



FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Artificial Intelligence-based Technologies for Biophotonic Data

Auf künstlicher Intelligenz basierende Technologien für
biophotonische Daten

**Dissertation
(Kumulativ)**

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat)

vorgelegt dem Rat der Chemisch-Geowissenschaftlichen Fakultät
der Friedrich-Schiller-Universität Jena

von Msc. Pranita Rajan Pradhan
geboren am 18. Juni 1991 in Thane, Indien

1. Gutachter

Prof. Dr. Jürgen Popp

Institut für Physikalische Chemie

Friedrich-Schiller Universität Jena

2. Gutachter

PD. Dr. Thomas Wilhelm Bocklitz

Institut für Physikalische Chemie

Friedrich-Schiller Universität Jena

Tag der öffentlichen Verteidigung: 2. Juni 2021

ABSTRACT

For decades, biophotonic technologies have been booming in fields as diverse as biology, medicine, pharmaceutical science, environmental science, and agriculture. These technologies reveal not only structural but also molecular and functional changes in the sample under investigation. Furthermore, biophotonic technologies have such prominent advantages as high molecular sensitivity, high usability, high compactness, and high spatial and temporal resolution. Due to these advantages, biophotonic technologies have great potential in clinical applications. A typical example of biophotonic technology in the clinic is optical coherence tomography (OCT), which is a powerful diagnostic tool in ophthalmology. Likewise, flow cytometry and endoscopy are routinely used for cancer diagnosis. Nowadays, researchers emphasize the use of biophotonic technologies for point-of-care testing in clinics and the *in vivo* imaging of live cells for automating the disease diagnosis workflow. Furthermore, researchers are also focusing on integrating multiple biophotonic technologies in a single unit for understanding diseases at the cellular, molecular, and tissue level. Such ever-increasing developments in biophotonic technologies result in a massive amount of biophotonic data, and analysis of large biophotonic data by a human being is challenging. Therefore, algorithms that can automatically analyze biophotonic data to extract useful “patterns” like an experienced person are crucial. Extracting patterns from data using algorithms which can imitate human intelligence by learning from the data itself is categorized into a field of “artificial intelligence” (AI). Utilizing AI to analyze data from biophotonic technologies like Raman spectroscopy, coherent anti-Stokes Raman scattering (CARS) microscopy, two-photon excitation fluorescence (TPEF) microscopy, and second-harmonic generation (SHG) microscopy is the main highlight of this thesis. Concisely, this thesis will use AI and biophotonic data for biomedical applications like the prediction of disease, segmentation of various regions in tissue, and transformation of one modality into another modality. The results in this thesis will show that utilizing AI, along with biophotonic technologies, can benefit the field of biomedicine and the life sciences.

