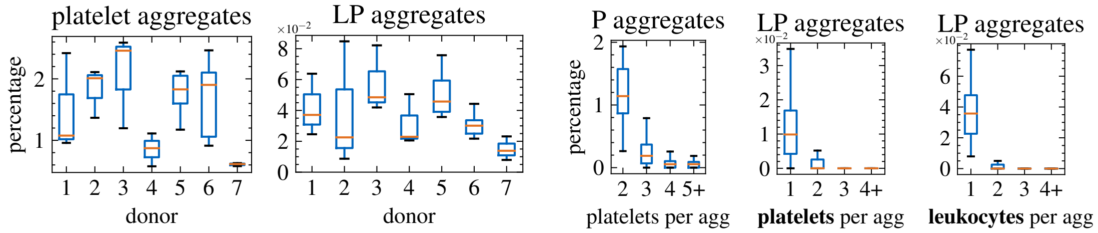
Appendix D. Activated platelets spiked in whole blood



(*a*) Detected platelet percentage (*b*) Detected erythrocyte percentage

Figure 10: Activated platelets spiked in whole blood: The dots show the observations and the curve represents the fitted function.

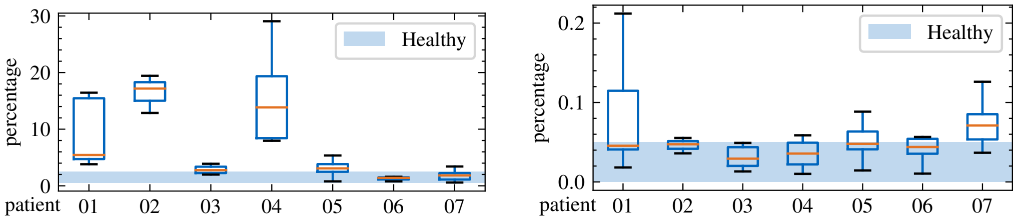
Appendix E. Healthy Reference



(*a*) Platelet (P) and leukocyte-platelet (LP) aggregate occurrence (*b*) Aggregate composition of platelet (P) and leukocyte-platelet (LP) aggregates

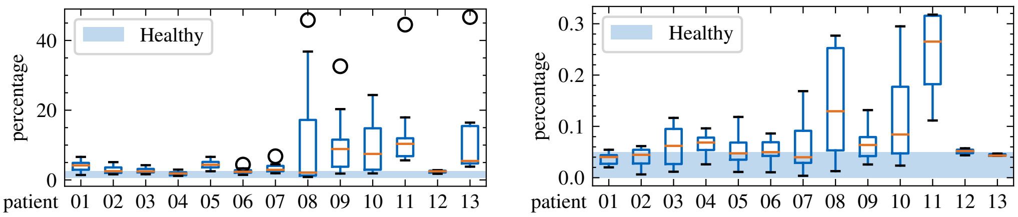
Figure 11: Aggregates in samples from healthy donors

Appendix F. Sepsis and COVID-19 in Detail



(*a*) Platelet aggregates (*b*) Leukocyte-platelet aggregates

Figure 12: Aggregate occurrence in samples from patients with sepsis



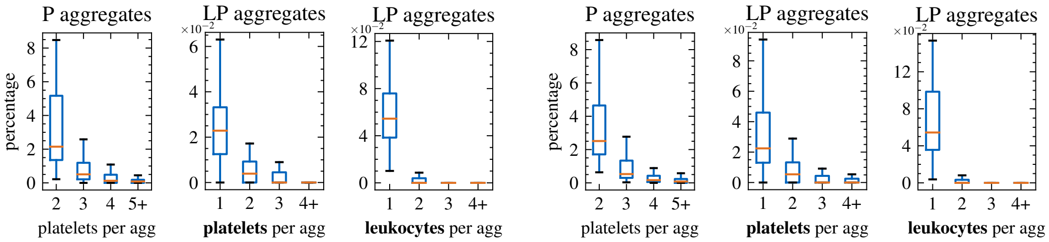
(*a*) Platelet aggregates (*b*) Leukocyte-platelet aggregates

Figure 13: Aggregate occurrence in samples from patients with COVID-19

Appendix G. Aggregate Composition

In order to gain a better understanding of the characteristics of aggregates, we analyzed the size distribution of aggregates of both the sepsis and COVID-19 samples. For the assessment of platelet aggregate size, the samples are divided into two categories, samples with lower platelet aggregate levels and samples with higher platelet aggregate levels, separated by a threshold of 10%. For the sepsis cohort, patients with sepsis with fewer observed aggregates show a similar distribution of aggregate size to healthy donors, while samples with higher aggregate levels also show comparatively more larger aggregates, as shown in Figure 14(*a*). The same is observable for COVID-19 patients, as shown in Figure 14(*b*).

To analyze the amount of platelets in leukocyte-platelet aggregates a threshold of 0.2% was chosen to divide the samples into two categories of lower and higher observed leukocyte-platelet aggregates. Analysis of the amount of platelets in leukocyte-platelet aggregates shows similar results to those of platelet aggregates for sepsis patients, as shown in Figure 14(*a*), as well as COVID-19 patients, as shown in Figure 14(*b*). Patients with more observed aggregates also showed comparatively more larger aggregates. However, the analysis of leukocyte amounts in leukocyte-platelet aggregates showed slightly different results, as shown in Figure 14(*a*) and 14(*b*). Both in healthy donors, in patients with fewer observed leukocyte-platelet aggregates, and in patients with more leukocyte-platelet aggregates almost only aggregates containing a single leukocyte are observed, all of these featuring a similar distribution, just with different levels in general.



(*a*) In sepsis samples (*b*) In COVID-19 samples

Figure 14: Aggregate composition of platelet (P) and leukocyte-platelet (LP) aggregates

Reprint Permission

Title:	Composition Counts: A Machine Learning View on Immunothrombosis using Quantitative Phase Imaging
Author:	David Fresacher, Stefan Röhrl, Christian Klenk, Johanna Erber, Hedwig Irl, Dominik Heim, Manuel Lengl, Simon Schumann, Martin Knopp, Martin Schlegel, Sebastian Rasch, Oliver Hayden and Klaus Diepold
Publication:	Proceedings of the 8th Machine Learning for Healthcare Conference
Publisher:	Proceedings of Machine Learning Research
Date:	August 16, 2024
Online:	https://proceedings.mlr.press/v219/fresacher23a.html

Core Publication IV

Towards Interpretable Classification of Leukocytes based on Deep Learning

Stefan Röhrl, Johannes Groll, Manuel Lengl, Simon Schumann, Christian Klenk, Dominik Heim, Martin Knopp, Oliver Hayden, Klaus Diepold

Summary This paper addresses uncertainty calibration and communication in holographic cytology to build trust in AI-driven systems. It focuses on interpretability by using local visual explanations to reveal patterns in cell detection, similar to human biologists examining a blood smear. Comparing two lightweight neural network architectures, AlexNet and LeNet5, it uses variational inference for confidence estimation without compromising predictive accuracy. Visual explanation techniques reveal the networks' recognition strategies, although the results may not perfectly match the established biological features. The networks demonstrate reliability in leukocyte classification and robustness to outliers, even identifying potentially mislabeled cells, highlighting their ability to provide accurate classifications.

Own Contributions

- Initiating the research idea
- Gathering of related work and assessing the scientific context of this work
- Planning the experiments and generating the corresponding datasets
- Writing the complete manuscript and designing the figures and plots
- Revising and completing the software framework
- Validating the experiments and discussing the results
- Defending the work and integrating the reviewer feedback

S. Röhrl, J. Groll, M. Lengl, S. Schumann, C. Klenk, D. Heim, M. Knopp, O. Hayden, and K. Diepold. "Towards Interpretable Classification of Leukocytes Based on Deep Learning". In: *International Conference on Machine Learning 3rd Workshop on Interpretable Machine Learning in Healthcare.* 2023, p. 2311.14485

Towards Interpretable Classification of Leukocytes based on Deep Learning

Stefan Röhrl [*1] **Johannes Groll** [*1] **Manuel Lengl** [1] **Simon Schumann** [1] **Christian Klenk** [2] **Dominik Heim** [2]
Martin Knopp [2 1] **Oliver Hayden** [2] **Klaus Diepold** [1]

Abstract

Label-free approaches are attractive in cytological imaging due to their flexibility and cost efficiency. They are supported by machine learning methods, which, despite the lack of labeling and the associated lower contrast, can classify cells with high accuracy where the human observer has a little chance to discriminate cells. In order to better integrate these workflows into the clinical decision making process, this work investigates the calibration of confidence estimation for the automated classification of leukocytes. In addition, different visual explanation approaches are compared, which should bring machine decision making closer to professional healthcare applications. Furthermore, we were able to identify general detection patterns in neural networks and demonstrate the utility of the presented approaches in different scenarios of blood cell analysis.

1. Introduction

The complexity of deep learning models is growing in various fields. Medical imaging and clinical decision making are not exempt from this development (Holzinger et al., 2019; Guo et al., 2017). In recent years, deep learning has helped to support and automate many diagnoses, if it didn't even make them possible in the first place (Shen et al., 2017; Lundervold & Lundervold, 2019). Despite the potential benefits, doctors and patients remain skeptical about basing diagnosis and treatment on the output of black box models. Adversarial patterns and malfunctions could easily harm human life in these safety critical scenarios (Zeiler & Fergus, 2014; Rudin, 2019). Recently, the US Food & Drug Administration (FDA) approved several machine learning approaches in medical applications (Benjamens et al., 2020) but adoption could still be faster. From a scientific and regulatory perspective, the developers of such tools are particularly challenged to better address their target groups and to transparently communicate the performance as well as the limitations of their products (Holzinger et al., 2019; High-Level Expert Group on Artificial Intelligence, 2019; Rudin, 2019). In addition, the applications often lack appropriate customization for typical clinical workflows. Decisions are never made without sound evidence, and equipment must meet strict quality specifications. The target group is accustomed to particular types of visualizations that new technologies must adopt to have a chance of gaining trust (Evagorou et al., 2015; Vellido, 2020).

Quantitative Phase Imaging (QPI) is one of these new platform technologies that benefit greatly from the advances in computer vision and machine learning (Nguyen et al., 2022). Microscopes based on QPI are able to capture the optical height of cells without time consuming and costly fluorescence staining. Hence, many hematological (Go et al., 2018; Ozaki et al., 2019) and oncological (Nguyen et al., 2017; Ugele et al., 2018; Paidi et al., 2021) applications were demonstrated in this field. However, the resulting images are widely unknown to biomedical researchers and practitioners as they show only limited resemblance to light microscopy stained images (see Figure 1). Moreover, none of the previous publications put much emphasis on **visually explaining** the results of machine learning models or demonstrating the robustness of the approaches to perturbations.

In this work, we aim to transfer leukocyte classification as one of the most widely used laboratory tests (Horton et al., 2018) from molecular hematology to QPI and deep learning. To do this, we focus on rather small architectures like the *AlexNet* (Krizhevsky et al., 2012) and the *LeNet5* (LeCun et al., 1998), as the cell images do not require thousands of highly specialized filters, and even larger models would contradict our quest for transparency. In the following sections, we will provide a baseline for **differentiating four subtypes of leukocytes** with the proposed network architectures. Regarding **confidence estimation**, we will introduce modifications for variational inference and compare them to the frequentist approach. The predictions of the confidence calibrated models will then be used to test different **visual explanation tools** to support and communicate their decisions to the medical target group. Furthermore, we apply meta-aggregations to derive **general detection patterns** for the distinct cell classes dependent on their confidence

[*] Equal contribution [1] Chair of Data Processing, Technical University of Munich, Germany [2] Heinz-Nixdorf Chair of Biomedical Electronics, Technical University of Munich, Germany. Correspondence to: Stefan Röhrl <stefan.roehrl@tum.de>.

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

level. The more the network uses visual properties of the cell that are also important for human experts, the easier it becomes to justify the decisions. Finally, we will apply our findings to common obstacles in the cell analysis workflow and demonstrate the robustness and explainability of the architectures studied.

2. Background and Related Work

2.1. Confidence Estimation and Calibration

As in many safety critical scenarios, the safe use of clinical decision support systems (CDSSs) can only be ensured if the reliability and the limitations of the model can be accurately stated (High-Level Expert Group on Artificial Intelligence, 2019). Predictions with a low confidence level have to be checked by human experts, e.g. physicians, whereas the CDSS gets more autonomy in cases of high confidence. This approach borrows closely from human decision making, where trust is an important dimension of human interaction. Therefore, considering confidence estimations helps in interpreting predictions of deep learning algorithms and supports the development of a trustworthy interaction of a user with a CDSS (Guo et al., 2017).

A classification model is said to be *calibrated* if the prediction probability is equal to the actual probability of being correct. This behavior can be evaluated using a **reliability plot** (DeGroot & Fienberg, 1983; Niculescu-Mizil & Caruana, 2005), in which the accuracy of a model is plotted as a function of reliability. A perfectly calibrated model is represented as the identity function. For example, Guo et al. (2017) studied the confidence calibration of modern neural networks. While smaller neural networks, such as those proposed in LeCun et al. (1998) or Niculescu-Mizil & Caruana (2005), appear to produce well-calibrated confidence estimates, this is not true for more complex model architectures. Larger model architectures such as *AlexNet* (Krizhevsky et al., 2012) or *ResNet* (He et al., 2016) achieve better performance, they also tend to produce significantly higher confidence values compared to the achieved accuracy.

To counteract this behavior, several methods have been proposed for re-calibrating a model's confidence estimates in post-processing (Platt, 1999; Zadrozny & Elkan, 2002; Naeini et al., 2015; Kull et al., 2019). **Temperature scaling** has been shown to be effective for multiclass ($K > 2$) classification tasks. Here, the network logits $\mathbf{z}_i$ for the i-th sample for each class $k \in \{1, ..., K\}$ are scaled by a learned scalar parameter $T > 0$ before entering the softmax function

$$\sigma_{SM}\left(\frac{\mathbf{z}_i}{T}\right)^{(k)} = \frac{\exp\left(z_i^{(k)}/T\right)}{\sum_{j=1}^{K}\exp\left(z_i^{(j)}/T\right)}. \qquad (1)$$

Once the network is trained, T can be optimized based on the validation set. The scaling factor T does not affect the maximum of the softmax function and has in turn no negative impact on the model performance (Guo et al., 2017).

2.2. Visual Explanation of Deep Learning Models

In addition to calculating accurate confidence estimates, this work aims to improve the transparency of model predictions by providing visual explanations similar to Ghosal & Shah (2021) or Huang et al. (2021).

Model-agnostic methods impose no restrictions on the architecture or training of a model and are therefore flexible in their application. Furthermore, they do not affect the model performance while still offering intuitive explanations even for uninterpretable features of a black-box (Ribeiro et al., 2016b). The popular framework for Local Interpretable Model-agnostic Explanations (LIME) by Ribeiro et al. (2016a) approximates the local behavior of any machine learning model for a given input sample. For this, the interpretable representation of a sample $x \in \mathbf{R}^d$ is modeled as a binary vector $x' \in \{0,1\}^{d'}$ indicating the presence or absence of important features. For image classification, it is beneficial to apply this representation to contiguous patches, so-called *super-pixels*. Hence, the method is strongly dependent on the chosen segmentation algorithm like *Quickshift* (Vedaldi & Soatto, 2008), *SLIC* (Achanta et al., 2012), or *compact watershed* segmentation (Neubert & Protzel, 2014). The local behavior of the non-linear model $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is approximated by a surrogate model $g : \{0,1\}^{d'} \rightarrow \mathbb{R}$ in the linear form of $g(z') = w_g \cdot z'$, with z' being sampled from the neighborhood of x. An adaptation for LIME to work with Bayesian predictive models and approximate both mean and variance of an explanation from the underlying probabilistic model is given by Peltola (2018).

Propagation-based approaches, in contrast, use the internal structure of a neural network to determine the relevance of features to the model's internal decision-making. *Class Activation Mapping* (Zhou et al., 2016) demonstrated a way to retroactively add location information to a prediction, even though the convolutional layers solely acted as pattern detectors. Generalizing this approach to networks that additionally contain fully-connected layers, Selvaraju et al. (2017) proposed *Gradient-weighted Class Activation Mapping* (Grad-CAM). Another approach to extract information from the network's internals follows the principles of *Backpropagation*. This mechanism is commonly applied to train neural networks and trace back the output weights of a model to the actual feature map (Springenberg et al., 2015). Thus, gradient information that contributes to the prediction of a particular class, i.e., gradients with a positive sign, are propagated through the network and displayed as an explanation. *Guided Backpropagation* combines this gradient information with Grad-CAM to weight these potentially noisy explanations (Selvaraju et al., 2017).

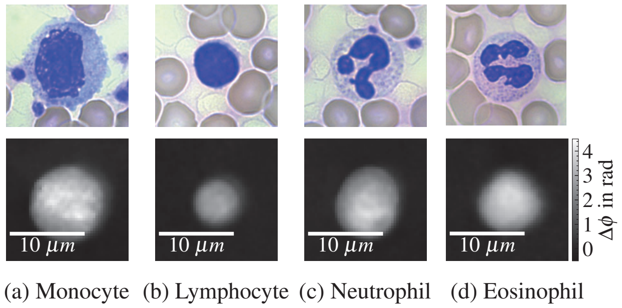


Figure 1. Subtypes of leukocytes: The upper row shows the cells under a light microscope with Giemsa staining (Barcia, 2007) on substrate. The lower row contains the corresponding phase images in suspension using monochromatic light at λ=528nm.

Meta-explanations are methods for aggregating individual explanations to extract general patterns and to draw conclusions about the model's overall behavior. This can be done by clustering the explanations (Lapuschkin et al., 2019), perform a layer-wise *relevance propagation* (Bach et al., 2015), or use *concept activation vectors* (Kim et al., 2018).

3. Methodology and Data Acquisition

3.1. Quantitative Phase Images of Leukocytes

The data used in this work was captured with a QPI microscope as used by Ugele et al. (2018) and Klenk et al. (2019). The liquid sample stream is focused by a microfluidics chip, allowing tens of thousands of cells to be imaged under near *in-vivo* conditions in a matter of minutes. The resulting phase images are 512×382 pixels in size, each containing multiple leukocytes. Background and noise subtraction is then performed to prepare the images for threshold segmentation and to separate the individual cells into single cell image patches. The entire preprocessing pipeline is described in Appendix A.1. Filtering out debris and defocused cells by requesting a *diameter* $\geq$ 4µm and a *circularity* $\geq$ 0.85 (see Appendix A.2), we obtained a set of N=11,008 leukocytes, balanced by their class label. They were randomly split into a training (70%), validation (20%) and test set (10%). Note that *Basophil* cells were excluded from the widely known *Five-Part Differential* data set, as it was not possible to prepare a sufficient number of cells, due to their natural sparsity and our limited number of healthy donors. Consequently, the data set consists of **Monocytes**, **Lymphocytes**, **Neutrophil** and **Eosinophil** cells, forming a *Four-Part Differential*. Typical examples of cell images are shown in Figure 1.

3.2. Experimental Setup and Metrics

In this work, we compare the performance of the larger *AlexNet* to the smaller *LeNet5* in the aforementioned four-part leukocyte differential. Thus, the last fully-connected layer was adapted to the four classes. As dropout layers are necessary to implement variational inference (VI), we introduced one after each fully-connected layer. The single cell patches of 50×50 pixels were scaled to the expected input dimensions. In case of AlexNet, the gray-scale phase images were replicated to three channels. For training, we used ADAM optimization (Kingma & Ba, 2015) with a cross-entropy loss function for N samples and $K = 4$ classes

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} y_{i,j} \log(p_{i,j}), \tag{2}$$

where $y_{i,j}$ is a binary indicator for a correct classification and $p_{i,j}$ is the prediction probability for an observation i of class j. The networks' classification performance is assessed using the measures of precision and recall

$$\text{Precision} = \frac{T_p}{T_p + F_p}, \quad \text{Recall} = \frac{T_p}{T_p + F_n} \tag{3}$$

as well as their harmonic mean

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

Predictions are gathered using *frequentist* deterministic forward-pushes, in the conventional case. The confidence estimation is derived from the softmax output. For variational inference, the dropout layers stay active during testing, resulting in a probabilistic behavior for a single input. These outputs are summarized as *mean*, *median* and *standard deviation* to form a prediction, and the values of 100 independent predictions to form the confidence score.

Besides the reliability plots described in Section 2.1, this work follows Naeini et al. (2015) for evaluating the confidence estimations. Clustering the described confidence estimations in $M = 10$ equally-spaced bins, we are able to estimate the expected calibration error

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \Big| \text{acc}(B_m) - \text{conf}(B_m) \Big| \tag{5}$$

as a term describing the average confidence/accuracy deviation of each bin B_m weighted by the number of contributing samples. To provide a lower quality bound, the maximum calibration error

$$\text{MCE} = \max_{m=1}^{M} \Big| \text{acc}(B_m) - \text{conf}(B_m) \Big| \tag{6}$$

was calculated analogously. Investigating the reliability of the demonstrated approaches, we conducted every experiment in 15 evaluation runs containing independent initializations and data splits.