KURZFASSUNG

In den letzten Jahrzehnten werden biophotonische Technologien in verschiedene Bereichen wie Biologie, Medizin, Pharmazie, Umweltwissenschaften und Landwirtschaft eingesetzt. Dies resultiert aus dem Fakt, dass biophotonische Technologien nicht nur strukturelle, sondern auch molekulare und funktionelle Veränderungen in der untersuchten Probe aufzeigen. Darüber hinaus besitzen biophotonische Technologien Vorteile wie eine hohe molekulare Empfindlichkeit, eine hohe Benutzerfreundlichkeit, die Kompaktheit der Messgeräte und hohe räumliche und zeitliche Auflösung. Aufgrund dieser Vorteile haben biophotonische Technologien ein großes Potenzial für klinische Anwendungen. Ein typisches Beispiel für biophotonische Technologien, welche in der Klinik eingesetzt werden, ist die optische Kohärenztomographie (OCT), die ein leistungsfähiges diagnostisches Werkzeug in der Augenheilkunde darstellt. Ebenso werden Durchflusszytometrie und Endoskopie routinemäßig zur Krebsdiagnose eingesetzt. Des Weiteren wird der Einsatz von biophotonischen Technologien für Point-of-Care-Tests in Kliniken und die In-vivo-Bildgebung lebender Zellen zur Automatisierung des Arbeitsablaufs bei der Krankheitsdiagnose erforscht. Darüber hinaus konzentrieren sich die Forscher auch auf die Integration mehrerer biophotonischer Technologien in einem einzigen Messgerät um ein Verständnis von Krankheiten auf zellulärer, molekularer und Gewebeebene zu erreichen. Solche immer weiter fortschreitenden Entwicklungen in der Biophotonik führen zu einer riesigen Menge biophotonischer Daten, und die manuelle Analyse großer biophotonischer Daten ist eine Herausforderung. Daher sind Algorithmen, die biophotonische Daten automatisch analysieren können, um nützliche "Muster" zu extrahieren, von entscheidender Bedeutung. Das Extrahieren von Mustern aus Daten mittels Algorithmen, welche die menschliche Intelligenz imitieren können indem sie aus den Daten selbst lernen, wird in einen Bereich der "künstlichen Intelligenz" (KI) eingeordnet. Die Nutzung der künstlichen Intelligenz zur Analyse von Daten aus biophotonischen Technologien wie der Raman-Spektroskopie, der kohärenten Anti-Stokes-Raman-Streuung (CARS), der Zwei-Photonen-Anregungs-Fluoreszenz-Mikroskopie (TPEF) und der Zweiten Harmonischen Generation (SHG) Mikroskopie ist das Gegenstand dieser Arbeit. In dieser Arbeit werden KI und biophotonische Daten für biomedizinische Anwendungen wie die Vorhersage von Krankheiten, die Segmentierung verschiedener Regionen im Gewebe und die Transformation einer Modalität in eine andere Modalität verwendet. Die Ergebnisse dieser Arbeit zeigen, dass die Nutzung von KI Verfahren zusammen mit biophotonischen Technologien in der Biomedizin und den Biowissenschaften ein großes Potential besitzen.

Contents

List of figures	iii
List of acronyms	iv
Mathematical notations	v
1 Introduction	1
1.1 Motivation	5
1.2 Organization of the thesis	6
2 Raman spectroscopy and non-linear multimodal imaging	7
2.1 Raman spectroscopy for biomedical applications	7
2.2 Non-linear multimodal imaging for biomedical applications	9
2.2.1 Coherent anti-Stokes Raman scattering	11
2.2.2 Two-photon excitation fluorescence microscopy	12
2.2.3 Second harmonic generation microscopy	13
2.3 The need of Chemometrics for biophotonics	13
3 Artificial intelligence for biophotonics	15
3.1 Fundamentals of AI and its applications	16
3.2 Basics of machine learning	17
3.2.1 Dataset	18
3.2.2 Supervised learning	19
3.2.3 Unsupervised learning	20
3.2.4 Cross-validation and performance measure	22
3.2.5 Challenges motivating deep learning	23
3.3 Basics of deep learning	24
3.3.1 Artificial neural networks: A biological motivation	24
3.3.2 Neural network architectures	25
3.3.3 Statistical learning and non-linear parameter optimization	29
3.3.4 The backpropagation algorithm	31
3.4 Basics of transfer learning	32
4 Selected work	33
4.1 Identification of sepsis using non-linear multimodal images	35

4.2	Semantic segmentation of non-linear multimodal images	37
4.3	Pseudo H&E staining of non-linear multimodal images	39
4.4	Interpretation of deep learning models	41
4.5	Transfer learning for breast cancer diagnosis using data fusion	43
5	Summary	45
6	Zusammenfassung	49
7	Future research directions	55
	Bibliography	57
8	Publications, manuscripts and conference proceeding	67
P1	Towards an Interpretable Classifier for Characterization of Endoscopic Mayo Scores in Ulcerative Colitis Using Raman Spectroscopy	69
P2	Deep learning a boon for Biophotonics?	85
P3	Nonlinear Multimodal Imaging Characteristics of Early Septic Liver Injury in a Mouse Model of Peritonitis	111
P4	Semantic Segmentation of Non-linear Multimodal Images for Disease Grading of Inflammatory Bowel Disease: A SegNet-based Application	127
P5	Computational tissue staining of non-linear multimodal imaging using Generative Adversarial Networks	139
P6	Data fusion of histological and immunohistochemical image data for breast cancer diagnostics using Transfer learning	165
9	Appendix	179
A1	Computational tissue staining of non-linear multimodal imaging using supervised and unsupervised deep learning	181
A2	Data Fusion of Histological and Immunohistochemical Image Data for Breast Cancer Diagnostics using Transfer Learning	209
	Peer-Reviewed Publications	223
	Conferences and summer school	224
	Workshop	225
	Teaching and supervision activity	226
	Acknowledgements	227
	Declaration	229