4. Experiments

4.1. Model Performance with Variational Inference

As a baseline for the succeeding experiments, we compare our two model architectures in a frequentist and variational

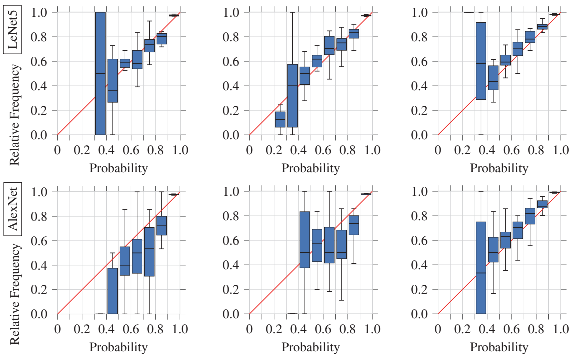
	Dropout	VI	Metric	Precision	Recall	F_1	Accuracy
LeNet5	p=0.00	✗	–	0.922 (1.0e-2)	0.923 (9.2e-3)	0.922 (9.4e-3)	0.927 (9.0e-3)
	p=0.25	✗	–	0.924 (9.9e-3)	0.926 (1.1e-2)	**0.925** (1.0e-2)	0.930 (9.1e-3)
	p=0.50	✗	–	0.910 (1.0e-2)	0.913 (1.0e-2)	0.911 (1.0e-2)	0.917 (9.4e-3)
	p=0.25	✓	mean	0.925 (8.4e-3)	0.927 (9.7e-3)	**0.926** (8.8e-3)	0.931 (7.6e-3)
	p=0.50	✓	mean	0.910 (1.1e-2)	0.916 (9.8e-3)	0.913 (1.0e-2)	0.918 (9.8e-3)
	p=0.25	✓	median	0.925 (9.1e-3)	0.926 (1.1e-2)	**0.924** (1.0e-2)	0.930 (8.8e-3)
	p=0.50	✓	median	0.909 (1.1e-2)	0.915 (1.0e-2)	0.911 (1.0e-2)	0.917 (9.4e-3)
AlexNet	p=0.00	✗	–	0.965 (5.2e-3)	0.962 (4.8e-3)	**0.963** (4.9e-3)	0.967 (4.1e-3)
	p=0.25	✗	–	0.963 (5.2e-3)	0.960 (6.1e-3)	0.962 (5.5e-3)	0.966 (4.4e-3)
	p=0.50	✗	–	0.963 (6.8e-3)	0.959 (5.9e-2)	0.961 (6.3e-3)	0.965 (5.9e-3)
	p=0.25	✓	mean	0.963 (5.2e-3)	0.960 (6.0e-3)	**0.962** (5.5e-3)	0.966 (4.4e-3)
	p=0.50	✓	mean	0.963 (6.8e-3)	0.959 (5.9e-3)	0.961 (6.3e-3)	0.965 (5.9e-3)
	p=0.25	✓	median	0.963 (5.2e-3)	0.960 (6.1e-3)	**0.962** (5.5e-3)	0.965 (4.4e-3)
	p=0.50	✓	median	0.963 (6.8e-3)	0.959 (5.9e-3)	0.961 (6.3e-3)	0.966 (5.9e-3)

Table 1. Classification results for the test set over 15 runs using frequentist (VI=✗) and variational inference (VI=✓). The table shows the averaged results. Standard deviation is stated in brackets.

inference setting. To consider both, precision and recall characteristics of the tested models, the F_1-score was used as key performance metric. In the case of a frequentist model, the model output was normalized using a softmax function and considered as the prediction value. The prediction values of the probabilistic models were calculated as the mean or median value of 100 independent forward pushes for each sample. Table 1 lists the performance on the test set of 15 independent runs. All model and training configurations showed convergence. In the frequentist setting the AlexNet (F_1=96.3%) reaches a slightly better performance than the LeNet5 (F_1=92.5%) and featured less variance. The impact of dropout regularization was rather low. For the LeNet5 a moderate dropout rate of p=0.25 even improved the classification performance. Hence, in further experiments we use a dropout rate p=0.50 for AlexNet and p=0.25 for LeNet5 architectures, if not stated differently.

4.2. Confidence Calibration

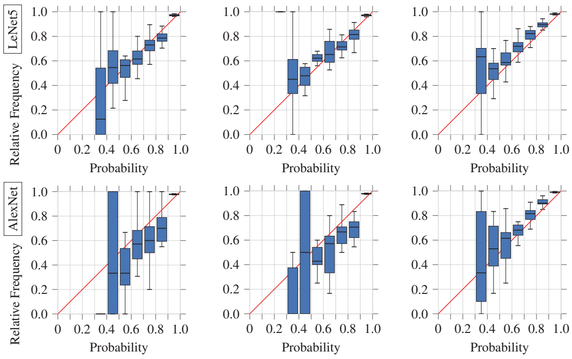
For qualitative analysis, we use reliability plots, which show the accuracy of a model as a function of a confidence score. To this end, the predictions were grouped into M=10 equal bins based on their respective confidence estimation. In case of a perfectly calibrated model, the empirical frequency should be an identity of the probability, as indicated with a red line in the following plots. If the frequency for a bin is below this line, the predictions are less accurate than the estimated confidence and the model becomes overconfident. We noticed that the frequencies show a high variance at lower probabilities and thus extended the vanilla reliability plots to box plots in the following figures. In the **frequentist** setting, Figure 2 reveals more stability and a better default calibration of the LeNet5. The larger AlexNet, in contrast, exposes unstable behavior and overconfidence. Applying temperature scaling to both of the models provided a more reliable estimate. The overconfidence reduces tremendously and especially the stability of AlexNet improves. Table 2 registers the effect of the calibration on the ECE and MCE, which in case can be improved by up to 53.8%.



(a) p=0.25 (b) p=0.50 (c) calibrated

Figure 2. Reliability plots for calibration of frequentist models

The probabilistic behavior of a **variational** model allows the generation of multiple independent predictions for each input sample. The mean and median of the observed output distributions were calculated, and the calibration of these metrics was analyzed. In addition, the standard deviation was interpreted as a measure of uncertainty, as opposed to a confidence measure. Similar to the frequentist approach, the LeNet5 provided fairly well calibrated confidence scores for mean and median predictions. The reliability plots in Figure 3 present only larger deviations for lower prediction values, which can be explained by the smaller number of relevant predictions. Also the AlexNet presents a better initial calibration than in the frequentist setting but is still slightly overconfident. The standard deviation of variational predictions provides useful information. Unlike the confidence scores shown before, the standard deviation does not contribute to the decision making process of the model but is interpreted as uncertainty measure. While mean and median are moderately suitable for calibration, standard deviation and temperature scaling provided the best variational confidence optimization in terms of ECE and MCE for both models. As listed in Table 2, especially the MCE as a worst case scenario could be reduced, which is crucial for the underlying medical application.



(a) p=0.25 (b) p=0.50 (c) calibrated

Figure 3. Reliability plots for calibration of variational models

				uncalibrated		calibrated		improvement	
	Dropout	VI	Metric	ECE	MCE	ECE	MCE	ECE	MCE
LeNet5	p=0.00	✗	–	0.34	3.8	**0.17**	3.0	50.0% ↑	21.1% ↗
	p=0.25	✗	–	0.25	4.0	0.18	3.7	28.0% ↑	7.5% →
	p=0.50	✗	–	**0.21**	3.0	0.20	2.8	4.8% →	6.7% →
	p=0.25	✓	mean	0.26	3.0	0.18	2.8	30.8% ↑	6.7% →
	p=0.50	✓	mean	0.27	2.9	0.21	2.6	22.2% ↗	10.3% →
	p=0.25	✓	median	0.26	3.1	0.18	2.9	30.8% ↑	6.5% →
	p=0.50	✓	median	0.25	3.0	0.24	2.8	4.0% →	6.7% →
	p=0.25	✓	std.dev.	**0.21**	**2.6**	0.19	2.6	9.5% →	0.0% →
	p=0.50	✓	std.dev.	0.25	3.2	0.25	**2.4**	0.0% →	25.0% ↑
AlexNet	p=0.00	✗	–	0.26	4.9	0.22	4.0	15.4% ↗	18.4% ↗
	p=0.25	✗	–	0.26	4.9	0.18	3.8	30.8% ↑	22.4% ↗
	p=0.50	✗	–	0.26	4.0	**0.12**	3.4	53.8% ↑	15.0% ↗
	p=0.25	✓	mean	0.25	4.5	0.14	3.5	44.0% ↑	22.2% ↗
	p=0.50	✓	mean	0.25	4.7	0.14	3.5	44.0% ↑	25.5% ↑
	p=0.25	✓	median	0.25	4.0	0.14	3.7	44.0% ↑	7.5% →
	p=0.50	✓	median	0.25	4.6	0.13	3.9	48.0% ↑	15.2% ↗
	p=0.25	✓	std.dev.	0.23	**3.2**	0.19	**3.0**	17.4% ↗	6.3% →
	p=0.50	✓	std.dev.	**0.18**	3.4	0.15	3.1	16.7% ↗	8.8% →

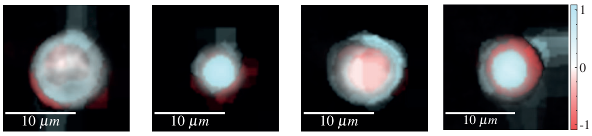
Table 2. Expected and maximum calibration error for all tested confidence measures averaged over 15 independent evaluation runs. Error values are stated in a magnitude of 10^{-1}.

In summary, the results of examining frequentist and variational inference methods for LeNet5 and AlexNet architectures are consistent with the observation that confidence estimates from larger models tend to be miscalibrated (Guo et al., 2017). The smaller LeNet5 generated well-calibrated confidence estimates with considerably low and consistent deviations from ideal behavior. The more complex AlexNet architecture provided better classification results, but also produced overconfident predictions. Temperature scaling enabled the implementation of a large AlexNet with good classification performance and well-calibrated confidence estimates. Consulting the results of Table 1 and 2, the experiments showed that the calibrated AlexNet architecture with the dropout rate of $p = 0.50$ achieved the best F_1-scores and the lowest ECE values of all tested models.

4.3. Visual Explanations

As not all of the tested explanation approaches provided useful results for quantitative phase images, which are not as rich in features as macroscopic images, we will only provide the results for LIME and Guided Backpropagation. The analysis of Occlusions, Backpropagation and Grad-CAM are stated in the appendix in section A.4.

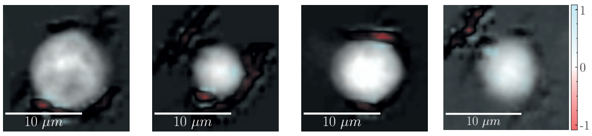
LIME explanations were not promising either, as their quality is highly dependent on the image segmentation approach used. Inspired by the principles of *tile coding* (Sherstov & Stone, 2005), best results were achieved by combining several sets of segmentations into one explanation. The interpretability was further improved by neglecting the original binary setting (Ribeiro et al., 2016a) and emphasizing the contributions of the individual areas according to their weight in the surrogate model. Figure 4 displays the results of the weighted outputs of LIME explanations on four superimposed SLIC segmentations (Further details on the optimization of the segmentation can be found in A.3). The blue areas indicate a positive correspondence of the underlying cell structures and the predicted label, red areas show an opposing relation. Where the exemplary Monocyte and Lymphocyte exhibit a supporting explanation, larger red areas for the Neutrophil and Eosinophil examples might require a double check by a physician or biologist.



(a) Monocyte (b) Lymphocyte (c) Neutrophil (d) Eosinophil

Figure 4. LIME explanations using SLIC-segmentation

Guided Backpropagation combines two approaches, by weighing the results of backpropagation with the class activation maps of Grad-CAM. Therefore, Guided Backpropagation cannot be applied to LeNet5. In most cases, the generated explanations in Figure 5 highlight only small parts of the cells, which could imply the detection of nucleus structures. For the samples of classes Lymphocyte and Eosinophil, the explanation also emphasizes minor gradients surrounding the actual cell, which could indicate that the size of the cell plays a role as other background parts in the distant corners are not affected.



(a) Monocyte (b) Lymphocyte (c) Neutrophil (d) Eosinophil

Figure 5. Explanations derived from Guided Backpropagation

4.4. Aggregated Meta-Explanations

With the huge number of cells to be analyzed, biomedical researchers only need the individual explanations in special cases. Usually, the general predictive behavior of the models is of greater interest. Therefore, in the following paragraphs, we will examine the models for general predictive patterns. One is based on ground truth labels and confidence scores, and the other is based on clustering methods. As LIME and Guided Backpropagation seemed to produce the most interpretable explanations, we will focus on those two approaches. To calculate the confidence estimates the variational scenario is used.

Aggregation based on labels and confidence estimations For aggregating the individual explanations, the confidence estimates were grouped into six equally sized bins and separated by their class label. The resulting averaged meta-explanations for the calibrated LeNet5 can be seen in Figure 6. Especially for the most certain group, two distinct patters can be observed: Monocytes and Eosinophils are

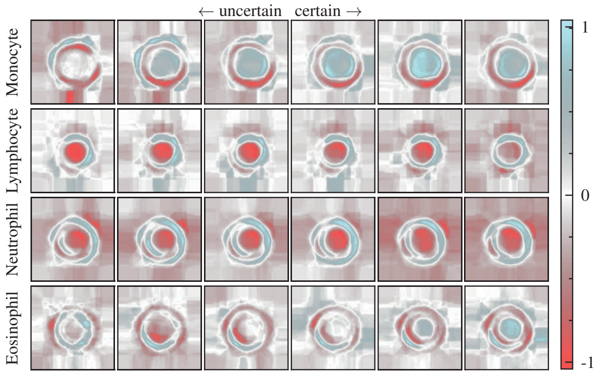


Figure 6. Aggregated LIME explanations based on calibrated confidence estimations for LeNet5

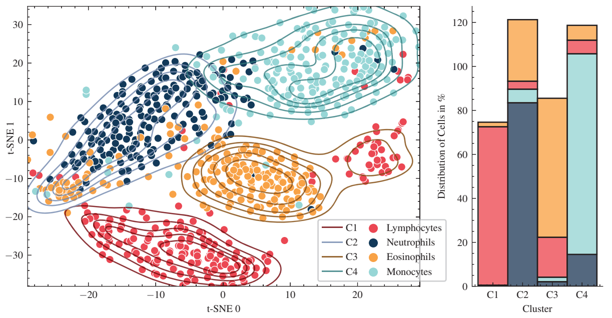


Figure 7. Clustering of t-SNE embedded LIME explanations from AlexNet by a k-Means algorithm. The clusters could not exactly assign all four classes. Therefore, the cumulative sum for some clusters is higher than 100% in the chart on the right.

represented by a stronger positive contribution of the inner part of the cells. In contrast, Neutrophils and Lymphocytes clearly depict a blue circle, which indicates the importance of the cell membrane. Furthermore, this behavior correlates with the biological appearance of the cells: The large Monocytes and small Lymphocytes can be easily differentiated from the other classes purely considering their size. For the more similar Neutrophils and Eosinophils, the network has learned to consider the cells' interior for one group to make a distinction.

In general, the prediction patterns for AlexNet and LeNet5 are similar. The LIME aggregations for AlexNet, displayed in Figure 19 (Appendix), entail an overall higher mean value, which makes the detected features less prominent. Also, the meta-explanations using guided backpropagation reveal a similar behavior. Figure 20 (Appendix) presents the same patterns for distinguishing the cells in their size as well as in their interior. For all classes the patterns get more precise with an increasing confidence estimate. Particularly, Eosinophils demonstrate the need for very confident estimates, to ensure that the correct parts of the image were analyzed for the clinical decision making.

Aggregation based on explanation clustering An alternative to the aggregation based on ground truth labels is the aggregation by unsupervised clustering. Here, we will see whether there are unique classification strategies that correspond to a particular cell type. In order to remain in a dimension that is manageable for humans, the high dimensional LIME explanations are embedded in a 2D space using t-SNE (van der Maaten & Hinton, 2008). This embedding is visualized in Figure 7, in which the color of the dots illustrates the respective cell class. Applying a k-Means clustering, with k equal to the number of classes, on this 2D space reveals distinct detection patterns for each individual class. Solely cluster C3, which is dominated by Eosinophils, incorporates an apostate group of Lymphocytes. This might be due to a limited capture quality, reduced sample purity or the fact that the network uses two distinct strategies to discover the small Lymphocytes. For the same reasons, also other clusters, especially C2, exhibit some mismatches as can be seen in the bar chart in Figure 7. Nevertheless, there is always one dominant cell class which supports our assumption that the network mainly relies on disjoint detection patterns for each of them.

5. Applications

After calibrating the confidence estimations and extracting patterns for the general predictive behavior of the networks, the presented techniques need to prove useful in real-world applications. Therefore, we confronted the variational setup for the LeNet5 with unknown data from familiar and unfamiliar domains and tested its classification confidence and the according visual explanations. We expect a high confidence only for leukocyte samples so the influence by unwanted objects stays at a minimum. In cases of overconfidence, the visual explanation should help to detect a violation of the general detection pattern in order to mitigate the interferences.

For an initial overview, we applied a train test split closer to real-world scenario to the leukocyte data. The test set of 1024 cells now consists exclusively of data from an independent donor, which was not present during training. To test resilience to typical error sources, we introduced two additional test sets: **Erythrocytes** make up 99% of human blood (Alberts, 2017), hence, it is likely that they find their way into leukocyte images. They should not be classified as leukocytes and need to exhibit a low confidence score. As the viscoelastic focusing by the microfluidics chip cannot guarantee perfect focus for all cells, **Defocused** examples should also be discarded by their low confidence score. In addition to erythrocytes and defocused cells, we picked two deviant test sets to simulate unfamiliar if not confusing inputs for the classifiers: The well known **MNIST** (LeCun et al., 2010) data set provides a similar image size but stands out with prominent edges. A data set of images with the

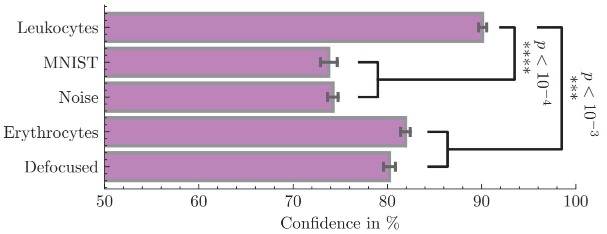
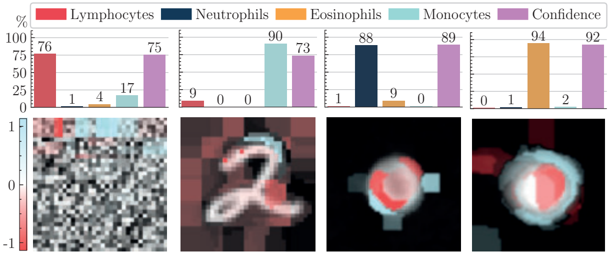


Figure 8. Confidence estimates by the LeNet5 for the different test sets. The error bars describe the standard error.

same dimensions but consisting purely of white **Noise** completes the list of challenges. The network could solve the leukocyte classification task for the new individual with an accuracy of 92,8% and a high confidence in its predictions, as Figure 8 displays. Additionally, the results demonstrate a general robustness against too deviant inputs. Calibrated on the standard deviation from variation inference, the confidence score shows a tremendous drop for MNIST and Noise images. Also, for the more closely related test sets, there is still a significant difference in the networks confidence, as determined by a *Kruskal-Wallis test* (Kruskal & Wallis, 1952) and post hoc analysis using *Bonferroni correction* (Armstrong, 2014).

5.1. Visual Inspection of Unknown Data

Even if the confidence estimation works well for most data and a clinical decision can be based exclusively on the most confident predictions, Figure 8 uncovers that there are still abnormal objects, which also reach a high confidence. Hence, Figure 9 investigates examples of unknown objects that could falsely contribute to the four-part differential. Here, the visual explanations of the noise patterns (a) and the MNIST image (b) show a totally divergent appearance which does not fit our general detection behavior. Some erythrocytes (c), nevertheless, could be too similar to leukocytes, as their outer cell membrane contributes positively to the prediction. Though, the inner torus shape should oppose a confident prediction.

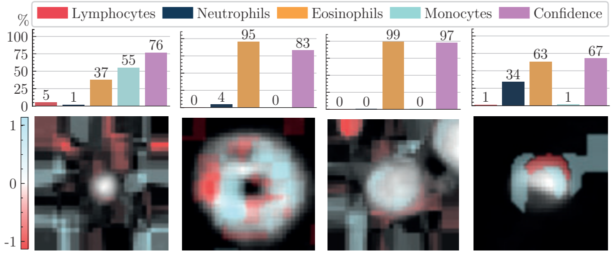


(a) Noise (b) MNIST (c) Erythrocyte (d) Leukocyte

Figure 9. Examples of unknown data might have a relatively high confidence estimate but stand out by their visual explanations. The respective bar plots visualize the predicted class estimates and the according confidence score in %.

5.2. Outlier Detection by Visual Inspection

Moving on from unknown data, the network also has to deal with cells and structures which were present during training but should not influence the classification results as they are no valid leukocytes. For this purpose, Figure 10 displays some examples of those *outliers*, their predicted class, and the according explanation. These are thrombocytes (a), defocused cells (b) or ruptured cells (d). Micro-Thrombotic events, also called aggregates (c), might have their relevance for certain diseases (Nishikawa et al., 2021) but are inconvenient for the four-part differential. Thrombocytes and ruptured cells should not be a big problem, as they show a conflicting explanation pattern. However, the thrombocyte has a rather high confidence score, which could be problematic. Also the defocused cell gets recognized, which is rather exceptional. The biggest problem still are aggregates as they contain more than one cell. The explanation in Figure 10c therefore has two contribution regions resulting in a high confidence, but two cells of different types would cancel each other out. Consequently, the proposed method offers only limited help for outlier detection, but the concerned objects are easily detectable using other methods. Stricter filter rules or more advanced techniques for this use case as presented by Röhrl et al. (2022) are strongly recommended.