List of Figures

2.1	Raman spectroscopy.	9
2.2	Non-linear multimodal imaging.	12
3.1	Overview of AI, ML, DL.	17
3.2	Classification vs. Regression.	18
3.3	Biological motivation of Artificial neural networks.	25
3.4	Artificial neural networks or multi-layer perceptron.	26
3.5	Convolutional neural network.	27
3.6	Visualization of non-convex error function of ANNs.	30
4.1	AI-based applications for spectroscopic data.	34
4.2	Image classification of non-linear multimodal images.	36
4.3	Semantic segmentation of non-linear multimodal images.	38
4.4	Image translation of non-linear multimodal images.	40
4.5	Spectra classification using deep convolutional neural network.	42
4.6	Transfer learning for deep learning models.	44

List of acronyms

AE	Autoencoders
ANN	Artificial neural networks
CARS	Coherent anti-Stokes Raman scattering
CGAN	Conditional generative adversarial network
CNN	Convolutional neural network
CPU	Central processing unit
GAN	Generative adversarial network
GPU	Graphical processing unit
HE	Hematoxylin and eosin
MLP	Multi-layer perceptron
NIR	Near Infra-red
NLM	Non-linear multimodal imaging
MPM	Multi-photon microscopy
ReLU	Rectified linear unit
SERS	Surface-enhanced Raman spectroscopy
SGD	Stochastic gradient descent
SHG	Second harmonic generation
TPEF	Two-photon excitation fluorescence

Mathematical notations

a, A	A scalar random variable
\mathbf{a}	A column vector
\mathbf{A}	A matrix
a_i	Element i of vector \mathbf{a}
A_{ij}	Element i, j of matrix \mathbf{A} where i and j represents row index and column index
$A_{(i,:)}$	Row i of matrix \mathbf{A}
$A_{(:,j)}$	Column j of matrix \mathbf{A}
A^T	Transpose of matrix \mathbf{A}
\mathcal{C}, \mathbb{R}	A set
$\mathbf{a}^{(\tau)}$	A time point from a set of variables
\mathbf{a}^l	A layer of a deep learning model

*It's not about the destination to success,
it's about the journey to satisfaction.
-Unknown*

1

Introduction

With the introduction of the light microscope back in the 17th century until the systematic development by Carl Zeiss, Ernst Abbé, and Otto Schott in Jena in the 19th century, light-based or optical technologies have greatly influenced the life sciences and medicine. These light-based technologies are an integral part of the multidisciplinary research subject referred to as “biophotonics” [1]. Biophotonics is a field of science that deals with optical processes in biological systems. It utilizes light across the entire spectrum – from ultraviolet to the visible, infrared, and terahertz regions. Even though the word “biophotonics” has been coined recently, it dates to the 16th century when the first optical microscope to visualize biological tissue was invented. Since then, many inventions in biophotonic technologies and especially microscopes have been made. Today, the super-resolution microscope, the developers of which received the Nobel Prize in Chemistry in 2014, can visualize cellular structures even smaller than 20 nm [1]. These inventions greatly benefit the field of biomedicine.