(a) Thrombocyte (b) Defocused (c) Aggregate (d) Ruptured

Figure 10. Some kinds of outliers from the leukocyte data set might be difficult to detect as they show a high confidence and partly similar explanation patterns. The respective bar plots visualize the predicted class estimates and the according confidence score in %.

5.3. Mislabeled Data

Finally, there were some discrepancies during the training of the networks. Normally, we would expect the classification error to shrink with an increasing confidence, but for certain samples this was not the case. This is based on the fact that hardly any biological sample is of 100% purity. For the creation of the four-part differential training data set, this originates from the separation process of the individual cell types via *immunomagnetic isolation kits*. Here, paramagnetic antibodies are used to label and sort the cells. It might happen that some of the cells escape this labeling and contaminate the other classes. Modern isolation kits reach a purity of 95% and above (Son et al., 2017) and are constantly improved.

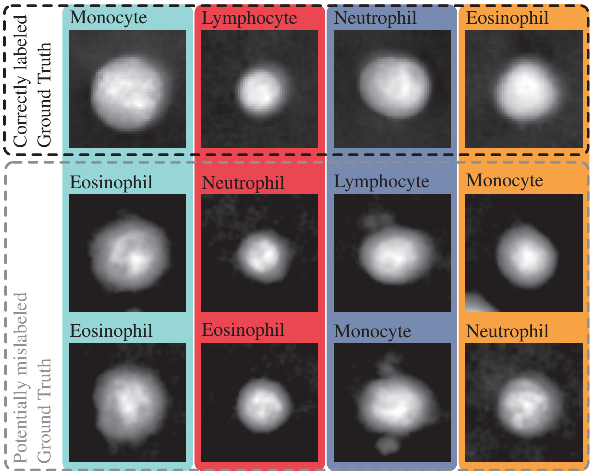


Figure 11. The top row shows valid representatives of the ground truth label. The lower rows contain potentially mislabeled cells that were assigned to another class by the LeNet5 with a confidence estimate $\geq 95\%$.

Nevertheless, our calibrated networks could demonstrate that they detected cells with a high confidence estimate for a potentially wrong class. Inspecting these images showed that the network might have become smarter than the ground truth, as Figure 11 reveals. Obviously, the cell in the top row has more resemblance to the cells drawn in the same column below than to the cell class the ground truth label would imply. Hence, the proposed method could be used in a *human-assisted labeling* setting (Holzinger, 2016) to further purify biological data sets.

6. Discussion and Conclusion

The goal of this work was to improve the interpretability of machine learning to overcome the limited applicability of algorithmic decision making in a clinical environment. For this purpose, we chose the ascending and label-free platform technology of QPI and performed a four-part differential of leukocytes, as there are many publications for the proof of concept but none which focus on transparency.

For the selected use case, the vanilla AlexNet model showed slightly better classification performance than the smaller LeNet5, but with a higher overestimation of its confidence. This drawback was overcome by introducing temperature scaling as an effective way to **calibrate** the confidence estimations. The application of variational inference further improved the consistency of the confidence estimation and reduced the ECE and MCE. Together with the high classification accuracy, these values can be used as a quality measure in a certification processes, when the presented techniques are integrated in a medical assay (Jin et al., 2023).

The comparison of state-of-the-art **visualization methods** for deep learning predictions outlined promising results for LIME and Guided Backpropagation. The methods facilitated the visualization of relevant decision factors for individual predictions. LIME was further adapted to convey the relative importance of individual image regions and to achieve explanations with higher granularity. It was possible to derive consistent meta-explanations and extract general detection patterns by aggregation or unsupervised clustering. Nevertheless, the appearing patterns had only limited resemblance with the **biological patterns** we hoped to find. As Figure 6 outlines, Monocytes and Lymphocytes are differentiated by their unique size. Eosinophils and Neutrophils generally have a similar appearance, but the networks are able to tell them apart based on their interior which seems to be more emphasized in the case of Eosinophils.

Applying the optimized technology to unknown data in **real-world scenarios** revealed high robustness against deviant cell structures and contamination. Certainly, leukocytes from new donors were accurately classified with a high degree of confidence. The calibrated confidence estimation even allowed the detection of mislabeled cells in the ground truth. For outlier detection like thrombocytes, aggregates, defocused or ruptured cells, the methods did not perform well and we recommend to use other methods for this purpose.

The high and robust classification performance without special labeling procedures demonstrates the maturity of the technologies presented in this and related work (Ugele et al., 2018; Shu et al., 2021; Fanous et al., 2022). However, if the essential explainability of the decisions is still missing and there is no relation to the established biological features, the clinical acceptance and a market entry will remain challenging. Therefore, in **future work** we want to use newer methods for visual explanations, such as *Smooth Grad-CAM++* (Omeiza et al., 2019), *EVET* (Oh et al., 2021), or *FIMF score-CAM* (Li et al., 2023) to help QPI make the expected breakthrough (Nguyen et al., 2022). We hope that others will also decide to integrate visual explanation and confidence calibration instead of focusing only on accuracy, in order to promote the discourse and enable interdisciplinary work on this topic (Yang et al., 2020).

All in all, the work contributes to making deep learning more transparent and communicable for the investigated use case. It represents a development in the right direction of **interpretable** machine learning in this field and lays the foundation for subsequent user studies in biomedical research and clinical application.

Acknowledgements

The authors would like to especially honor the contributions of J. Groll for the software implementation and experiments. This research was funded by the German Federal Ministry for Education and Research (BMBF) with the funding ID ZN 01 | S17049.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

Alberts, B. *Molecular biology of the cell*. WW Norton & Company, 2017.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. A unified view of gradient-based attribution methods for Deep Neural Networks. In *NIPS Workshop on Interpreting, Explaining and Visualizing Deep Learning*, 2017.

Armstrong, R. A. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, 34(5):502–508, 2014.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Barcia, J. J. The Giemsa stain: Its History and Applications. *International Journal of Surgical Pathology*, 15(3):292–296, 2007.

Benjamens, S., Dhunnoo, P., and Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *Nature Partner Journals: Digital Medicine*, 3(1):1–8, 2020.

DeGroot, M. H. and Fienberg, S. E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society*, 32:12–22, 1983.

Evagorou, M., Erduran, S., and Mäntylä, T. The role of visual representations in scientific practices: from conceptual understanding and knowledge generation to 'seeing'how science works. *International Journal of STEM Education*, 2(1):1–13, 2015.

Fanous, M. J., He, S., Sengupta, S., Tangella, K., Sobh, N., Anastasio, M. A., and Popescu, G. White blood cell detection, classification and analysis using phase imaging with computational specificity (pics). *Scientific reports*, 12(1):20043, 2022.

Fong, R. C. and Vedaldi, A. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3429–3437. 2017.

Ghosal, S. and Shah, P. A deep-learning toolkit for visualization and interpretation of segmented medical images. *Cell reports methods*, 1(7):100107, 2021.

Go, T., Kim, J. H., Byeon, H., and Lee, S. J. Machine learning-based in-line holographic sensing of unstained malaria-infected red blood cells. *Journal of Biophotonics*, 11(9):e201800101, 2018.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pp. 630–645. 2016.

High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. European Commission, 2019.

Holzinger, A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131, 2016.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4): e1312, 2019.

Horton, S., Fleming, K. A., Kuti, M., Looi, L.-M., Pai, S. A., Sayed, S., and Wilson, M. L. The Top 25 Laboratory Tests by Volume and Revenue in Five Different Countries. *American Journal of Clinical Pathology*, 151(5):446–451, 2018.

Huang, J.-H., Yang, C.-H. H., Liu, F., Tian, M., Liu, Y.-C., Wu, T.-W., Lin, I., Wang, K., Morikawa, H., Chang, H., et al. DeepOpht: Medical Report Generation for Retinal Images via Deep Models and Visual Explanation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2442–2452, 2021.

Jin, W., Li, X., Fatehi, M., and Hamarneh, G. Guidelines and evaluation of clinical explainable ai in medical image analysis. *Medical Image Analysis*, 84:102684, 2023.

Kasprowicz, R., Suman, R., and O'Toole, P. Characterising live cell behaviour: Traditional label-free and quantitative phase imaging approaches. *International Journal of Biochemistry and Cell Biology*, 84:89–95, 2017.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *Proceedings of the 35th International Conference on Machine Learning*, 6:4186–4195, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Klenk, C., Heim, D., Ugele, M., and Hayden, O. Impact of sample preparation on holographic imaging of leukocytes. *Optical Engineering*, 59(10):102403, 2019.

Kotsiantis, S. B., Kanellopoulos, D., and Pintelas, P. E. Data preprocessing for supervised leaning. *International Journal of Computer and Information Engineering*, 1(12): 4104–4109, 2007.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25, pp. 1097–1105, 2012.

Kruskal, W. H. and Wallis, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

Kull, M., Nieto, M. P., Kangsepp, M., Silva Filho, T., Song, H., and Flach, P. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems*, pp. 12295–12305. 2019.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 11:2278–2324, 1998.

LeCun, Y., Cortes, C., and Burges, C. MNIST handwritten digit database, 2010. URL `http://yann.lecun.com/exdb/mnist`.

Li, J., Zhang, D., Meng, B., Li, Y., and Luo, L. FIMF score-CAM: Fast score-CAM based on local multi-feature integration for visual interpretation of CNNS. *IET Image Processing*, 17(3):761–772, 2023.

Lundervold, A. S. and Lundervold, A. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015.

Neubert, P. and Protzel, P. Compact watershed and preemptive SLIC: On improving trade-offs of superpixel segmentation algorithms. In *Proceedings of the 22nd International Conference on Pattern Recognition*, pp. 996–1001. IEEE, 2014.

Nguyen, T. H., Sridharan, S., Macias, V., Kajdacsy-Balla, A., Melamed, J., Do, M. N., and Popescu, G. Automatic Gleason grading of prostate cancer using quantitative phase imaging and machine learning. *Journal of Biomedical Optics*, 22(3):036015, 2017.

Nguyen, T. L., Pradeep, S., Judson-Torres, R. L., Reed, J., Teitell, M. A., and Zangle, T. A. Quantitative phase imaging: Recent advances and expanding potential in biomedicine. *American Chemical Society Nano*, 16(8): 11516–11544, 2022.

Niculescu-Mizil, A. and Caruana, R. Predicting Good Probabilities With Supervised Learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632. 2005.

Nishikawa, M., Kanno, H., Zhou, Y., Xiao, T.-H., Suzuki, T., Ibayashi, Y., Harmon, J., Takizawa, S., Hiramatsu, K., Nitta, N., et al. Massive image-based single-cell profiling reveals high levels of circulating platelet aggregates in patients with COVID-19. *Nature Communications*, 12(1): 7135, 2021.

Oh, Y., Jung, H., Park, J., and Kim, M. S. EVET: Enhancing Visual Explanations of Deep Neural Networks Using Image Transformations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3578–3586, 2021.

Omeiza, D., Speakman, S., Cintas, C., and Weldermariam, K. Smooth Grad-CAM++: An enhanced inference level visualization technique for deep convolutional neural network models. *arXiv preprint arXiv:1908.01224*, 2019.

Ozaki, Y., Yamada, H., Kikuchi, H., Hirotsu, A., Murakami, T., Matsumoto, T., Kawabata, T., Hiramatsu, Y., Kamiya, K., Yamauchi, T., Goto, K., Ueda, Y., Okazaki, S., Kitagawa, M., Takeuchi, H., and Konno, H. Label-free classification of cells based on supervised machine learning of subcellular structures. *PloS one*, 14:1–20, 01 2019.

Paidi, S. K., Raj, P., Bordett, R., Zhang, C., Karandikar, S. H., Pandey, R., and Barman, I. Raman and quantitative phase imaging allow morpho-molecular recognition of malignancy and stages of B-cell acute lymphoblastic leukemia. *Biosensors and Bioelectronics*, 190:113403, 2021.

Peltola, T. Local Interpretable Model-agnostic Explanations of Bayesian Predictive Models via Kullback-Leibler Projections. *arXiv preprint arXiv:1810.02678*, 2018.

Platt, J. C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, 10:61–74, 1999.

Ribeiro, M. T., Singh, S., and Guestrin, C. ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining SIGKDD*, pp. 1135–1144, 2016a.

Ribeiro, M. T., Singh, S., and Guestrin, C. Model-Agnostic Interpretability of Machine Learning. *arXiv preprint arXiv:1606.05386*, 2016b.

Röhrl, S., Ugele, M., Klenk, C., Heim, D., Hayden, O., and Diepold, K. Autoencoder features for differentiation of leukocytes based on digital holographic microscopy (dhm). In *Computer Aided Systems Theory EUROCAST*, pp. 281–288, Cham, 2020. Springer International Publishing.

Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, pp. 206–215, 2019.

Röhrl, S., Hein, A., Huang, L., Heim, D., Klenk, C., Lengl, M., Knopp, M., Hafez, N., Hayden, O., and Diepold, K. Outlier detection using self-organizing maps for automated blood cell analysis. In *Interpretable Machine Learning in Healthcare Workshop, ICML*, 2022.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Shen, D., Wu, G., and Suk, H.-I. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221, 2017.

Sherstov, A. A. and Stone, P. Function approximation via tile coding: Automating parameter choice. In *Proceedings of the International Symposium on Abstraction, Reformulation, and Approximation*, pp. 194–205. Springer, 2005.

Shu, X., Sansare, S., Jin, D., Zeng, X., Tong, K.-Y., Pandey, R., and Zhou, R. Artificial-intelligence-enabled reagent-free imaging hematology analyzer. *Advanced Intelligent Systems*, 3(8):2000277, 2021.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the 2nd International Conference on Learning Representations*, pp. 1–8, 2014.

Singh, D. and Singh, B. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97:105524, 2020.

Son, K., Mukherjee, M., McIntyre, B. A., Eguez, J. C., Radford, K., LaVigne, N., Ethier, C., Davoine, F., Janssen, L., Lacy, P., and Nair, P. Improved recovery of functionally active eosinophils and neutrophils using novel immunomagnetic technology. *Journal of Immunological Methods*, 449:44–55, 2017.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *Proceedings of the 3rd International Conference on Learning Representations*, pp. 1–14, 2015.

Suzuki, S. and Abe, K. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics and Image Processing*, 30(1):32–46, 1985.

Ugele, M., Weniger, M., Stanzel, M., Bassler, M., Krause, S. W., Friedrich, O., Hayden, O., and Richter, L. Label-Free High-Throughput Leukemia Detection by Holographic Microscopy. *Advanced Science*, 5(12):1800761, 2018.

van der Maaten, L. and Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605, 2008.

Vedaldi, A. and Soatto, S. Quick shift and kernel methods for mode seeking. In *European Conference on Computer Vision*, pp. 705–718. 2008.

Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, 32(24): 18069–18083, 2020.

Yang, F., Huang, Z., Scholtz, J., and Arendt, D. L. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 189–201, 2020.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

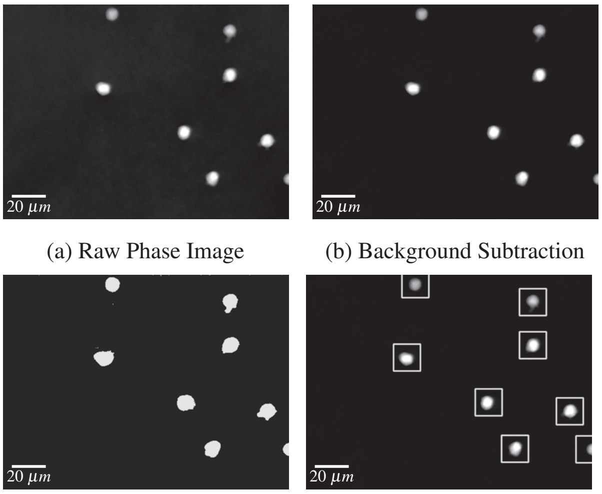
Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Visio*, pp. 818–833. 2014.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pp. 2921–2929, 2016.

A. Appendix

A.1. Image Preprocessing

To achieve satisfactory segmentation and classification results, it is crucial to perform preprocessing. Figure 12a shows that the QPI configuration produces phase images of numerous cells with dimensions of 512×382 pixels. From this, we need to extract patches of 50×50 pixels containing only single cells in order to classify them properly.



(a) Raw Phase Image (b) Background Subtraction

(c) Threshold Segmentation (d) Cell Image Patches

Figure 12. Preprocessing steps to achieve single cell image patches from a raw phase image

Background Subtraction To eliminate unwanted artifacts and background noise, the median of 100 images is computed and then subtracted from each frame. The transition from Figure 12a to Figure 12b visualizes the achieved smoothness in the image background. This operation is possible since the imaging setup and microfulidics channel are regarded as stationary.

Segmentation Cell detection in the acquired images involves two steps: thresholding and contour finding. First, the phase images are subjected to binary thresholding. Next, contours are extracted from the binary images using the algorithm introduced by Suzuki & Abe (1985). These extracted contours are then subjected to filtering based on a minimum contour area. Finally, each cell represented by a contour is saved as an image patch with dimensions of 50×50 pixels for further analysis. Compare Figures 12c and 12d.

Normalization Normalization is an essential step in preparing data for machine learning, especially when using neural networks, as it standardizes the feature or image values to a uniform range. The most common techniques are either mean and standard deviation based (such as z-score normalization) or minimum-maximum based (Singh & Singh, 2020; Kotsiantis et al., 2007). In this work, the images were first clipped to limit the range of values, since images produced by holographic microscopes theoretically have an unlimited range of values. Specifically, a minimum clipping value of 0.2 (due to the background) and a maximum clipping value of 4 were used, which proved effective in capturing leukocytes while minimizing cell clipping and utilizing the entire value range. *Min-Max normalization* was then used to transform the image values to the interval $[0, 1]$, which is ideal for neural networks.

A.2. Morphological Features

Hand crafted features are widely spread in the cytology community. Therefore, we adapted their use in our work to perform some kind of quality control. Table 3 shows a subset of the features introduced by Kasprowicz et al. (2017), Ugele et al. (2018), and Paidi et al. (2021) which are sufficient to filter out artifacts and impurities of the blood samples. These features are manly based on OpenCV contours[1] and the contained pixel values. They typically look like the contour drawn in red on the cell image in Figure 13. As texture features have proven to be insufficient for robust cell classification we do not consider them in this work (Röhrl et al., 2020).

	Feature	Explanation	Unit
P	# pixels	Number of pixels per cell contour	-
ϕ_i	phase shift	measured phase shift of the i-th pixel	rad
λ	wave length	wave length of the light source (528nm)	nm
A	area	$P \cdot$ pixel area	µm^2
d	diameter	$\sqrt{\frac{4A}{\pi}}$	µm
V	optical volume	$\sum^P \phi_i \cdot \frac{\lambda}{2\pi}$	µm^3
L	perimeter	OpenCV `arcLength()` of the cell contour	µm
C	circularity	$\frac{4\pi \cdot A}{L^2}$	-

Table 3. Morphological Features (excerpt) adapted from (Kasprowicz et al., 2017; Ugele et al., 2018; Paidi et al., 2021)

[1] https://docs.opencv.org/3.4/dd/d49/tutorial_py_contour_features.html

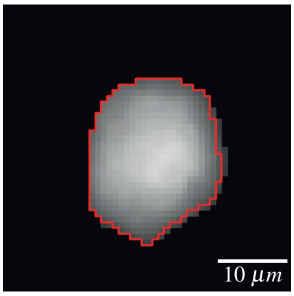


Figure 13. Cell image with its detected contour

A.3. LIME Segmentation

The generation of meaningful LIME explanations requires an interpretable data representation. In a first step, different algorithms were implemented and compared to calculate a consistent segmentation of the cell images. To evaluate these, segmentation results for individual samples of all relevant classes of leukocytes were manually reviewed. The evaluation revealed that all tested algorithms require careful tuning of the respective parameters to achieve satisfactory outputs for all relevant cell types. Due to the high contrast between the actual cell and the background of an image, reasonable segmentation had to be ensured to differentiate the individual parts within a cell. This was necessary to obtain granular explanations that take into account both the background of an image and the internal structure of the captured cells. (Compare Figure 14)

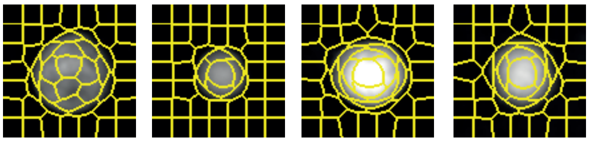
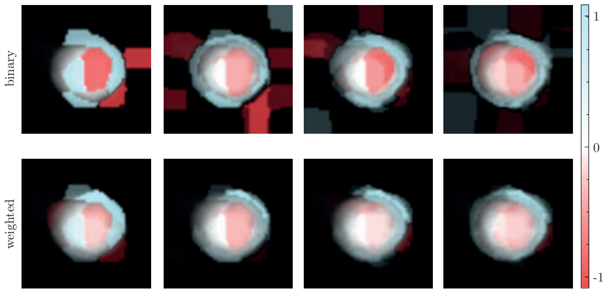

(a) Monocyte (b) Lymphocyte (c) Neutrophil (d) Eosinophil

Figure 14. SLIC segmentation of cell images as a pre-processing step for LIME explanations

In order to enable LIME to also evaluate more granular regions, we tested the options to increase the number of segments per explanation or to combine the outputs from several segmentations for the same sample image. The first approach, to simply increase the number of segments and thus yielding a more detailed resolution, resulted in noisy and complex interpretations, which were counterintuitive. Combining and weighing several segmentations with different but constant numbers of segments was promising as can be seen in Figure 15. The final segmenter consists of S=4 individually configured SLIC approaches. Detailed settings for the SLIC segmentations can be found in Table 4. The weighted results of the according LIME explanations are then merged into one mask via averaging.



(a) S=1 (b) S=2 (c) S=3 (d) S=4

Figure 15. Effects of varying the number of S SLIC segmentations on the weighted and binary LIME explanations

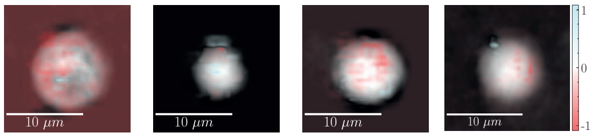
Name	Segments	Compactness	Sigma
SLIC$_1$	15	10	3.0
SLIC$_2$	25	10	2.5
SLIC$_3$	35	25	3.0
SLIC$_4$	50	15	5.0

Table 4. Parameterization of the used SLIC segmentations

A.4. Visual Explanation

The visual explanation methods in this section were implemented and tested on the leukocyte data set presented. Unfortunately, they did not prove to be very helpful for our use case, but we would still like to show the results for comparison.

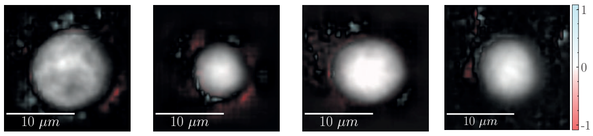
Perturbation-based approaches provide explanations by analyzing the effects of local changes on a model's response. These can also be model-agnostic as in case of simple occlusions (Zeiler & Fergus, 2014). Here, different image areas are systematically covered to determine the influence of the respective feature. To also detect cross-relationships between different areas, model-specific gradient information needs to be considered (Simonyan et al., 2014; Ancona et al., 2017). So called *meaningful perturbation* was introduced by Fong & Vedaldi (2017) to achieve more natural and plausible imaging. Instead of covering individual areas of an image with a black square, random noise and blur are applied to erase information in these specific areas. In the following example, simple occlusion was used with a patch of the size of 6×6 pixels to iteratively cover certain parts of an input image of 50×50 pixels. By observing the resulting changes in the prediction values, we calculate a sensitivity value for each pixel as shown in Figure 16. While the explanations roughly highlight relevant areas of an image, it is difficult to correlate the results with the underlying cells. Additionally, we noticed a high impact of the chosen patch size on the resulting sensitivity values, leading to inconsistent results.



(a) Monocyte (b) Lymphocyte (c) Neutrophil (d) Eosinophil

Figure 16. Explanations derived from Occlusion

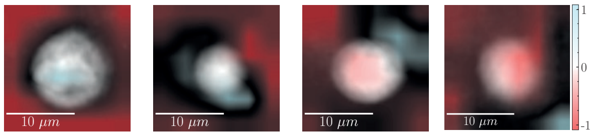
Backpropagation uses the inner structure of the analyzed deep learning model to pipe back the prediction value to the initial input space. The resulting explanations give an indication of which patterns in the cell image triggered the activation of the neural networks. Therefore, the explanations presented are highly dependent on the actual size of the input space. The explanations for an AlexNet model, displayed in Figure 17, have a higher resolution compared to a LeNet5 model and are thus easier to interpret. Although the interpretation of these patterns is not obvious, certain parts can be attributed to either an internal structure of a cell or the high contrast of the outer membrane.



(a) Monocyte (b) Lymphocyte (c) Neutrophil (d) Eosinophil

Figure 17. Explanations derived from Backpropagation

Grad-CAM explanation focuses on important regions of the image and produces much smoother results than the previously shown Backpropagation approach. However, this technique requires that the dimension of the last convolution block of the model is multidimensional, thus preventing its application to the LeNet5. The final convolutional layer of the implemented AlexNet architecture consists of filters with a size of 13×13. The total activation of this filter was aggregated and interpolated to fit the original, higher-dimensional input space. Therefore, the class activation maps had a low resolution, which directly depended on the underlying model architecture. As shown in Figure 18, Grad-CAM can be used as a basic method to validate relevant domains for a model but at the same time, the information is limited and does not allow for further differentiation.



(a) Monocyte (b) Lymphocyte (c) Neutrophil (d) Eosinophil

Figure 18. Explanations derived from Grad-CAM

A.5. Aggregation

For the AlexNet architecture it was also possible to extract general detection patterns for the different leukocyte classes. The pattern for LIME does not change that much as plotted in Figure 19. On the other hand, Guided Backpropagation produces clearly evolving patterns with an increasing confidence, which can be seen in Figure 20. First, the detection pattern focuses much more on the background, whereas for a higher confidence score, the attention moves towards the actual cells.

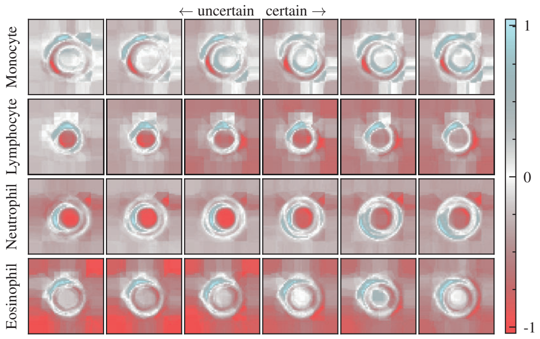


Figure 19. Aggregated LIME explanations based on calibrated confidence estimations for AlexNet

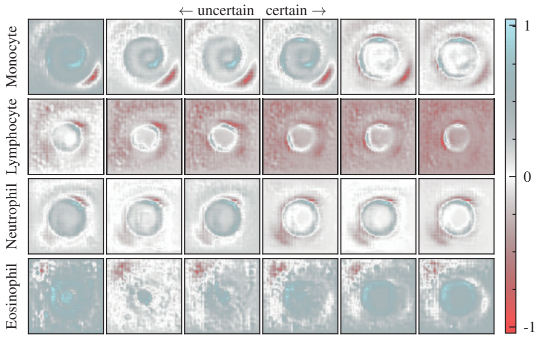
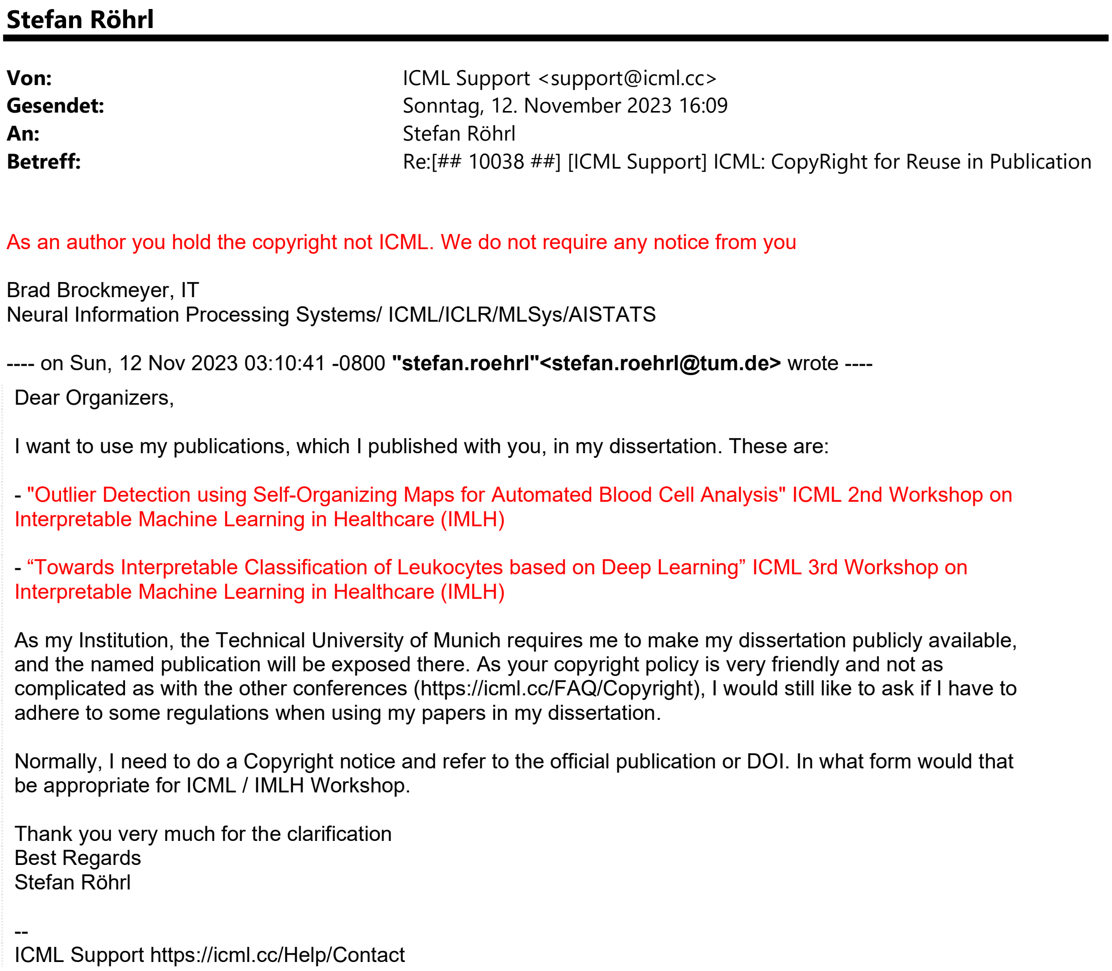


Figure 20. Aggregated Guided Backpropagation explanations based on calibrated confidence estimations for AlexNet

Copyright

No permission from the publisher is necessary as the authors hold the exclusive copyright. No rights were transferred to ICML.



Stefan Röhrl

Von:	ICML Support <support@icml.cc>
Gesendet:	Sonntag, 12. November 2023 16:09
An:	Stefan Röhrl
Betreff:	Re:[## 10038 ##] [ICML Support] ICML: CopyRight for Reuse in Publication

As an author you hold the copyright not ICML. We do not require any notice from you

Brad Brockmeyer, IT
Neural Information Processing Systems/ ICML/ICLR/MLSys/AISTATS

---- on Sun, 12 Nov 2023 03:10:41 -0800 **"stefan.roehrl"<stefan.roehrl@tum.de>** wrote ----

> Dear Organizers,
>
> I want to use my publications, which I published with you, in my dissertation. These are:
>
> - "Outlier Detection using Self-Organizing Maps for Automated Blood Cell Analysis" ICML 2nd Workshop on Interpretable Machine Learning in Healthcare (IMLH)
>
> - "Towards Interpretable Classification of Leukocytes based on Deep Learning" ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)
>
> As my Institution, the Technical University of Munich requires me to make my dissertation publicly available, and the named publication will be exposed there. As your copyright policy is very friendly and not as complicated as with the other conferences (https://icml.cc/FAQ/Copyright), I would still like to ask if I have to adhere to some regulations when using my papers in my dissertation.
>
> Normally, I need to do a Copyright notice and refer to the official publication or DOI. In what form would that be appropriate for ICML / IMLH Workshop.
>
> Thank you very much for the clarification
> Best Regards
> Stefan Röhrl
>
> --
> ICML Support https://icml.cc/Help/Contact

Core Publication V

Explainable Artificial Intelligence for Cytological Image Analysis

Stefan Röhrl, Hendrik Maier, Manuel Lengl, Christian Klenk, Dominik Heim, Martin Knopp, Simon Schumann, Oliver Hayden, Klaus Diepold

Summary This paper introduces a prototype of an eXplainable Artificial Intelligence (XAI) dashboard to increase confidence and understanding of machine learning results. The dashboard incorporates various explanation methods and design adaptations tailored to biomedical research concepts. A user study involving data scientists and biomedical researchers evaluates the dashboard's effectiveness in improving understanding, bias detection, and trustworthiness. Results indicate significant improvements over unexplained performance reports, with certain modules more appealing to specific user groups. However, there is a tendency for users to overestimate the trustworthiness of algorithms compared to their understanding of their behavior and bias detection capabilities. The results highlight the importance of domain-specific explanations and diverse approaches to facilitate collaborative interdisciplinary research.

Own Contributions

- Initiating the research idea
- Gathering of related work and assessing the scientific context of this work
- Planning the design of dashboard prototype
- Planning of the user study and gathering the study cohort
- Writing the complete manuscript and designing the figures and plots
- Analyzing and interpreting the study results
- Defending the work and integrating the reviewer feedback

S. Röhrl, H. Maier, M. Lengl, C. Klenk, D. Heim, M. Knopp, S. Schumann, O. Hayden, and K. Diepold. "Explainable Artificial Intelligence for Cytological Image Analysis". In: *Artificial Intelligence in Medicine.* Springer Nature Switzerland, 2023, pp. 75–85 Online: https://link.springer.com/chapter/10.1007/978-3-031-34344-5_10

Explainable Artificial Intelligence for Cytological Image Analysis

Stefan Röhrl[1](✉), Hendrik Maier[1], Manuel Lengl[1], Christian Klenk[2], Dominik Heim[2], Martin Knopp[1,2], Simon Schumann[1], Oliver Hayden[2], and Klaus Diepold[1]

[1] Chair of Data Processing, Technical University of Munich, Munich, Germany
stefan.roehrl@tum.de

[2] Heinz-Nixdorf Chair of Biomedical Electronics, Technical University of Munich, Munich, Germany

Abstract. Emerging new technologies are entering the medical market. Among them, the use of Machine Learning (ML) is becoming more common. This work explores the associated Explainable Artificial Intelligence (XAI) approaches, which should help to provide insight into the often opaque methods and thus gain trust of users and patients as well as facilitate interdisciplinary work. Using the differentiation of white blood cells with the aid of a high throughput quantitative phase microscope as an example, we developed a web-based XAI dashboard to assess the effect of different XAI methods on the perception and the judgment of our users. Therefore, we conducted a study with two user groups of data scientists and biomedical researchers and evaluated their interaction with our XAI modules, with respect to the aspects of behavioral understanding of the algorithm, its ability to detect biases and its trustworthiness. The results of the user tests show considerable improvement achieved through the XAI dashboard on the measured set of aspects. A deep dive analysis aggregated on the different user groups compares the five implemented modules. Furthermore, the results reveal that using a combination of modules achieves higher appreciation than the individual modules. Finally, one observes a user's tendency of overestimating the trustworthiness of the algorithm compared to their perceived abilities to understand the behavior of the algorithm and to detect biases.

Keywords: XAI · Quantitative Phase Imaging · Blood Cell Analysis

1 Introduction

The current gold standard sending and presenting hematological laboratory results is in tabular form with numbers and benchmarks. There are neither detailed insights provided on the methodology nor the possibility to interpret or question the results. With current medical analysis, this information it is not relevant for physicians and patients, as they are mainly interested in the plain results. However, as machine learning (ML) comes into place, the need for additional information increases. The workflow will also change for laboratory personnel and pathologists who are directly interacting with the algorithms

J. M. Juarez et al. (Eds.): AIME 2023, LNAI 13897, pp. 75–85, 2023.
https://doi.org/10.1007/978-3-031-34344-5_10

and responsible for the correctness of the result. Since one decade, deep learning models have pushed the boundaries in various fields of ML [11] and have shown a superior performance also in computer vision [9]. However, the increasing model complexity comes at the cost of interpretability. This lack of transparency is a problem that has been recognized broadly in legislation as well as academia in the recent years [6,18,20].

After the GDPR was introduced, in 2018, an independent expert group was set up by the European Commission to further evaluate implications for the research and deployment of AI, resulting in the *Ethics Guidelines For Trustworthy AI* [5]. They note that especially transparency, which state of the art models are missing, is the key to establish trust and accordingly they demand traceability, explainability and adequate communication.

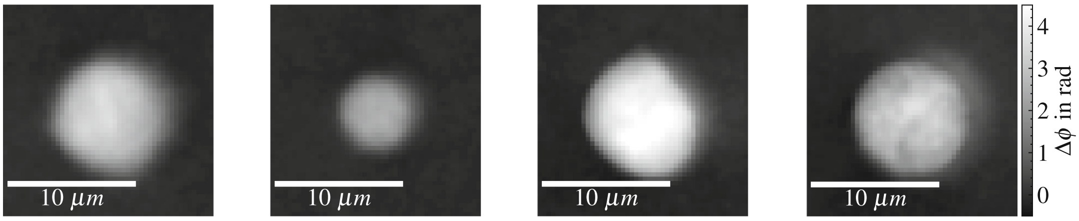
To demonstrate our work, we investigate the use case of hematological analysis, which is one of the most common laboratory tests [7], as it delivers comprehensive information about the health status of an organism. In contrast to conventional blood analysis via molecular labeling [16] or the gold standard blood smear [2], we focus on the ascending *quantitative phase imaging* approach combined with *microfluidics*. As a quantitative phase microscope offers a higher dynamic range than an unstained bright-field microscope, this technology works label-free and, therefore, needs no time-consuming sample preparation. The sample presentation via a microfluidics channel leverages the approach to a high statistical power, while keeping the cells near *in vivo* conditions. Various publications demonstrate its diverse potentials and versatility in the domains of oncology [10,13], hematology [15,19] and beyond [14].

Alongside with all its advantages, this new platform technology comes with several challenges. Besides the inexactness of viscoelastic focusing and orientation, the problem changes from classical cell sorting to a computer vision problem which is preferably solved by machine learning [9,14].

2 Background

2.1 Differential Blood Count Using Quantitative Phase Imaging

Blood contains a vast amount of information about the state of human health. Especially the composition of white blood cells (WBCs) and their functions form the basis for the detection of hematological, oncological or immunological diseases. However, many biomarkers remain hidden due to the technical limitations of conventional analyzers. This may be due to volatility of some biomarkers, the insufficient contrast or resolution provided by optical methods or the lack of suitable antibodies for fluorescent staining [16]. Using a quantitative phase microscope, the problem transitions into the domain of object detection, pattern recognition and classification. Skipping the tedious sample preparation, the measurement can be performed within 15 min after blood draw, which paves the way to the analysis of the kinetics of intra-cellular changes closer to the point of care [2,8]. Though, before pursuing the analysis of internal cell structures and morphological changes, the *Five-Part Differential* of WBCs has to be established using computer vision and machine learning techniques. For healthy individuals, Neutrophils (62%) make up the biggest proportion, followed by Lymphocytes (30%), Monocytes (5.3%), Eosinophils (2.3%) and Basophils (0.4%) [1]. Manual staining or costly molecular labeling are currently employed to solve this problem and have coined the biomedical community [2]. In contrast, phase images are largely unfamiliar and new visualizations and interpretations must be found to support the clinical decision-making process. Figure 1 shows typical examples of WBCs inside a phase microscope. In these grayscale images the brighter parts correspond to higher optical phase shifts $\Delta\phi$.



(a) Monocyte (b) Lymphocyte (c) Neutrophil (d) Eosinophil

Fig. 1. Quantitative phase images of WBC subtypes (Brightness ∼ Phase Shift)

2.2 Image Acquisition and Data Processing Pipeline

For our experiments, we use a custom-made differential holographic microscope by *Ovizio Imaging Systems* like [8,19]. It is equipped with a microfluidics chip to align the diluted blood sample stream in the focal plane of the microscope. Starting on the left of Fig. 2, the raw phase images, with a size of 512×384 pixels, undergo a simple background subtraction and threshold segmentation, before feeding the individual images of single cells into the next stages. Here, the path splits into several possibilities. The most transparent one is the extraction of handcrafted morphological features, which describe e.g. the `optical volume` of a cell or its `granularity` [15,19]. A subsequent interpretable classifier like a *naive Bayes* or *Random Forest* can use these features to predict the cell's class affiliation. On the other hand, we can pass the image to a deep (convolutional) neural network, which can learn important features to optimize cell classification without prior expert knowledge. Considering solely the classification accuracy, we noticed that the data driven black box models outperformed the classical approaches [11,14]. Therefore, we trained an *AlexNet* [9] architecture to perform the described WBC classification task on 7706 cells, which were balanced according to their class label. Due to their rare occurrence, we were not able to obtain a decent number of basophil cells near *in vivo* conditions. Therefore, this group of WBCs is not part of the data set. The fine tuned *AlexNet* classifier achieved an F_1-score of 0.963 and an accuracy of 96.7% on the unknown test set of 1000 cells with 250 cells per class.

Note that this work is not intended to improve the existing approaches regarding their accuracy. We investigate how the established methods and their

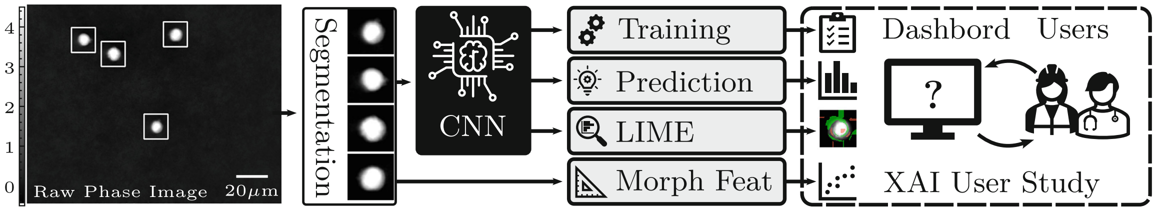


Fig. 2. The XAI dashboard needs to provide different means of explanations to optimally communicate the data processing pipeline to the domain experts.

outputs should be communicated and visualized for the different target groups to maximize the knowledge gain and acceptance of this emergent platform technology.