One of the popular inventions in the biophotonic field is optical coherence tomography (OCT), which is a gold-standard technique in ophthalmology for detecting diseases like glaucoma in the eye. Similarly, technologies like light microscopy (e.g., fluorescence microscopy) and spectroscopy (e.g., infrared (IR) spectroscopy, Raman spectroscopy) are gaining popularity in healthcare. Some of the examples in healthcare include intraoperative recognition of tumor borders by fluorescence spectroscopy and Raman spectroscopy; early detection of tumors by one- and two-photon fluorescence spectroscopy; and identification of biomarkers associated with diseases and their progression by Raman spectroscopy [1]. This already creates a clear picture that biophotonic technologies can potentially improve biomedical diagnostics. However, it

is not practical to present all the biophotonic technologies due to time and resource limitation. Hence, this thesis will focus only on two biophotonic technologies, namely: Raman spectroscopy (see section 2.1) and non-linear multimodal imaging (see section 2.2). Before discussing Raman spectroscopy and non-linear multimodal imaging in detail, it is important to discuss light-matter interactions.

Light interacts with matter by either getting absorbed or emitted, but a part of the incident light is also scattered (1 out of 10^6 or 10^9 photons in incident light) [1]. Light scattering takes place in two forms: elastic and inelastic scattering of light. Elastic scattering of light occurs when the wavelength of the incident light is unchanged. Two examples of elastic scattering of light include both Rayleigh scattering caused by the interaction of light or photons with atoms, molecules, or phonons, and Mie scattering caused by spherical objects [2]. When the molecule is much smaller than the wavelength of light, Rayleigh scattering occurs [3]. It is because of the Rayleigh scattering of sunlight with the molecules in the Earth's atmosphere that the sky appears blue during daytime. Inelastic scattering of light occurs when the wavelength of scattered light is changed with respect to the wavelength of the incident light. The photons in inelastically scattered light either gain or lose energy. Example of this includes Raman scattering caused by discrete quantum states of molecules or phonons [2]. Raman scattering, or the Raman effect, is the basis of Raman spectroscopy and from here onwards Raman spectroscopy refers to vibrational Raman spectroscopy, i.e. Raman scattering on vibrational states of molecules.

As mentioned above, in Raman scattering, an incident photon with frequency ν_0 and energy $E_0 = \frac{hc}{\lambda_0} = \hbar\omega_0 = h\nu_0$ (where h is the Planck's constant; $\hbar = \frac{h}{2\pi}$) interacts with a molecule to either gain or lose energy. The change in energy $\Delta E = E_1 - E_2$ of the scattered photon depends on the frequency of vibration of the molecule ν_m . The vibrational frequency of the molecule is dependent on the strength of the chemical bond and masses of the atoms in the molecule. Therefore, the vibrational frequencies are unique to the functional group of the molecule. As the vibrational frequencies are unique to the molecule and significantly influence the energy change in the scattered photon, Raman scattering retains "fingerprint" information of the molecule [3].

The change in the energy of the scattered photon can result in Stokes and anti-Stokes Raman scattering. Specifically, if the energy of the scattered photon is less than the incident photon, then Stokes Raman scattering occurs in a way that corresponds to a frequency $\nu_0 - \nu_m$. If the energy of the scattered photon is larger than the incident photon, then anti-Stokes Raman scattering occurs with a frequency $\nu_0 + \nu_m$ [4]. If the energy of the scattered and the incident photon is the same, then Rayleigh scattering occurs with a frequency equal to the incident frequency ν_0 (see figure 2.1A). Raman spectroscopy mostly refers to Stokes Raman scattering unless specified otherwise, because it's the dominant process at room temperature. The energy difference between the incident photon and the inelastically scattered photon is expressed as

wavenumber (cm^{-1}). A plot of the intensity of the inelastically scattered light as a function of the wavenumber is called the Raman spectrum (see figure 2.1B). More details, including the classical theory of Raman effect and biomedical applications of Raman spectroscopy are discussed in chapter 2.