2.3 Evaluating Explainability and Interpretability

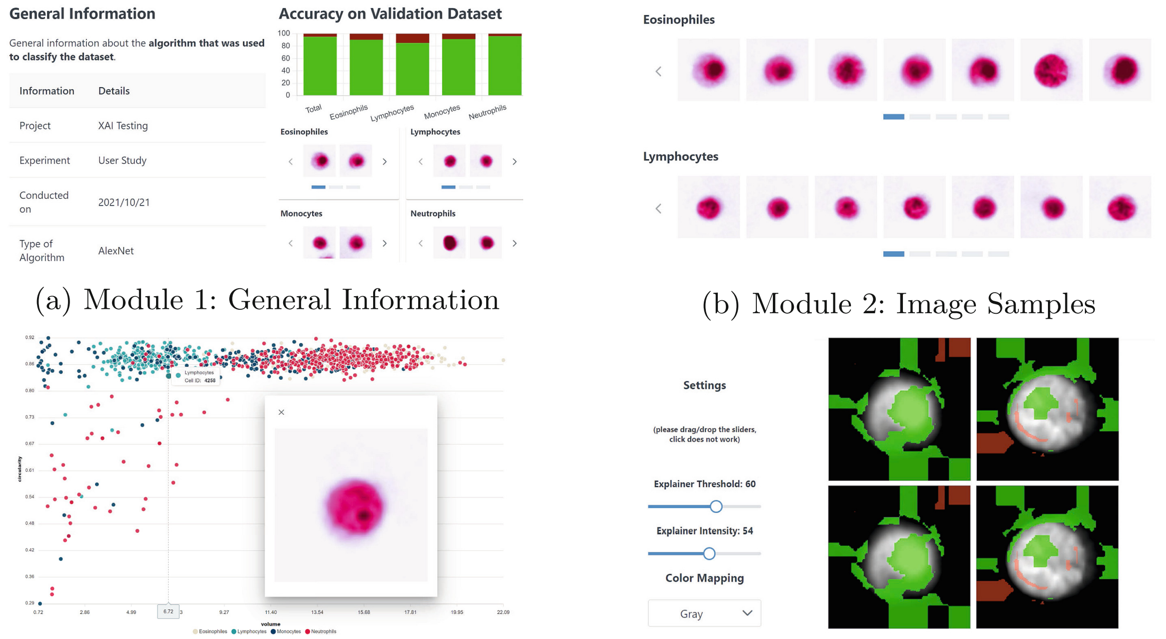
While it is widely agreed that trust and acceptance can be generated through transparency, explainability and robustness, there is still a broad discourse on how to define and how to determine these indicators [12]. Possible dimensions include (1) measuring the quality that is subjectively perceived by an individual or (2) measuring proxies for the sufficiency of the model. For the first dimension, we apply a *human-grounded metric*, where an evaluation should depend only on the "quality of the explanation, regardless of whether the explanation is the model itself or a post hoc interpretation of a Black Box model, and regardless of the correctness of the associated prediction" [3]. Here, we implemented an adapted form of the *binary forced choice*, where different visualizations are rated by humans according to a selection of questions. At the end, the users need to decide for their personal favorite. The second dimension focuses on the opacity of the chosen classifier and therefore evaluates its proxy model in the following aspects: First, the completeness compared to the original model, i.e., how closely it approximates the model to be explained. Second, the ability of the model to detect biases in the original model. And third, the ability of humans to "evaluate explanations for reasonableness, that is how well an explanation matches human expectations" [4].

3 XAI Dashboard Prototype

Based on preceded expert interviews and taking into account the technologies' boundary conditions, we identified four suitable interpretation approaches which we implemented as so-called modules in the prototype of an XAI dashboard.

3.1 Module 1: General Information on Training and Validation

The first XAI module displayed in Fig. 3a provides background information on the algorithm that was deployed for the prediction. On the left, a table lists



General Information

General information about the **algorithm that was used to classify the dataset**.

Information	Details
Project	XAI Testing
Experiment	User Study
Conducted on	2021/10/21
Type of Algorithm	AlexNet

Settings

(please drag/drop the sliders, click does not work)

Explainer Threshold: 60

Explainer Intensity: 54

Color Mapping

Gray

(a) Module 1: General Information

(b) Module 2: Image Samples

(c) Module 3: Morphological Features

(d) Module 4: LIME Visualizations

Fig. 3. Screenshots of the modules in the XAI Dashboard Prototype (Color figure online)

general information about the algorithm, where on the right, a barplot shows the performance of the algorithm on a validation data set. This offers the user an impression of the overall capability of the algorithm after the training has been completed. Finally, this module summarizes information about the training data set. The user has the option to view sample images for each class of WBCs specifically requested by interviewees with a biomedical background.

3.2 Module 2: Image Samples of Classified Cells

The second module shows cell samples from the actual prediction results. The cells are grouped by their predicted class, what can be seen in Fig. 3b. Since the underlying data are phase images (see Sect. 2.1), the images are not colorized by default. However, as a suitable color map can reveal more of the inner structure of the cell, the user has the option to chose a custom coloring. This element is based on a significant need of biomedical users to be able to have a look at cells. As it was identified in the preceding interviews, it would allow this target group to visually double-check if the classifications are meaningful.

3.3 Module 3: Morphological Features in a Scatter Plot

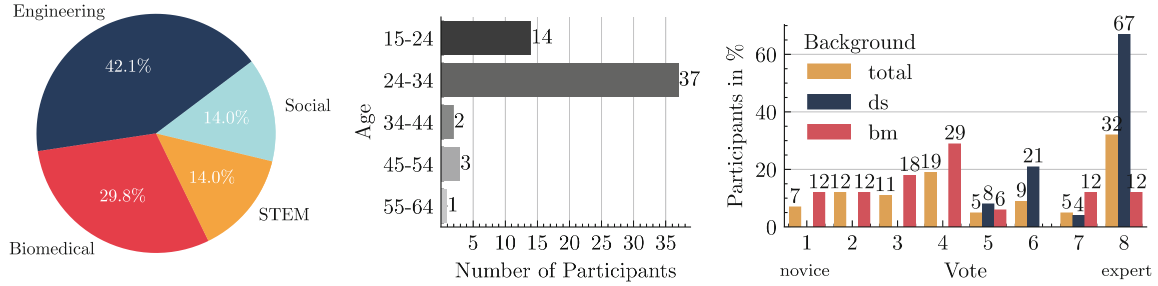
Many publications and interviews report the importance of morphological features for differentiating cells [8,14,15,19]. Furthermore, they are often used as input features for computer vision algorithms or for dimensionality reduction techniques. In contrast to the second module, which displays only individual predictions, the third module takes into account all predictions and gives a neat way to display and analyze the overall result. Figure 3c shows its implementation in form of a scatter plot of the individual cells. The axes represent selected morphological features, that can be dynamically adjusted by the user. The four predicted classes of WBCs are distinguished by the color of the dots. For closer inspection, the user can click on a dot to open a pop-up window containing the original image of the respective cell.

3.4 Module 4: Revealing Relevant Areas of an Image Using LIME

The fourth module, shown in Fig. 3d, reveals which parts of the image are relevant for the employed neural networks. For this purpose, the LIME library [17,21] has become one of the most popular tools. It provides insights into the behavior of a model by measuring the contribution of each input feature to the overall prediction of the sample. The visualization is created by perturbing the input data and observing the resulting effects on the model prediction. Furthermore, it is a model-agnostic linear proxy which identifies areas that contributed positively to the predicted class (green) and the ones which opposed that decision (red). Users interact with the method by being able to view only a minimum effect strength, tuning the overlays transparency to inspect the corresponding cell structures and finally focusing only on explanations concerning a specific class. This module might be the closest to human perception but also the most complex one in our prototype.

4 User Study Results

In total, the study cohort consists of 57 people from different scientific backgrounds and comprises students as well as researchers from various local institutes. Their distribution is displayed in Fig. 4a. The demographic composition of the cohort shows as shift towards the younger generation, as most of the participants are younger than 35. To simplify the observations we split the participants into a **biomedical (bm)** group and a **data science (ds)** group, in which people had to state a ML experience $\geq$ 5 on a scale from 1 to 8 and not being accounted



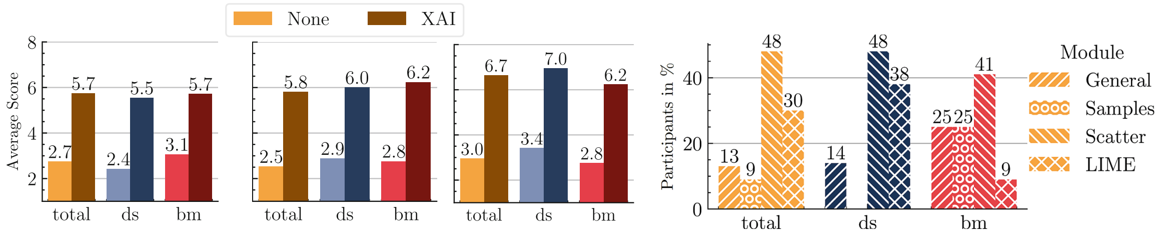
(a) Scientific Background (b) Age (c) Experience in AI

Fig. 4. Demographics of the participants

to the other group. This leads to Fig. 4c, which shows the distribution of participants' experience with machine learning algorithms. Naturally, data scientists have the most experience, with two-thirds of them reporting the highest score. They are therefore expected to provide the reference values for the study. In the biomedical group, the full range of experience is represented.

4.1 Evaluation of the Overall XAI Dashboard

For the estimation of the overall impact of the dashboard, a set of three questions was asked a first time when the users where confronted with the bare classification results and a second time when they had interacted with the XAI dashboard. The questions relate to the user's (a) understanding of the algorithm's behavior, (b) ability to detect biases, and (c) impression of the trustworthiness. We recorded their answers on an evenly distributed Lickert scale from 1 to 8. As the subplots in Fig. 5 show, the first take away is the dashboard's positive effect on all three aspects. This confirms the assumption that both user groups have difficulties to judge and understand the ML algorithm without the XAI spyhole. When looking at the **behavioral understanding**, the data scientist experience the highest improvement whereas the biomedical group states a slightly higher overall value even without the dashboard. When it comes to the **detection of biases** the user groups show an inverted influence. Another aspect to bear in mind is that the ability to judge the **trustworthiness** is rated higher than the understanding of the algorithm as well as the ability to detect biases.



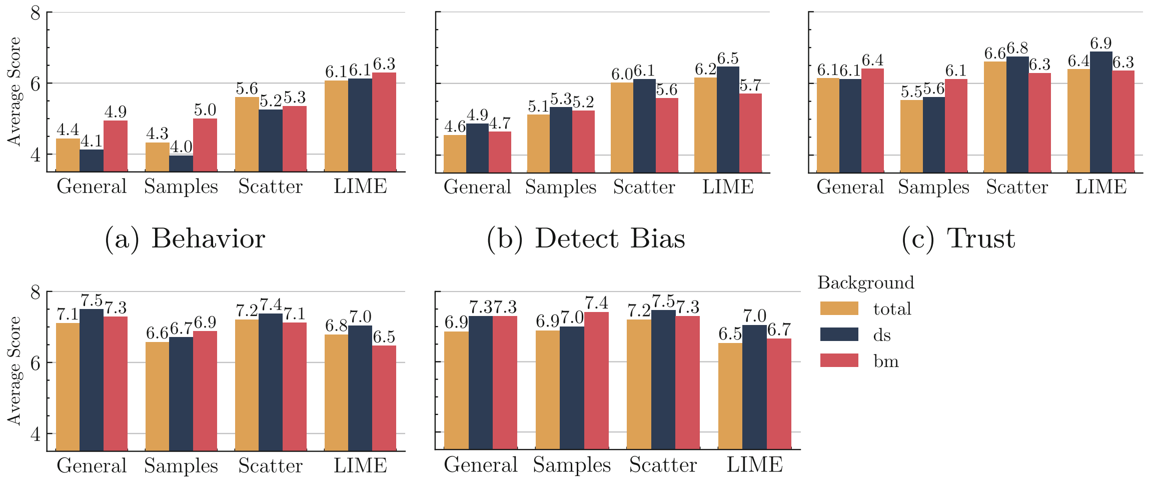
(a) Behavior (b) Detect Bias (c) Trust (d) Favorite XAI Module

Fig. 5. Overall improvements through XAI dashboards

At the end of the survey, the participants were asked to pick their **favorite** module of the dashboard. The results in Fig. 5d expose module 3 as the preferred one for both user groups. The highly complex LIME explanation ranks second best at the total group and among the data scientists. In contrast, it ranks last for the biomedical group. This is to be expected, since LIME is a method focusing on the needs of data scientist rather than biomedical people. Those are more interested in the representation of the familiar morphological features and the modules about the training/validation of the algorithm as well as the cell samples. Note that the data science group could not profit from cell images.

4.2 Comparison of the XAI Modules

In the survey, each module is examined by five aspects. In addition to the three questions introduced in the previous section, we asked (d) if the shown module is relevant for the task and (e) if the users understand the displayed information. These additional questions elaborate whether the previous are influenced by other factors and to ensure that the modules used are comprehensible in themselves. Concerning the perceived **relevancy**, Fig. 6d shows that all modules score relatively high. We partly attribute this to the fact that all modules were designed based on the needs of the various user groups. When comparing the modules, it is noticeable that the modules that refer globally to the algorithm (1, 3) are more relevant than the modules that refer to a local explanation (2, 4). Users are more interested in general information than investing time in individual examples. A good **understanding** of what is being shown is paramount to any data presentation and, in this case, a prerequisite for all other aspects. Equally to the relevancy, the modules are understood by the users pretty well as indicated by Fig. 6e. Data scientists generally exhibit a little higher understanding compared to biomedical, which is comprehensible considering this being a dashboard about ML. An exception is the module on showing classified cell samples. Surprisingly, across all modules biomedical users indicate a higher level of **understanding of the algorithm's behavior** than the data scientists. As it is unlikely that they could gather this lead by our dashboard, there must be an unobserved variable, which needs further investigation. Besides most of the models only achieve a mediocre rating, the LIME module stands out being the only one scoring above 6. We suspect the reason is that this is the only module that gives direct insight into the algorithm instead of indirectly examining the prediction results. Therefore, it is rather unusual that the users also consider LIME to be the best module for its capabilities to **detect biases**. The globally operating scatter plot is only second. Apparently, there is no tendency such as



(a) Behavior

(b) Detect Bias

(c) Trust

(d) Relevancy

(e) Understanding

Fig. 6. Overall improvements through XAI dashboards

global or local methods are superior to detect biases. Detecting biases scores a little higher than the aspect about understanding the behavior, but still, there is room for improvement. Users with a biomedical background seem to prefer the modules 1 and 2 when it comes to judging the **trustworthiness** of an algorithm. For the other modules, it is the opposite. Again, module 3 (Morphology, Scatter Plots) and module 4 (LIME) on average score highest for all user groups. On top, biomedical participants believe that the algorithm training and validation module helps them most in assessing trustworthiness.

5 Discussion and Conclusion

The aim of the dashboard is to provide users with a tool for interpretation, explanation and to judge an algorithm's trustworthiness. For this purpose, with respect to the diversity of the target groups, it proved beneficial to combine individual modules into a so-called XAI dashboard, as the overall dashboard was rated higher in all aspects than the single modules. When asking participants, if they would use the presented tools, the average participant is very positive with 7 out of 8 points.

In the qualitative interviews conducted in advance, respondents indicated that data scientists were more interested in more technical approaches such as the LIME module, while biomedical scientists would be more interested in morphological features and cell samples. Although this is confirmed to some extent, interestingly, the scatter plot of morphological features also emerged as the favorite module of data scientists. Thus, it can be concluded that while there are differences, there is certainly overlap in the relevance of the dashboard for both user groups simultaneously.

However, the users' assessment of understanding the algorithm's behavior and their ability to detect biases is only mediocre and the ultimate goal of making black box models transparent can only be approximated. Moreover, we observed that the domain knowledge might cause the users to be more skeptical versus technology they are familiar with. On the other hand, the explanations and interactivity might convey an exaggerated sense of security and lead users to overestimate their own trust, as can be seen in high ratings of trustworthiness despite the lack of understanding and detecting biases. Here, the responsibility lies with the developers. They must be careful not to mislead the users and stay as transparent as possible. Also legislation and academia require that the system's level of accuracy and its limitations are communicated [4,5].

All in all, there is a high demand for explainability and the ability to understand the decision making process is a crucial prerequisite for the deployment of ML. The shown model-agnostic, surrogate XAI modules, be they local or global, are considered suitable for this purpose by the different user groups. Nevertheless, there are still many other techniques (especially model-specific ones), which could provide even deeper insight. Tracing their impact in the proposed aspects will help to establish high potential technologies like quantitative phase imaging combined with ML in the biomedical domain. From our perspective XAI approaches will be indispensable for interdisciplinary research and clinical decision support systems.

Acknowledgment. The authors would like to especially honor the contributions H. Meier for the implementation of the user interface prototype and the execution of the studies. We also thank of J. Groll for the training and providing the underlying machine learning models.

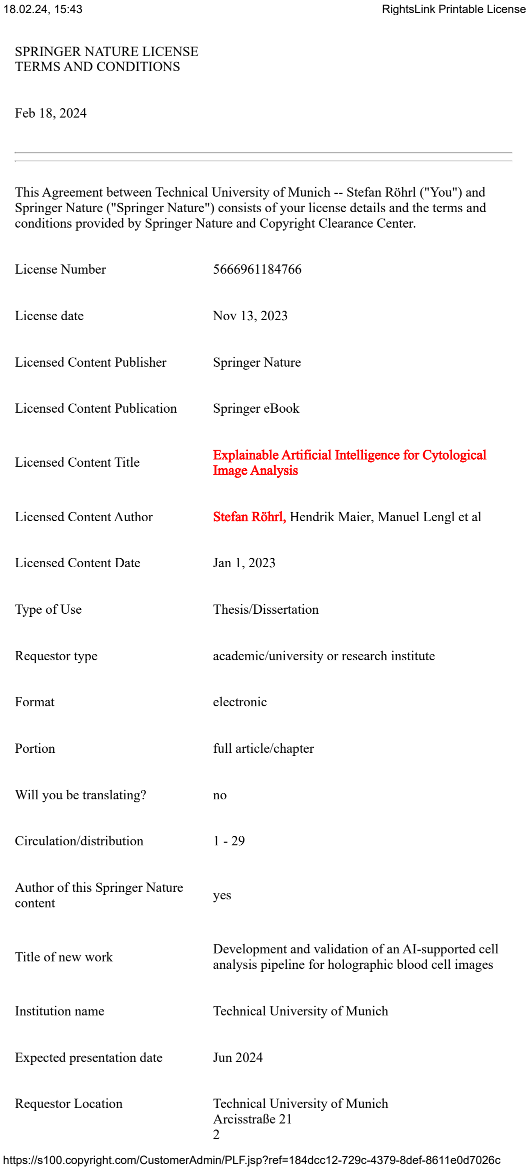
References

1. Alberts, B.: Molecular biology of the cell. WW Norton & Company (2017)
2. Barcia, J.J.: The giemsa stain: its history and applications. Int. J. Surg. Pathol. **15**(3), 292–296 (2007)
3. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv:1702.08608 (2017)
4. Gilpin, L.H. et al.: Explaining explanations: an overview of interpretability of machine learning. In: 5th International Conference on Data Science and Advanced Analytics, pp. 80–89 (2018)
5. High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI. European Commission (2019)
6. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9(4) (2019)
7. Horton, S., et al.: The Top 25 laboratory tests by volume and revenue in five different countries. Am. J. Clin. Pathol. **151**(5), 446–451 (2018)
8. Klenk, C., Heim, D., Ugele, M., Hayden, O.: Impact of sample preparation on holographic imaging of leukocytes. Opt. Eng. **59**(10), 102403 (2019)
9. Krizhevsky, A. et al.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105 (2012)
10. Lam, V.K., et al.: Machine Learning with Optical Phase Signatures for Phenotypic Profiling of Cell Lines. Cytometry A **95**(7), 757–768 (2019)
11. LeCun, Y., et al.: Deep learning. Nature **521**(7553), 436–444 (2015)
12. Murdoch, W.J., et al.: Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. **116**(44), 22071–22080 (2019)
13. Nguyen, T.H., et al.: Automatic Gleason grading of prostate cancer using quantitative phase imaging and machine learning. J. Biomed. Opt. **22**(3), 036015 (2017)
14. Nguyen, T.L., et al.: Quantitative Phase Imaging: Recent Advances and Expanding Potential in Biomedicine. Am. Chem. Soc. **16**(8), 11516–11544 (2022)
15. Paidi, S.K., et al.: Raman and quantitative phase imaging allow morpho-molecular recognition of malignancy and stages of B-cell acute lymphoblastic leukemia. Biosens. Bioelectron. **190**, 113403 (2021)
16. Park, Y., Depeursinge, C., Popescu, G.: Quantitative phase imaging in biomedicine. Nat. Photonics **12**(10), 578–589 (2018)
17. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining SIGKDD, pp. 1135–1144 (2016)

18. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Mach. Intell. **1**, 206–215 (2019)
19. Ugele, M. et al.: Label-Free High-Throughput Leukemia Detection by Holographic Microscopy. Advanced Science 5(12) (2018)
20. Vellido, A.: The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput. Appl. **32**(24), 18069–18083 (2020)
21. Zhang, Q.S., Zhu, S.C.: Visual interpretability for deep learning: a survey. Front. Inf. Technol. Electron. Eng. **19**(1), 27–39 (2018)

Copyright

The publisher granted permission to reprint the publication in this document on November 13th, 2023.