Raman scattering is a very weak effect [4]. Therefore, attempts to improve the Raman scattering signal have been witnessed in the past few years. One such attempt is coherent anti-Stokes Raman spectroscopy (CARS) [5], which is a non-linear optical process used to enhance the Raman signal. This thesis will address CARS and other two non-linear optical techniques, namely: two-photon excited fluorescence (TPEF) microscopy and second harmonic generation (SHG) microscopy. Biophotonic technologies like CARS, TPEF, and SHG are collectively referred to as non-linear multimodal (NLM) imaging. The term “non-linear” used for these technologies is due to their non-linear dependence on the incident or excitation (or laser) light intensity. In other words, the inelastic (or Raman) scattering of light discussed so far is a spontaneous process, i.e. the excitation of molecules is generated by a single exciting (or incident) frequency $\omega_0 = 2\pi\nu_0$ [1]. However, in CARS, the molecules in a sample are excited by at least two different incident frequencies (ω_p and ω_s) using two laser beams. The frequency difference of the two laser beams is tuned to match the molecular vibration frequency ω_m (see figure 2.2A). The absorption of two photons results in the excitation of coherent molecular vibrations, which is recorded as the CARS signal [6]. CARS microscopy can be applied to visualize any molecular vibration; however, it is mainly used to image the aliphatic C-H-stretching vibration of methylene groups for visualizing lipids [6]. Like CARS microscopy, SHG microscopy is also a non-linear scattering process in which two incident photons ω_p are coherently scattered into a photon of twice the frequency ($2\omega_p$). SHG microscopy is used to visualize quasi-crystalline structures in tissues like collagen, tubulin, and cholesterol. Lastly, TPEF microscopy, which is neither a coherent nor a scattering process [6], is caused by absorbing two photons by a fluorophore to induce molecular transition. TPEF microscopy is used to visualize prominent autofluorophores in animal tissue, including proteins like elastin and keratin, pigments like melanin, and enzymes like NADH and flavines. Researchers have already shown that the TPEF and SHG signals can be acquired along with the CARS signal without damaging the sample under investigation [6]. Likewise, it has also been shown that the combination of these three optical technologies provides not only various biomolecular information but also high spatial resolution and considerable penetration depth, making it suitable for biological studies. More information about CARS, TPEF, and SHG is provided in section 2.2.

So far, two biophotonic technologies have been discussed in this thesis: Raman spectroscopy and NLM imaging. Now the stage is set to introduce the dataset obtained by these biophotonic technologies and the challenges encountered during analysis of

their dataset. Raman spectroscopy provides data in the form of a spectrum. Each Raman spectrum is a plot of the intensity of the inelastically scattered light along the y-axis against the wavenumber along the x-axis (see figure 2.1B). The NLM imaging provides RGB images in which the red channel is the CARS signal, the green channel is the TPEF signal, and the blue channel is the SHG signal (see figure 2.1B). Both the Raman spectroscopic and the NLM imaging data, in statistical terms, can be categorized as multivariate data because each observation comprises more than two inter-related variables. Taking the Raman spectrum as an example, the number of variables in a single spectrum is equal to the number of wavenumbers. Furthermore, complex relationships between these variables or wavenumbers can exist. Likewise, a non-linear multimodal image provides information in the form of color, texture, and intensity of pixels. The intensity values of a pixel, i.e. three intensity values in the case of an RGB image, may or may not be related to the intensity values of neighboring pixels. Thus, interpretation of the pixel intensity values is difficult, and its analysis by a non-expert poses several challenges. A detailed explanation of the challenges encountered for analyzing Raman spectroscopic and NLM imaging data is given in section 2.3.

The multivariate nature of Raman spectroscopic and NLM imaging data requires systematic recognition of important “patterns” or “features”. Additionally, both datasets inherently exhibit the “curse of dimensionality” due to the enormously large dimension of variables, which affects the classification and organization of these datasets. Thus, analysis of such multivariate high-dimensional data is a crucial task. For purpose of analysis, a broad field of chemistry called “chemometrics” plays an important role. The first successful application of chemometrics was for the analysis of the fluorescence spectra of mixtures in which the number of fluorescent components in a mixture was quantitatively determined [7]. Subsequently, various applications for interpreting biological data using chemometric methods were reported in literature [1, 8, 9] (e.g., chemometric approaches for pre-processing Raman spectra and, specifically, for removing noise or unwanted spikes, instrument calibration effects, and unwanted shifts in the baseline [10]). Chemometric approaches are also used for the classification of Raman spectra into similar groups and studying the heterogeneous nature of biomolecules such as proteins, lipids, carbohydrates, and nucleic acids in spectroscopic data. Likewise, various clinical applications of spectroscopic data like disease recognition [11, 12, 13], tumor detection [14, 15], and tissue characterization [16, 17] are implemented using chemometrics. A recent trend in chemometrics is to use advanced algorithms for performing biophotonic data analysis. These algorithms are categorized under the field of artificial intelligence (AI).