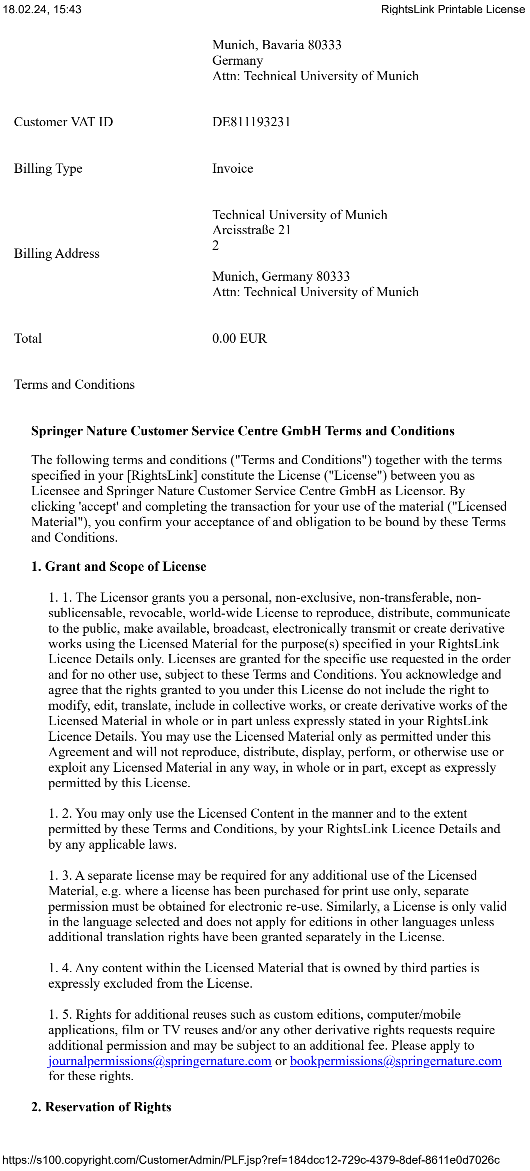


SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Feb 18, 2024

This Agreement between Technical University of Munich -- Stefan Röhrl ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5666961184766
License date	Nov 13, 2023
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Explainable Artificial Intelligence for Cytological Image Analysis
Licensed Content Author	Stefan Röhrl, Hendrik Maier, Manuel Lengl et al
Licensed Content Date	Jan 1, 2023
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes
Title of new work	Development and validation of an AI-supported cell analysis pipeline for holographic blood cell images
Institution name	Technical University of Munich
Expected presentation date	Jun 2024
Requestor Location	Technical University of Munich Arcisstraße 21 2



	Munich, Bavaria 80333 Germany Attn: Technical University of Munich
Customer VAT ID	DE811193231
Billing Type	Invoice
Billing Address	Technical University of Munich Arcisstraße 21 2 Munich, Germany 80333 Attn: Technical University of Munich
Total	0.00 EUR

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

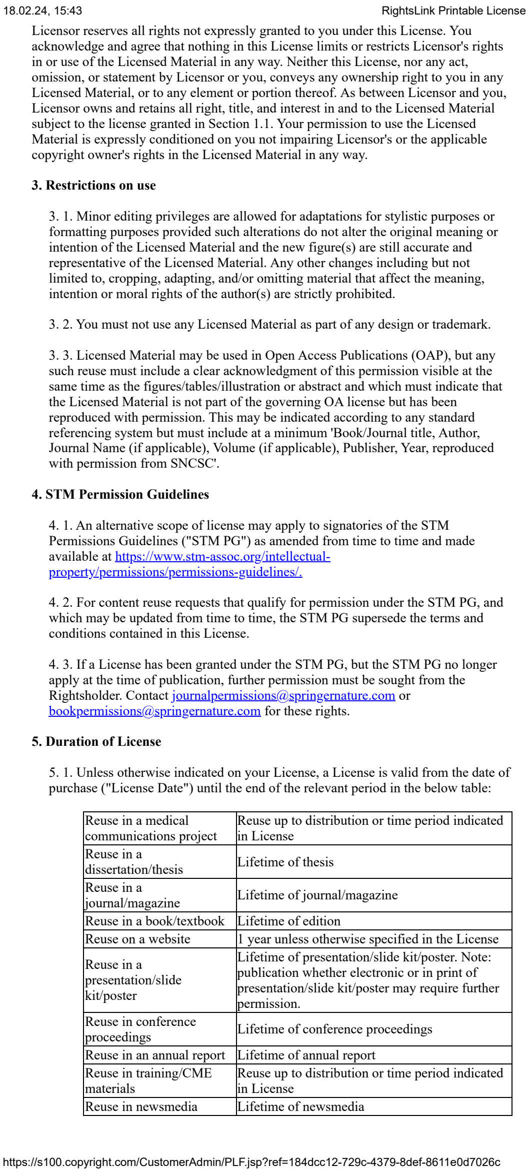
1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights



18.02.24, 15:43 RightsLink Printable License

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

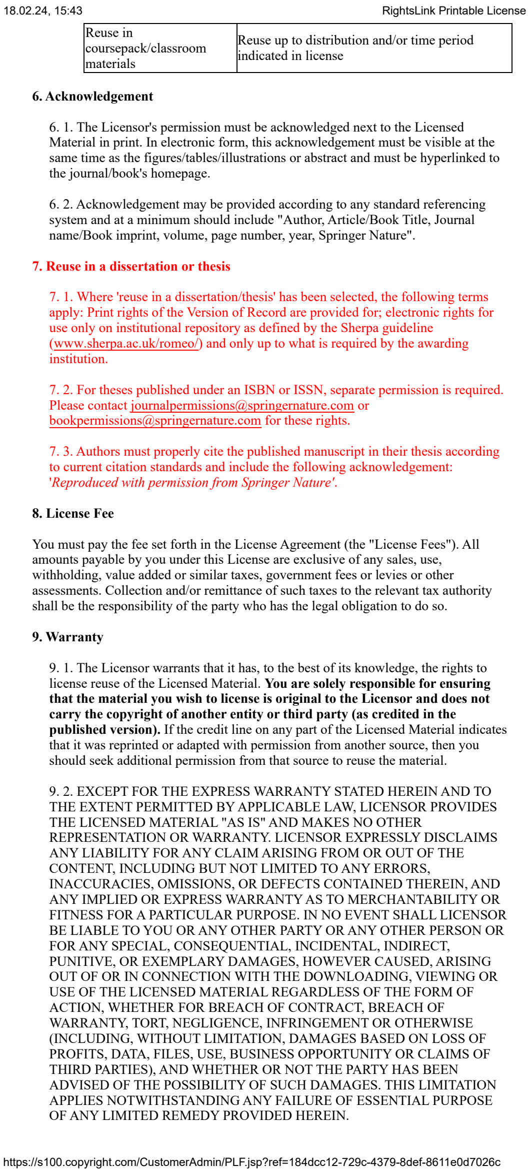
4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia

https://s100.copyright.com/CustomerAdmin/PLF.jsp?ref=184dcc12-729c-4379-8def-8611e0d7026c



18.02.24, 15:43 RightsLink Printable License

Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: *'Reproduced with permission from Springer Nature'.*

8. License Fee

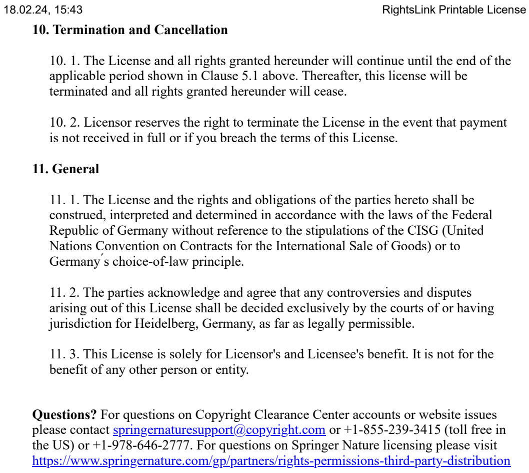
You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

https://s100.copyright.com/CustomerAdmin/PLF.jsp?ref=184dcc12-729c-4379-8def-8611e0d7026c



18.02.24, 15:43 RightsLink Printable License

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany´s choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Core Publication VI

Rethinking Usability Heuristics for Modern Biomedical Interfaces

Stefan Röhrl, Christian Janotte, Christian Klenk, Dominik Heim, Manuel Lengl, Alice Hein, Martin Knopp, Oliver Hayden, Klaus Diepold

Summary This paper contributes to enhancing the usability of AI-driven interfaces in biomedical research, focusing on blood cell segmentation and classification. It develops a tailored heuristic rule set specifically for biomedical AI interfaces and compares it to two established usability heuristics. An independent software prototype is created, incorporating an active learning approach to minimize human effort and improve usability. Extensive interviews and literature reviews inform the compilation of 15 heuristic rules critical to AI-driven biomedical user interfaces. Evaluation of the three usability rule sets demonstrates the effectiveness of the newly developed biomedical AI heuristics in identifying essential usability issues, particularly in domains requiring human-AI interaction. The paper emphasizes the importance of domain-specific heuristics in addressing the unique challenges of biomedical interfaces and aims to apply these design rules to advance AI-based technologies in research and healthcare.

Own Contributions

- Initiating the research idea and the prototype development
- Gathering of related work and assessing the scientific context of this work
- Planning the expert interviews and acquiring the study cohort
- Supervising the user study
- Writing the complete manuscript and designing the figures and plots
- Analyzing and interpreting the study results
- Defending the work and integrating the reviewer feedback

S. Röhrl, C. Janotte, C. Klenk, D. Heim, M. Lengl, A. Hein, M. Knopp, O. Hayden, and K. Diepold. “Rethinking Usability Heuristics for Modern Biomedical Interfaces”. In: *Proceedings of the 16th International Conference on Advances in Computer-Human Interactions.* IARIA, 2023, pp. 77–84 Online: https://personales.upv.es/thinkmind/dl/conferences/achi/achi_2023/achi_2023_4_30_20027.pdf

Rethinking Usability Heuristics for Modern Biomedical Interfaces

Stefan Röhrl*, Christian Janotte*, Christian Klenk†, Dominik Heim†, Manuel Lengl*, Alice Hein*,
Martin Knopp†*, Oliver Hayden† and Klaus Diepold*

*Chair of Data Processing, Technical University of Munich, Arcisstr. 21, 80333 Munich, Germany
†Heinz-Nixdorf Chair of Biomedical Electronics, Technical University of Munich, Einsteinstr. 25, 81675 Munich, Germany

{stefan.roehrl, christian.janotte, christian.klenk, dominik.heim, m.lengl, alice.hein, martin.knopp, oliver.hayden, kldi}@tum.de

***Abstract*—High usability is the ultimate goal in user interface development. In order to test this, user studies are often carried out at great expense. An alternative to this is offered by more favorable implementation guidelines and heuristic evaluation that get by with a smaller number of tests. Tools in the area of biomedical research face major challenges here, as they are extremely crucial, the users are highly demanding, and the advent of Artificial Intelligence (AI) requires researchers to take a powerful leap of faith. Since general heuristics are often insufficient for this domain, we introduce new Biomedical Research AI Heuristics and evaluate them among others using a prototype user interface in the domain of blood cell analysis. The comparative study shows our specialized approach competes very well with Nielsen's well-established general heuristics and a recent publication of rules for AI development. Our set finds the most relevant usability issues and can support the review process for the growing number of biomedical systems that will use artificial intelligence technologies in the future.**

***Keywords*—Usability Heuristics; Blood Cell Analysis; Human Assisted Labeling; Quantitative Phase Imaging.**

I. Introduction

One of the current challenges in biomedical research is to interpret the increasing amount of data available from new imaging and analysis techniques. To utilize the new information, more and more Artificial Intelligence (AI) is finding its way into this field. It is being used to facilitate differential diagnostics and to improve the understanding of medical conditions. Here, a new platform technology promises major changes in the field of blood analysis. A microscope working with Quantitative Phase Imaging (QPI) does not require expensive reagents and therefore no time-consuming sample preparation [1][2]. Combining this approach with a microfluidics channel, the optical amplitude and phase information of millions of cells can be recorded within minutes. The simplicity, high statistical power and speed of this approach allow statements about the composition of the blood, morphological changes of the cells and thus the kinetics of diseases [3]–[5]. Nevertheless, the resulting images are rather unknown in the medical domain and reference databases as well as sufficient ground truth data is missing, which hinders the efficient training of machine learning algorithms. To overcome these problems, we have to provide an easy way for researchers to work hand in hand with the machine to explore this new field of hematological analysis based on computer vision and AI.

For successful human-computer interaction, the user interface represents the common language the interdisciplinary researchers and developers have to speak. Misunderstandings can prevent such emerging technologies from being successful, as they cannot rely on the trust and the establishment of the gold standard methods [6]. Here, we would like to introduce and compare new **rule set for heuristic evaluation**, which are specifically designed for the development of AI-infused interfaces in biomedical research. As the target group of biomedical researchers and practitioners stands out for a busy schedule and demand high standards in the aspects of explainability [7], transparency [8] and causality [9], having a set of tailor-made heuristics promises a quicker translation of new technologies to the point of care. While most of the usability heuristics used in the past have been of a rather general nature [10], domain-specific ones have become more prominent in the last decades [11]–[13].

In this work, we propose a new labeling platform for holographic cell images where humans and AI work closely together in (inter-)active learning scenarios. This will facilitate the generation of verified ground truth data and be a valuable representative for this kind of biomedical user interfaces. Our primary interest, however, is to validate the newly developed usability heuristics against the existing ones, and thus to meet the need for guidance in the development process of AI-infused biomedical systems.

In the following, the work is divided into the appropriate sections: Section II motivates the choice of the clinical application and introduces the concepts for comparing heuristic rule frameworks. Then, Section III presents the specially developed web-based prototype of a user interface. The three sets of heuristic rules are introduced in Section IV, followed by their evaluation by experts as well as by user tests in Section V. The results of the study are described and visualized in Section VI. Finally, Section VII discusses the findings and draws conclusions for possible future work.

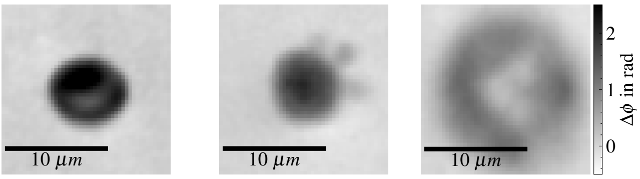
II. Background and Related Work

Before introducing the prototype, we will investigate the medical relevance of the chosen use case and the methods for evaluating sets of heuristic rules.

A. Medical Relevance of Quantitative Phase Imaging

The process of blood analysis in general is one of the most requested laboratory tests [14] and has been extensively studied in the past, leading to technically advanced solutions.

Copyright (c) IARIA, 2023. ISBN: 978-1-68558-078-0



(a) Red Blood Cell (b) Aggregate (c) Defocused Cell

Figure 1. Phase images of different cell classes

As a result, most state-of-the-art instruments work with a blood processing scheme based on marker materials [6]. Although these devices being very precise, they come with several downsides, as they require non-specific and costly labeling as well as time-consuming sample preparation such as *hemolysis* [15]. Using QPI methods combined with machine learning, the exercise translates into a computer vision task, which offers more flexibility. The morphological and internal patterns of blood cells provide insights for oncological [3][16], parasitic infections [17] and other diseases [4]. Also, the aggregation of blood cells can deliver crucial information [5][18].

However, before the images can be automatically interpreted and classified, they must be segmented and labeled by experts. Figure 1 shows representatives of typical cells and structures as they look like under a quantitative phase microscope. Red blood cells (a) are quite simple to detect, whereas aggregates of white blood cells and platelets (b) are more difficult to find due to their complex structure and the associated rarer occurrence. The algorithm as well as the human also have to learn, which objects need to be discarded (c). Note that medical experts are usually only trained on stained thin films and are therefore unfamiliar to this representation [6]. The brightness information directly correlates with the optical phase shift $\Delta\phi$ caused by the cells. Greater detail about the microscope can be found in [2][3].

B. Active Learning for Human Assisted Labeling

Manually labeling large amounts of data such as images is tedious and sometimes even challenging for skilled personnel, as the previous section describes. Therefore, crowd sourcing is not an option. As biomedical experts are expensive and limited in time, the Active Learning (AL) approach seems promising [19]. In AL, an algorithm is trained on a very sparse data set to learn a classification problem. However, instead of leaving the user with the task of correcting a predicted class label when the system is uncertain, the algorithm attempts to minimize the actions that need to be taken [20]. Moreover, AL shows suitable behavior for imbalanced data sets like ours to build a *human-in-the-loop* system [21], as we do in our prototype.

C. Quality Assessment of Usability Heuristics

The developed user interface represents the precedent to put our newly developed heuristics into practice. To make the heuristics more comparable, we need to introduce quality assessment measures as well as standard procedures to obtain these measures. Hartson et al. [22] propose to apply the different evaluation methods to the target system and compare the found usability problems to a baseline of “real” usability problems. In our work, we will determine the baseline by conducting *asymptotic user testing* [22]. As not every usability problem is as crucial as the other, we will further rate each problem then by a *severity score* proposed by Nielsen [23]. Table I shows the weighting of the apparent usability problems in order to compare the heuristics on their ability to prevent major usability issues.

TABLE I. SEVERITY RATINGS FOR USABILITY PROBLEMS [23]

	$s(p)$	**Description**
Rating	0	Violates a heuristic but is not a usability problem
	1	Cosmetic or unimportant usability problem
	2	Minor usability problem
	3	Significant usability problem
	4	Usability catastrophe

Starting from there, Sears [24] defines the **thoroughness** criterion (also known as recall in other disciplines)

$$T = \frac{|E \cap F|}{|E|}, \tag{1}$$

where $|E \cap F|$ denotes the number of problems F found by the heuristics from the baseline set of real usability problems E. Using our mapping of severity scores we can calculate the **weighted thoroughness**

$$T_w = \frac{\sum_i s(f_i)}{\sum_j s(e_j)} \text{ with } f_i \in E \cap F \text{ and } e_j \in E, \tag{2}$$

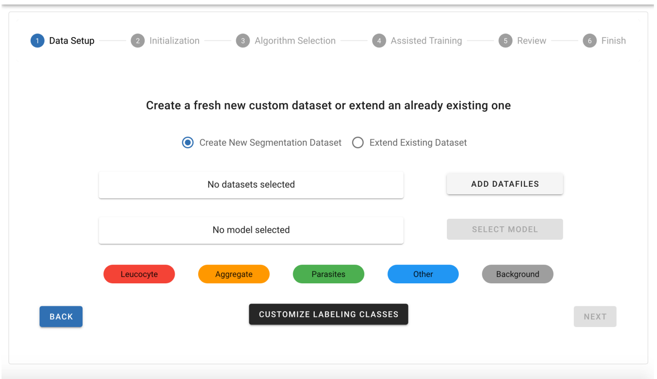
where $s(p)$ assigns every usability problem its rating according to Table I. Finally, the **validity** criterion [25] (also called precision)

$$V = \frac{|E \cap F|}{|F|} \tag{3}$$

helps us to judge how many of the identified problems F where real and no false alarms.

III. HUMAN ASSISTED LABELING PROTOTYPE

In order to provide an easily accessible and customizable user interface, we developed a web-based prototype for this study, which is divided into different views.



Create a fresh new custom dataset or extend an already existing one

● Create New Segmentation Dataset ○ Extend Existing Dataset

No datasets selected

No model selected

Figure 2. View 1 - Data Setup

The **data setup** view displayed in Figure 2 contains the functionality for setting up the general properties of the project. At the top, there is an option button that allows the user to choose whether to start with an empty data set or expand an existing data set based on a previously trained algorithm. Below, the user finds means to load the respective data containers or models. The lower part of the page displays the currently available classes of cell types. Each of them has its own color scheme and can be customized, added or deleted by clicking the button below them.

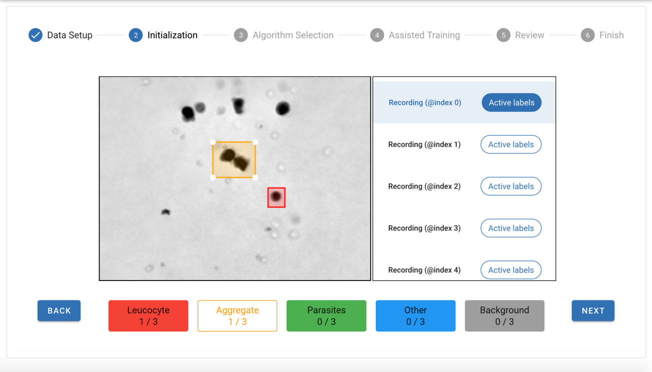
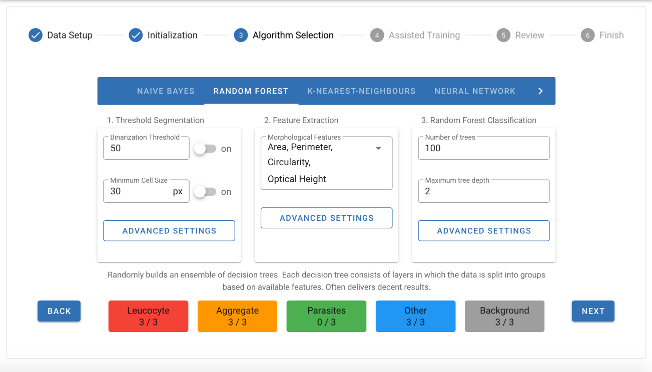


Figure 3. View 2 - Initialization

After having specified the labeling task and the expected classes of cells, clicking the "Next" button opens the **initialization** view (Figure 3). As the name suggests, it is used to provide an initial training set for the later algorithm. A large canvas is the main component of this view, displaying the selected set of cells, but also providing an area for drawing and annotating. In the bottom part of the view, there is a footer that displays the available classes. Clicking on one of them activates the class which is illustrated by highlighting. The user can now click and drag the mouse to draw bounding rectangles around the cells in the image. This combination of location and class is later called a label.



1. Threshold Segmentation

Binarization Threshold: 50 — on

Minimum Cell Size: 30 px — on

2. Feature Extraction

Morphological Features: Area, Perimeter, Circularity, Optical Height

3. Random Forest Classification

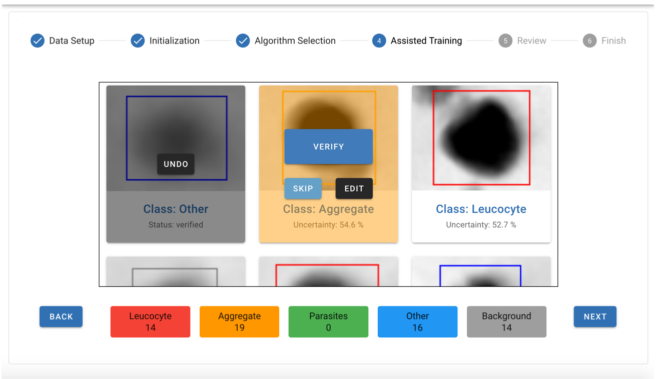
Number of trees: 100

Maximum tree depth: 2

Randomly builds an ensemble of decision trees. Each decision tree consists of layers in which the data is split into groups based on available features. Often delivers decent results.

Figure 4. View 3 - Algorithm Selection

In the **algorithm selection** view (Figure 4), users can specify the type of algorithm they want to use to classify cells in the records by selecting the appropriate tab at the top. Currently, users can choose from *Naive Bayes*, *Random Forest*, *k-Nearest-Neighbors* and a small *Neural Network* [26][27]. Depending on the type of classifier, necessary segmentation and feature extraction steps can be customized in the respective tab.



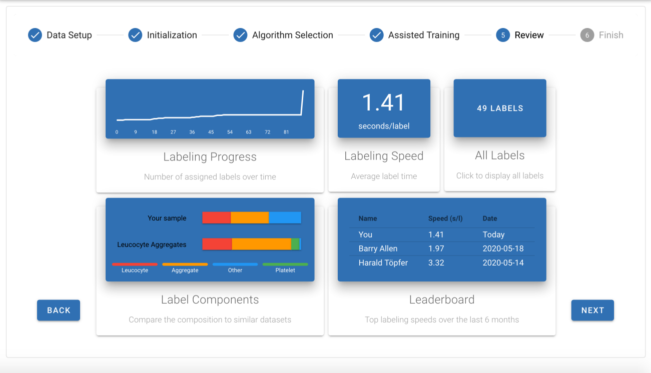
Class: Other
Status: verified

Class: Aggregate
Uncertainty: 54.6 %

Class: Leucocyte
Uncertainty: 52.7 %

Figure 5. View 4 - Assisted Training

When the algorithm has made its first predictions based on the initial training set the **assisted training** part starts in this reoccurring view (Figure 5). A gallery appears showing the proposed labels that the algorithm found in the data. As suggested by the AL principles from Section II-B, they are ordered by their uncertainty from the highest to the lowest value. Here, users can intervene and verify or correct the algorithm and hence, enlarge the training set without manually scanning the raw data and drawing rectangles. Furthermore, human assistance is only required for difficult objects, reducing the wasted time on already mastered samples. The algorithm can then be periodically retrained on the extended training set and can quickly reach a satisfying performance on the complete data set.



Name	Speed (s/l)	Date
You	1.41	Today
Barry Allen	1.97	2020-05-18
Harald Töpfer	3.32	2020-05-14

Figure 6. View 5 - Review

Finally, the **review** view (Figure 6) summarizes the labeling progress over time and the current performance. It compares the composition of the data set to other similar data sets and displays the percentages of detected cell classes. As a kind of gamification element, it also shows the labeling speed and ranks it with the performance of other users.

IV. Usability Heuristics

This section gives an overview of the state of the art in usability heuristics and introduces our new set of rules.

A. Nielsen's Heuristics

The general usability heuristics by Nielsen and Molich have been known for decades and are still used today. They are based on some of the most fundamental rules for user interface development. Their strength in finding many usability problems has been demonstrated in the past. Due to their generality, they can be applied to almost any type of system, but they have the disadvantage of not always finding as many usability problems as a set developed specifically for the system's domain. Nevertheless, they are a good starting point and will be a strong competitor and thus valuable for comparing them with our own set of heuristics. For our comparison, we used the rules from Nielsen's most cited publications [10], [28]. Also, minor modifications in wording [29], done in the last years, were considered.

B. Human-AI Interaction Heuristics

With the advent of AI in recent decades, it was only a matter of time before user interface developers began to address the specific requirements of AI-infused interfaces. In collaboration with Microsoft, a group of researchers led by Amershi recently proposed a set of usability guidelines for the development of such systems [30]. Guidelines and heuristics are not technically the same thing, since guidelines are used during the implementation of an interface and heuristics during verification. However, for short lists of guidelines such as this one, they can often be used interchangeably [31]. For our experiments, we converted the guidelines into a set of heuristic rules and provided them with examples for the experts so that this set can be used equivalently for the evaluation. This set of heuristics additionally distinguishes whether a problem occurs **immediately**, while the user is using the tool, or if it appears over a **longer period**. It is to be expected that some of the listed rules of this set will have minor relevance for our labeling interface like Rule 6 that is about mitigating social biases. To be consistent, we will keep the set unaltered.

C. Biomedical Research AI Heuristics

The main idea of this work is not only to compare the proven heuristics by Nielsen and Molich and the recently published AI guidelines interpreted as heuristics. We intend to create our own set of heuristics specifically targeted at biomedical research applications that use AI. The amount of software in this area will increase in the coming years, and it may be beneficial to have custom heuristics at hand for evaluation to save valuable testing time. Table II shows a set of 15 rules grouped in four categories, which constitute our *Biomedical Research AI Heuristics*. They are inspired by several publications in the domain of user interface design, biomedical and AI applications over the last decades. We completed those rules by hints and suggestions from preceding interviews with experts from local institutions working in the field of biomedical research.

Table II. Heuristics for AI in biomedical research

	#	Name	Short Description
Structure	1	Streamline main task	Focus on the main task that a system was created for and make the system easy to learn [32].
	2	Provide full control	Provide global control of important model parameters and the data pipeline [33][34].
	3	Orientation	Always show users where they are, what is currently going on and what they can do next [10].
Interaction	4	Guide attention	Keep the users focused on their task and only alarm them in urgent cases [35][36].
	5	Provide comparisons	Let users compare among similar data or parameters when they need to judge an outcome or make a decision.
	6	Show impact	Users need to see how their actions influence the system and its performance [37].
	7	User over System	Allow users to correct errors of the AI efficiently at all times and even turn off the AI if needed [35].
Presentation	8	Familiar language	Use non-technical language if possible. Pay attention to use correct terminology for medical concepts [38].
	9	Precise language	Avoid ambiguous wording for labels and commands that could trigger confusion [10].
	10	Familiar look	Use ways of presentation for the interface that users know from other tools.
	11	Appeal	Give the users the feeling of using a state-of-the-art and high-quality product.
Explainability	12	Explain data	Foster the interpretability of the data and how it differs from other data sources [39].
	13	Explain processing	There needs to be a high-level explanation for the overall procedure that is performed by the system [9].
	14	Explain reasoning	There has to be an explanation why and how the system derived a certain result or prediction [9].
	15	Strengths / Limitations	Show what the strengths and weaknesses of the system are and what expectations are realistic. [40]

V. Usability Evaluation

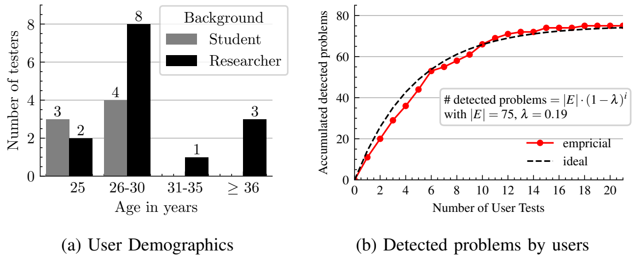
Once all the prerequisites are met, the prototype is tested by means of heuristic evaluation and user testing.

A. Heuristic Evaluation

For the evaluation of heuristics, we will compare the three heuristics with different aims and origins presented in the previous section. Their performance will be compared to determine whether general or domain-specific heuristics perform better in the domain of AI-infused interfaces for biomedical research. Most usability researchers like Nielsen classify potential expert evaluators into three different categories: *novices*, *single experts* and *double experts* [41]. Novices are new to usability concepts but often have knowledge in the domain where the user interface will be deployed. In contrast, single experts already have experience in the field of usability engineering but lack knowledge of the designated domain. Double experts are evaluators who are proficient both in usability engineering and the domain. On average, a novice finds only 22% of issues in a system, while single experts manage to find 41% and double experts even around 60% [42]. The experts participating in our review are neither novice evaluators nor have they been conducting such reviews for years. Nevertheless, they have a sound knowledge of usability concepts and have conducted a heuristic evaluation before. In addition, some of them also have a basic understanding of the domain of the system. Each heuristic is applied to our prototype user interface by five different evaluators, a number often recommended for user interface development because of its cost-efficiency [43]. In order to keep focus on the most relevant usability problems, we use the severity rating system introduced in Table I. During the expert review process, each expert will assign a level of potential impact to the usability problems they have discovered. After a final list of aggregated usability problems is compiled for all heuristics, each expert will also assign ratings to the problems found by their peers. In the end, the ratings among the experts will be averaged and rounded.

B. *User Testing*

In order to compare the different heuristics in this work, we need to gather knowledge about the real usability problems E inherent in our prototype interface. For this, asymptotic user testing [22] is selected as a test procedure. With a conservative detection rate of 19% per user [22][44], the relation between the number of testers and the percentage of discovered usability problems seems to level off at around 20 testers, which is very late. This is shown by the ideal curve in Figure 7b. However, to increase the chances of overlooking as few problems as possible, we decided to conduct a test series with at least this number of testers. Eventually, we found 21 representative users with a biomedical background who were willing to participate. Their demographics are displayed in Figure 7a. The youngest tester was 21 and the oldest 59 years old. What almost all testers had in common was their lack of experience with machine learning. 76.2% said they had no experience at all. This was beneficial to see how they would react to something they had never used before.

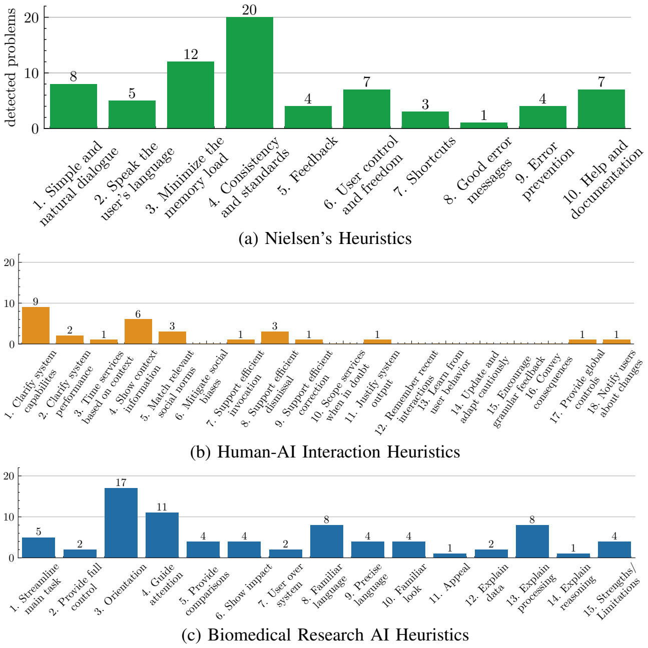


(a) User Demographics

(b) Detected problems by users

Figure 7. Composition and performance of the user tests

All users were given two tasks: 1) “In a small sample you are interested in the number of white blood cells, single platelets, and cell aggregates. Extract these components and perform some further evaluation to show them to a co-worker.” 2) “Your bigger recording is rich in white blood cell aggregates. You want to detect the same components as before but also keep track of other cells as they might become relevant later. Prepare and store your results for further evaluation.” Users were given as much time as they needed to complete the tasks and were encouraged to ask questions and think aloud throughout the test [45]. Meanwhile, the evaluator took informal notes that would later be summarized in a formal test protocol. Testers were also required to complete a short questionnaire after the test.



(a) Nielsen's Heuristics

(b) Human-AI Interaction Heuristics

(c) Biomedical Research AI Heuristics

Figure 8. Detected problems by the respective heuristics

VI. Results

This section summarizes findings and compares the results of the different testing strategies by the proposed metrics.

A. *Heuristic Evaluation*

The first set of rules we applied to our biomedical user interface consisted of **Nielsen's ten general usability heuristics**. They were developed without regard to a specific type of user interface. The review conducted by five evaluators using this rule set identified 60 violations within the system. This number was obtained by comparing and aggregating the results of the individual evaluators. Figure 8a shows the number of usability problems identified by each rule of Nielsen's heuristics. It is important to note that the sum of all bars is greater than the total number of problems identified, since a problem may relate to more than one heuristic. Prominently, Rule 4, which deals with user interface consistency and standards, is responsible for 20 usability problems, which is significantly more than any other rule. The second most problems are related to Rule 3, which focuses on reducing the user's memory load. It is not possible to say whether their numerous occurrence is due to the fact that these rules highlight important aspects of biomedical interfaces very effectively, or whether an unusual high number of violations occurred by chance. The rough list of usability problems merely indicates the presence of these violations. Conclusively, the five evaluators emphasized that they enjoyed working with the set and that it was easy to use. In addition, it is worth noting that each heuristic was applied at least once and no heuristic was omitted.

The second set of rules used in this project were the recently published **guidelines for human-AI interaction**. They were proposed as guidelines that can support the development of interfaces to let human users interact with AI. In our evaluation, the five experts discovered 26 violations and the corresponding usability problems, which is less than half the amount that Nielsen's heuristics helped to find. Figure 8b shows the number of heuristic violations per rule in this set of human-AI heuristics. The distribution of problems looks quite different from that resulting from Nielsen's heuristics. First, there are a number of rules that did not help uncover a usability problem at all. This is mainly due to the fact that these aim for long-term effects which do not apply to the tasks covered in our study. The two heuristics that have received the most attention are Rule 1, which deals with explaining what the system can do, and Rule 4, concerning context and relevancy of the displayed information. What is interesting about this second set of heuristics is the informal feedback from the evaluators. They pointed out that these rules were very difficult to apply to the system. The reason for this could be that they were not developed as heuristics, but as guidelines. As such, they might be too specific and not generally applicable.

The third set of rules we applied to the interface is the one we created specifically for the field of **biomedical research interfaces that use AI**. Here, the five experts reported a list of 55 usability problems. This is slightly less than what they discovered with the general heuristics, but still much more than what the heuristics for human-AI interaction identified. The distribution of usability issues across the different rules within our custom heuristics is shown in Figure 8c. All fifteen rules were found to have at least one violation. The two most frequent heuristics are Rule 3 and Rule 4, which are concerned with providing orientation and guiding the user's attention. The third place is shared by Rule 8 and Rule 13. It is interesting to note that these four heuristics are all aimed at reducing the complexity of AI for the biomedical users or enabling them to better deal with it. Evaluators noted that the set was easy to use and that they felt it covered most usability issues with a large impact on the user experience. This feeling is supported by the fact that it detected the most usability issues with the highest impact among the three heuristic sets, with fourteen violations of the maximum severity level.

B. User Testing

This would lead us to the quality assessment metrics introduced in Section II-C, but before we can apply them we need the baseline of real usability problems E determined by our user tests. With respect to the asymptotic behavior of the usability problem discovery process, we assumed that about 20 testers would be needed to find most of the problems. The test ultimately resulted in the detection of 75 usability problems over the course of 21 user tests. To support the claim that we almost reached an asymptotic upper bound, we plotted the occurrence of problems over tests in Figure 7b, indeed revealing the asymptotic shape of a Poisson process [43].

To obtain the severity ratings of the real usability problems, we sent the complete list of issues to our usability experts and summarized the ratings based on their judgment. Many of the entries in this list are common problems that can occur in any type of user interface, such as misleading button descriptions and lack of loading indicators. However, there are also some problems (see Table III) that seem to be rather unique and that can serve as examples of typical problems in environments where users with a biomedical background interact with AI. These were concentrated to uncertainties about the specific workflow of the program and obscure consequences, which certain changes in the settings might have. Only 2 out of 21 users requested major changes before they would use such a system for their daily work. 19% stated that they would use it, but still suggested some minor changes. The majority of 71% of users indicated that they would use the system in the future exactly as it is, after becoming familiar with it.

TABLE III. EXEMPLARY USABILITY PROBLEMS

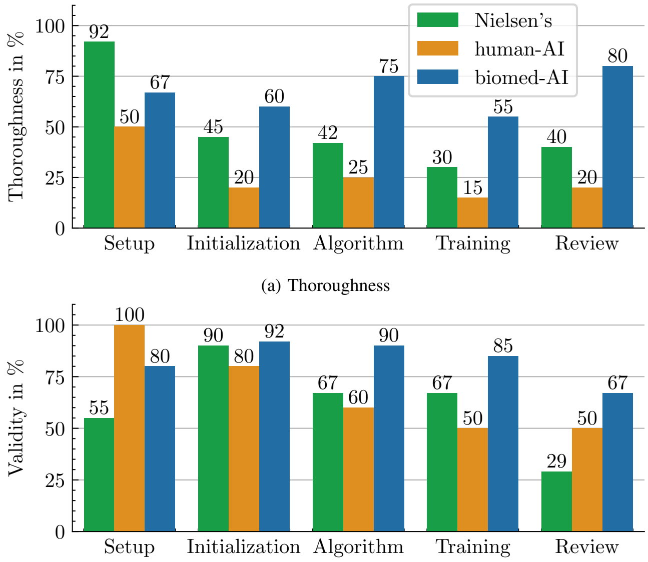
View	**Description** **Note:** The listed problems all have a maximum severity rating of 4. The numbers indicate the violated heuristic rule or the number of affected users respectively.	**Nielsen's**	**human-AI**	**biomed-AI**	**User Test**
1	There is no clear indicator that tells the user when the initialization is completed or what happens with empty classes. The "X/3" in the footer is not prominent enough.	5 10		4 6	2
3	The different algorithms are not sufficiently explained and the current explanations are hard to find. Users do not know which algorithm to choose.	1 2		4 6 13 15	1
3	The wording of some parameters and explanations is too technical to understand.			8	2
4	Users do not understand the training process, what they have to do and why multiple iterations with retraining make sense. The initial performance might be disappointing.	3 10	1 2	13	5

C. Metric-Based Comparison of Heuristics

Now that we have a baseline, we can relate it to the findings from different heuristics. This results in a list of 104 usability problems, with which we can compute the quality assessment metrics. As listed in Table IV, the three different sets of heuristics did not perform equally well. For almost all metrics, the domain focus of our set of heuristics is noticeable and provides improved results in the criteria **thoroughness** and **validity**. The general heuristics by Nielsen still occupy a stable second place, although it should be noted that all three heuristics were not able to predict usability problems seamlessly. Nevertheless, the high validity of our custom heuristics make them a reliable tool to alert developers of incipient and severe usability issues. We can further compute the thoroughness metric for high severity levels (3 & 4), as these should be addressed early in the development process. Among the highest level of severity (4), our biomedical heuristics account for a thoroughness of

TABLE IV. RESULTS OF THE QUALITY ASSESSMENT METRICS

Metric	**Nielsen's**	**human-AI**	**biomed-AI**
Thoroughness	50.1%	25.3%	**62.7%**
Weighted Thoroughness	54.0%	28.9%	**69.0%**
Average Severity	2.66	**2.84**	2.74
Validity	63.3%	73.1%	**85.5%**



(a) Thoroughness

(b) Validity

Figure 9. Quality assessment metrics for the individual views

93.3%. Nielsen's set improves to 73.3% and the human-AI interaction heuristics to 46.7%.