Artificial intelligence is a science which uses machines or algorithms to perform tasks like a human. For more than a decade, AI has been gaining in popularity in social media, E-commerce platforms like Amazon, travelling aids like Google Maps,

and the personal voice assistants on smartphones like Siri and Alexa [18, 19, 20]. Likewise, AI-based technologies are used to understand complex chemical systems related to proteomics, genomics, and metabolomics [21, 22, 23]. AI-based technologies to study molecular distributions in biological samples using biophotonic techniques is also gathering interest [24]. The above applications have gained success due to numerous (beneficial) characteristics of AI. One of the most important characteristics of AI is data ingestion, i.e. the way an AI algorithm handles a vast amount of data in multiple formats, from multiple sources generated at multiple times. The next important characteristic is the imitation of human cognition, i.e. the ability of AI algorithms to imitate the human mind and solve complex problems and tasks (e.g., self-driving cars). Lastly, machine learning and deep learning (see figure 3.1), which is the most upcoming field of AI, refers to the capability of learning features from data automatically without any manual intervention. Knowing the beneficial characteristics of AI, this thesis proposes AI algorithms, especially deep learning algorithms, as an alternative approach for analyzing biophotonic data.

Finally, the discussions made so far bring us to two essential questions: “*Why utilize biophotonics in clinics?*” and “*Why use AI for biophotonic data analysis?*”. These questions will be answered in the motivation of this thesis.

1.1 MOTIVATION

This thesis has two primary motivations that will answer the two preceding questions. Both motivations drive towards building a fast disease diagnosis workflow and an automatic decision-making system for clinics and hospitals.

The first motivation arises from a clinical perspective and answers the question “*Why utilize biophotonics in clinics?*”. Since biophotonic technologies are label-free and molecularly sensitive, provide high spatial resolution, and have large penetration depth, these technologies can substantially benefit health professionals for understanding diseases at the cellular, molecular, and tissue levels. Usage of biophotonic technologies can also assure point-of-care diagnosis by facilitating routine examinations of high-risk patients. Furthermore, these technologies are highly practical due to their small-sized devices, minimal sample preparation, and parallel imaging of multiple biomarkers. Minimal sample preparation is also beneficial for *in vivo* disease diagnosis. Finally, increasing the use of biophotonic technologies in clinics will also reduce the ever-rising burden on well-established radiological techniques like computed tomography (CT) and magnetic resonance imaging (MRI).

The second motivation arises from the chemometric perspective and answers the question “*Why use AI for biophotonic data analysis?*”. Using AI for biophotonic data analysis is one step forward for automatic disease diagnosis workflow. This can be achieved in the following way. A biophotonic technology that can be used *in*

in vivo should be provided with a built-in AI model to assure real-time disease diagnosis. Furthermore, as AI models can extract useful patterns or features similar to an experienced professional, these models can reduce the effort of a data analyst for performing feature extraction. Another advantage is that AI-based models can complement the knowledge of doctors and physicians and help them to gain new insights into numerous diseases. Lastly, aging patients with chronic diseases can use medical devices with built-in AI models to perform self-testing and avoid frequent visitation to clinics.

As AI for biophotonic data analysis is still in the infancy stage, this thesis motivates the use of AI, machine learning, and deep learning to analyze biophotonic data like Raman spectroscopy and NLM imaging. To the best knowledge of the author, this thesis presents for the first time the analysis of Raman spectroscopic and NLM imaging data using AI, specifically deep learning, for clinical applications.

1.2 ORGANIZATION OF THE THESIS

This thesis is organized into following chapters.