Since not all views of our prototype are equally influenced by AI, we investigate the performance in the individual components of the system. Accordingly, Figure 9a shows the **thoroughness** metric for each set of heuristics. The Nielsen heuristics have the highest thoroughness in the setup view, but as more interaction with the AI is emerging, their performance drops. Surprisingly, the AI emphasized human-AI heuristics score even worse. Our biomedical heuristics have the highest thoroughness in the initialization, algorithm, training and review views, as these require frequent interaction between the users and the machine learning algorithms.

Similarly, we can evaluate the **validity** of the heuristics depending on the view as displayed in Figure 9b. The validity of our biomedical heuristics is always higher as Nielsen's heuristics. This means that Nielsen's heuristics tend to find a lot of irrelevant usability issues in all views. However, the validity for the review view is particularly low, for all three sets.

VII. Discussion and Conclusion

The performance metrics from Section VI-C indicate that there is a noticeable difference in which heuristics we use for an expert evaluation of an AI-infused user interface within the biomedical domain.

Nielsen's well-known heuristics struggle when it comes to finding real usability problems in biomedical interfaces induced with AI. They showed only mediocre thoroughness in these parts of the prototype. However, they found the most genuine usability problems in the parts of the interface that were least affected by machine learning, resulting in a high performance in those views. Unfortunately, this seems to be accompanied by reduced validity. Nielsen's heuristics tend to find more expendable problems than the competing heuristics. All in all, the results suggest that these general heuristics are not always the best choice when it comes to finding usability problems in a specific domain like the one we studied. This is a result that also has been discussed in other publications [46].

The **heuristics for human-AI** interaction did not score particularly well in terms of thoroughness and validity. In addition, the experts in this study indicated that this set was most difficult to use for interface evaluation. This could be due to the fact that this set was originally designed as a guideline and also has large focus on long-term effects that are not relevant here.

The **heuristics for biomedical user-AI** interaction that we developed in this work provided the most compelling results. While their thoroughness was good but not great, their weighted thoroughness and thus their potential to uncover the most important problems in a user interface like our prototype was a positive discovery. This was further emphasized by the set's high thoroughness scores for high severity problems. Moreover, our set performed better than Nielsen's general heuristics, especially in the parts of the interface that focused on user-AI interaction.

When putting the heuristics' evaluation in a larger context, we expected that we could detect at least 70% of the real usability problems as foreseen in literature [11][22][43]. Our experts were not novices, but the best detection rate they could achieve was 62.7% with the biomedical heuristics and even less with the other sets. There is a possibility that this is due to inadequate evaluation of our experts. However, it is more likely that the main reason is that it is simply more difficult to find usability problems in the domain we analyzed. This assertion is supported by studies like [11], pointing out the need for domain-specific heuristics for domains where usability problems are immanently difficult to detect. This was also one of the basic assumptions on which this entire paper is based. As biomedical interfaces seem challenging, an unweighted thoroughness of 62.7% is a satisfactory result.

Finally, we aim to apply our new biomedical heuristics on more user interfaces in this domain. Tools that are used for making diagnoses and more complex reasoning could be of special interest. With a more diverse expert group, we hope to reduce the effort of conducting user tests and help to establish AI based technologies in biomedical research and healthcare.

Acknowledgment

The authors would like to especially honor the contributions C. Janotte for the implementation of the user interface prototype and the execution of the studies. Furthermore, thanks go to D. Heim and C. Klenk for the sample preparation and recording of the measurements. All expert reviewers should also be thanked at this point.

This research was funded by the German Federal Ministry for Education and Research (BMBF) with the funding ID ZN 01 | S17049.

References

[1] Y. Park, C. Depeursinge, and G. Popescu, "Quantitative phase imaging in biomedicine," *Nature Photonics*, vol. 12, no. 10, pp. 578–589, 2018.

[2] C. Klenk, D. Heim, M. Ugele, and O. Hayden, "Impact of sample preparation on holographic imaging of leukocytes," *Optical Engineering*, vol. 59, no. 10, p. 102403, 2019.

[3] M. Ugele, M. Weniger, M. Stanzel, M. Bassler, S. W. Krause, O. Friedrich, O. Hayden, and L. Richter, "Label-Free High-Throughput Leukemia Detection by Holographic Microscopy," *Advanced Science*, vol. 5, no. 12, 2018.

[4] T. L. Nguyen, S. Pradeep, R. L. Judson-Torres, J. Reed, M. A. Teitell, and T. A. Zangle, "Quantitative Phase Imaging: Recent Advances and Expanding Potential in Biomedicine," *American Chemical Society Nano*, vol. 16, no. 8, pp. 11 516–11 544, 2022.

[5] M. Nishikawa, H. Kanno, Y. Zhou, T.-H. Xiao, T. Suzuki, Y. Ibayashi, J. Harmon, S. Takizawa, K. Hiramatsu, N. Nitta *et al.*, "Massive image-based single-cell profiling reveals high levels of circulating platelet aggregates in patients with covid-19," *Nature Communications*, vol. 12, no. 1, pp. 1–12, 2021.

[6] J. J. Barcia, "The Giemsa stain: Its History and Applications," *International Journal of Surgical Pathology*, vol. 15, no. 3, pp. 292–296, 2007.

[7] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: contextualizing explainable machine learning for clinical end use," in *Machine Learning for Healthcare Conference*. PMLR, 2019, pp. 359–380.

[8] C. M. Cutillo, K. R. Sharma, L. Foschini, S. Kundu, M. Mackintosh, and K. D. Mandl, "Machine intelligence in healthcare - perspectives on trustworthiness, explainability, usability, and transparency," *Nature Partner Journals Digital Medicine*, vol. 3, no. 1, pp. 1–5, 2020.

[9] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[10] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1990, pp. 249–256.

[11] S. Hermawati and G. Lawson, "Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus?" *Applied Ergonomics*, vol. 56, pp. 34–51, 2016.

[12] A. W. Kushniruk, V. L. Patel, and J. J. Cimino, "Usability testing in medical informatics: cognitive approaches to evaluation of information systems and user interfaces." in *Proceedings of the American Medical Informatics Association Annual Fall Symposium*, 1997, p. 218.

[13] J. Zhang, T. R. Johnson, V. L. Patel, D. L. Paige, and T. Kubose, "Using usability heuristics to evaluate patient safety of medical devices," *Journal of Biomedical Informatics*, vol. 36, no. 1-2, pp. 23–30, 2003.

[14] S. Horton, K. A. Fleming, M. Kuti, L.-M. Looi, S. A. Pai, S. Sayed, and M. L. Wilson, "The Top 25 Laboratory Tests by Volume and Revenue in Five Different Countries," *American Journal of Clinical Pathology*, vol. 151, no. 5, pp. 446–451, 2018.

[15] A. Filby, "Sample preparation for flow cytometry benefits from some lateral thinking," *Cytometry Part A*, vol. 89, no. 12, pp. 1054–1056, 2016.

[16] S. K. Paidi, P. Raj, R. Bordett, C. Zhang, S. H. Karandikar, R. Pandey, and I. Barman, "Raman and quantitative phase imaging allow morpho-molecular recognition of malignancy and stages of B-cell acute lymphoblastic leukemia," *Biosensors and Bioelectronics*, vol. 190, p. 113403, 2021.

[17] M. Ugele, M. Weniger, M. Leidenberger, Y. Huang, M. Bassler, O. Friedrich, B. Kappes, O. Hayden, and L. Richter, "Label-free, high-throughput detection of P. falciparum infection in sphered erythrocytes with digital holographic microscopy," *Lab on a Chip*, vol. 18, pp. 1704–1712, 2018.

[18] M. Finsterbusch, W. C. Schrottmaier, J. B. Kral-Pointner, M. Salzmann, and A. Assinger, "Measuring and interpreting platelet-leukocyte aggregates," *Platelets*, vol. 29, no. 7, pp. 677–685, 2018.

[19] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.

[20] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2016.

[21] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" *Brain Informatics*, vol. 3, no. 2, pp. 119–131, 2016.

[22] H. R. Hartson, T. S. Andre, and R. C. Williges, "Criteria for evaluating usability evaluation methods," *International Journal of Human-Computer Interaction*, vol. 13, no. 4, pp. 373–410, 2001.

[23] J. Nielsen, "Severity ratings for usability problems," *Papers and Essays*, vol. 54, pp. 1–2, 1995.

[24] A. Sears, "Heuristic walkthroughs: Finding the problems without the noise," *International Journal of Human–Computer Interaction*, vol. 9, no. 3, pp. 213–234, 1997.

[25] W. D. Gray and M. C. Salzman, "Damaged merchandise? A review of experiments that compare usability evaluation methods," *Human Computer Interaction*, vol. 13, no. 3, pp. 203–261, 1998.

[26] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[27] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[28] J. Nielsen, "Enhancing the explanatory power of usability heuristics," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1994, pp. 152–158.

[29] ——. (2005) Ten usability heuristics. (accessed 2022.12.17). [Online]. Available: https://www.informaticathomas.nl/heuristicsNielsen.pdf

[30] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen *et al.*, "Guidelines for human-AI interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13.

[31] B. Shneiderman, *Designing the user interface: Strategies for effective human-computer interaction*. Addison-Wesley, 1998.

[32] H. Lieberman, "User interface goals, AI opportunities," *AI Magazine*, vol. 30, no. 4, pp. 16–22, 2009.

[33] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[34] T. G. Gill, "Expert systems usage: Task change and intrinsic motivation," *Management Information Systems Quarterly*, pp. 301–329, 1996.

[35] E. Horvitz, "Principles of mixed-initiative user interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1999, pp. 159–166.

[36] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *Nature Partner Journals Digital Medicine*, vol. 3, no. 1, pp. 1–10, 2020.

[37] A. D. Jameson, "Understanding and dealing with usability side effects of intelligent processing," *AI Magazine*, vol. 30, no. 4, pp. 23–23, 2009.

[38] C. Rzepka and B. Berger, "User interaction with AI-enabled systems: a systematic review of is research," in *Thirty Ninth International Conference on Information Systems*, vol. 39, 2018, pp. 1–17.

[39] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[40] High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, European Commission, 2019.

[41] J. Nielsen, "Finding usability problems through heuristic evaluation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1992, pp. 373–380.

[42] J. Noyes and C. Baber, *User-centred design of systems*. Springer Science & Business Media, 1999.

[43] J. Nielsen and T. K. Landauer, "A mathematical model of the finding of usability problems," in *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*, 1993, pp. 206–213.

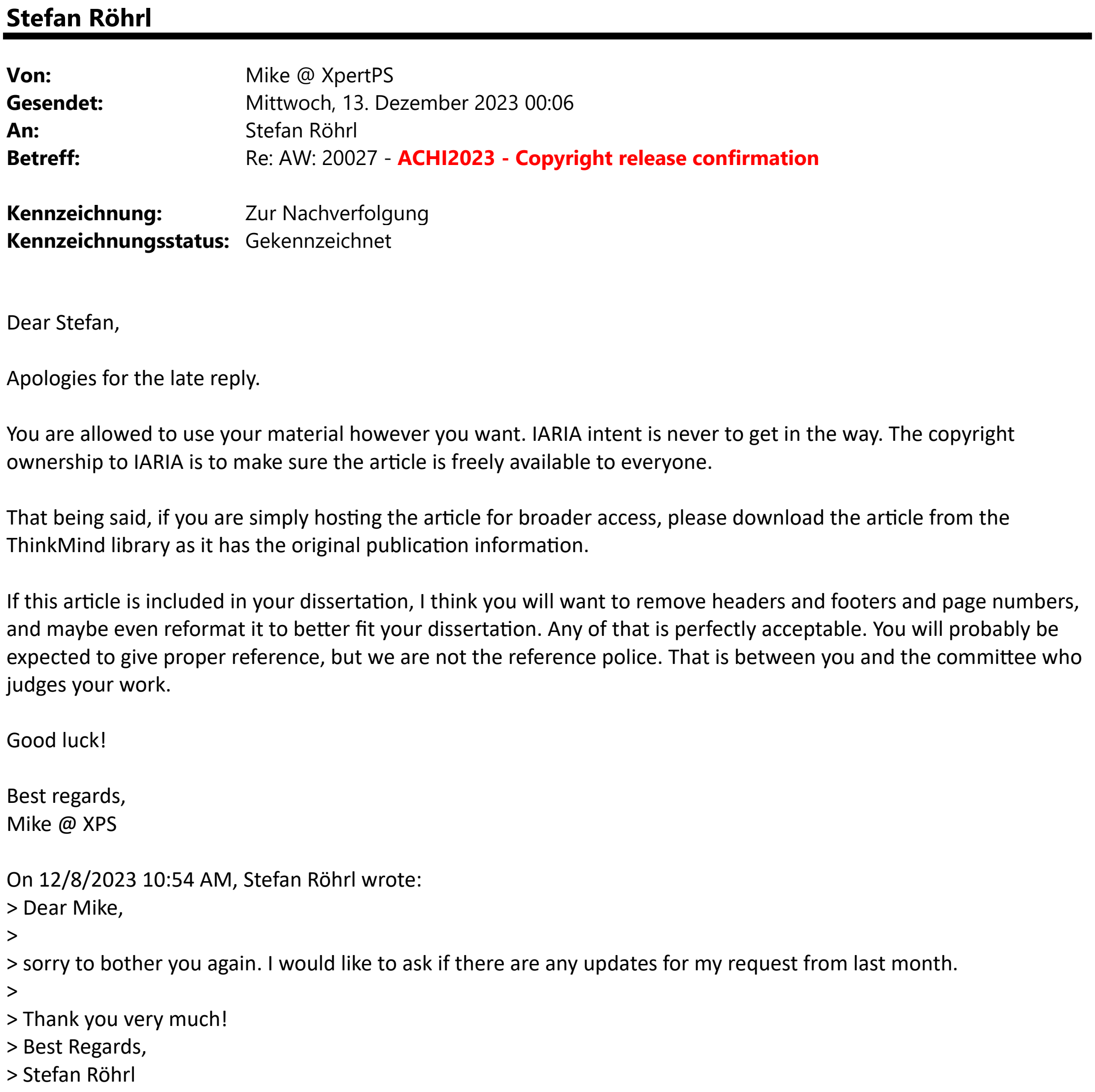
[44] J. R. Lewis, "Sample sizes for usability studies: Additional considerations," *Human factors*, vol. 36, no. 2, pp. 368–378, 1994.

[45] M. W. Jaspers, T. Steen, C. Van Den Bos, and M. Geenen, "The think aloud method: a guide to user interface design," *International Journal of Medical Informatics*, vol. 73, no. 11-12, pp. 781–795, 2004.

[46] C. Jimenez, P. Lozada, and P. Rosas, "Usability heuristics: A systematic review," in *2016 IEEE 11th Colombian Computing Conference (CCC)*. IEEE, 2016, pp. 1–8.

Copyright

The publisher granted permission to reprint the publication in this document on December 13th, 2023.



Stefan Röhrl

Von: Mike @ XpertPS
Gesendet: Mittwoch, 13. Dezember 2023 00:06
An: Stefan Röhrl
Betreff: Re: AW: 20027 - **ACHI2023 - Copyright release confirmation**

Kennzeichnung: Zur Nachverfolgung
Kennzeichnungsstatus: Gekennzeichnet

Dear Stefan,

Apologies for the late reply.

You are allowed to use your material however you want. IARIA intent is never to get in the way. The copyright ownership to IARIA is to make sure the article is freely available to everyone.

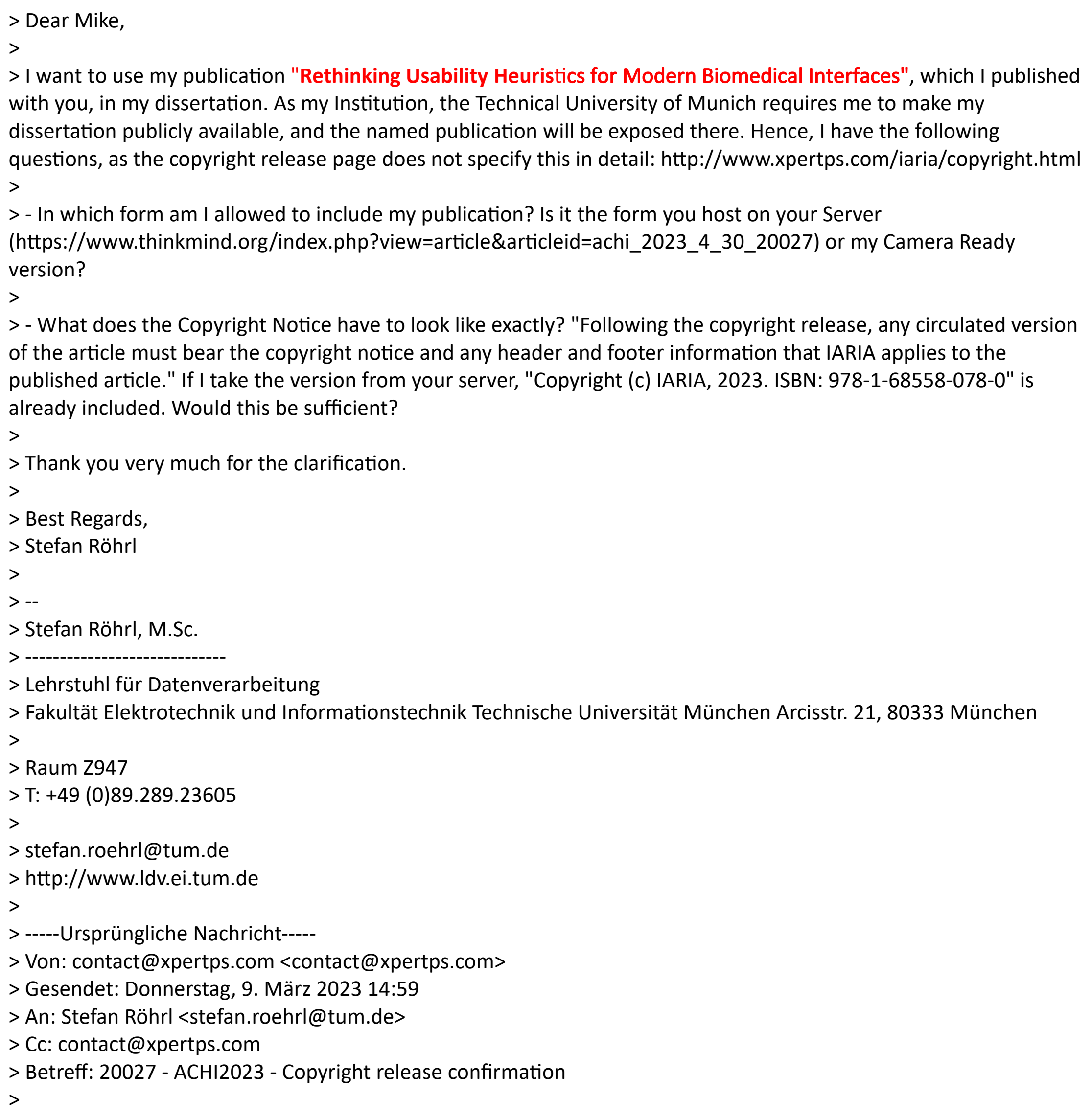
That being said, if you are simply hosting the article for broader access, please download the article from the ThinkMind library as it has the original publication information.

If this article is included in your dissertation, I think you will want to remove headers and footers and page numbers, and maybe even reformat it to better fit your dissertation. Any of that is perfectly acceptable. You will probably be expected to give proper reference, but we are not the reference police. That is between you and the committee who judges your work.

Good luck!

Best regards,
Mike @ XPS

On 12/8/2023 10:54 AM, Stefan Röhrl wrote:
> Dear Mike,
>
> sorry to bother you again. I would like to ask if there are any updates for my request from last month.
>
> Thank you very much!
> Best Regards,
> Stefan Röhrl



> Dear Mike,
>
> I want to use my publication **"Rethinking Usability Heuristics for Modern Biomedical Interfaces"**, which I published with you, in my dissertation. As my Institution, the Technical University of Munich requires me to make my dissertation publicly available, and the named publication will be exposed there. Hence, I have the following questions, as the copyright release page does not specify this in detail: http://www.xpertps.com/iaria/copyright.html
>
> - In which form am I allowed to include my publication? Is it the form you host on your Server (https://www.thinkmind.org/index.php?view=article&articleid=achi_2023_4_30_20027) or my Camera Ready version?
>
> - What does the Copyright Notice have to look like exactly? "Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article." If I take the version from your server, "Copyright (c) IARIA, 2023. ISBN: 978-1-68558-078-0" is already included. Would this be sufficient?
>
> Thank you very much for the clarification.
>
> Best Regards,
> Stefan Röhrl
>
> --
> Stefan Röhrl, M.Sc.
> -----------------------------
> Lehrstuhl für Datenverarbeitung
> Fakultät Elektrotechnik und Informationstechnik Technische Universität München Arcisstr. 21, 80333 München
>
> Raum Z947
> T: +49 (0)89.289.23605
>
> stefan.roehrl@tum.de
> http://www.ldv.ei.tum.de
>
> -----Ursprüngliche Nachricht-----
> Von: contact@xpertps.com <contact@xpertps.com>
> Gesendet: Donnerstag, 9. März 2023 14:59
> An: Stefan Röhrl <stefan.roehrl@tum.de>
> Cc: contact@xpertps.com
> Betreff: 20027 - ACHI2023 - Copyright release confirmation
>