- Chapter 1 briefly introduces the two biophotonic modalities: Raman spectroscopy and non-linear multimodal imaging. Furthermore, the motivation of the thesis is introduced in this chapter. Additionally, the acronyms and mathematical notations required for a clear understanding of the theoretical background are explained at the beginning of this thesis.
- Chapter 2 provides a theoretical explanation along with biomedical applications of Raman spectroscopy and non-linear multimodal imaging. This chapter also discusses chemometrics, which is crucial for analysing Raman spectroscopic and non-linear multimodal imaging data. Lastly, chapter 3 revisits fundamental concepts widely used in chemometrics like artificial intelligence, machine learning and deep learning.
- Chapter 4 is the essential part of this thesis and summarizes various scientific contributions developed to improve state-of-the-art approaches for the analysis of Raman spectroscopic and non-linear multimodal imaging data.
- Lastly, the conclusion of this thesis and is divided into four chapters: chapter 5 to chapter 8. Chapter 5 and chapter 6 are the summary of this thesis with English and German versions, respectively. Chapter 7 is the outlook of the dissertation with possible directions leading to future research. Some methods discussed here can potentially achieve the clinical translation of the state-of-the-art methods. Finally, chapter 8 comprises various publications in peer-reviewed journals, international conference proceedings, and manuscripts made during PhD tenure.

2

Raman spectroscopy and non-linear multimodal imaging

The previous chapter introduced various biophotonic technologies that could potentially benefit the biomedical field. It discussed the two biophotonic technologies used in this thesis: Raman spectroscopy and non-linear multimodal (NLM) imaging. The previous chapter also mentioned the use of chemometrics for analyzing biophotonic data. Now, this chapter will explain Raman spectroscopy and NLM imaging in detail, along with their biological applications. This chapter will further elaborate on the application of chemometrics in section 2.3, particularly for analyzing Raman spectroscopic data and NLM imaging data. In this regard, this chapter begins with Raman spectroscopy in section 2.1 and NLM imaging in section 2.2. Further, the subsections 2.2.1 to 2.2.3 elaborate on the following three non-linear optical modalities: CARS, TPEF, and SHG microscopy.

2.1 RAMAN SPECTROSCOPY FOR BIOMEDICAL APPLICATIONS

The basic idea of (vibrational) Raman scattering was already introduced in chapter 1. This section will present the classical theory of Raman scattering and biological applications of Raman spectroscopy.

The classical theory of the Raman effect can be explained by considering the electric nature of molecules. According to this theory, a molecule is a collection of oppositely charged particles (nucleus-electron dipole), the relative positions of which can be altered by the application of an external electric field of light. Because of the application

of the external electric field, an electric dipole moment is induced in the molecule. The induced dipole moment oscillates with the frequency of the applied electric field and attempts to restore its relative position. The ease with which an electric field can distort the relative position of the electron cloud around the nucleus in a molecule is called the electric “polarizability” (α). Another term in this context is “polarization” (\vec{P}) and refers to the total dipole moment per unit volume. The polarization \vec{P} in a molecule caused by an external electric field strength $\vec{E} = \vec{E}_0 \cos(2\pi\nu_0 t)$ of an electromagnetic wave (or laser beam) fluctuating with time t is given by

$$\vec{P} = \underbrace{\alpha_0 \vec{E}_0 \cos(2\pi\nu_0 t)}_{\text{Rayleigh scattering}} + \underbrace{\frac{1}{2} \left(\frac{\partial \alpha}{\partial q} \right)_0 q_0 \vec{E}_0 \cos(2\pi(\nu_0 - \nu_m)t)}_{\text{Stokes Raman scattering}} + \underbrace{\frac{1}{2} \left(\frac{\partial \alpha}{\partial q} \right)_0 q_0 \vec{E}_0 \cos(2\pi(\nu_0 + \nu_m)t)}_{\text{anti-Stokes Raman scattering}}, \quad (2.1)$$

where E_0 is the vibrational amplitude and ν_0 is the frequency of the external electric field. The vibrational frequency of the molecule is denoted as ν_m . The polarizability and the vibrational amplitude of the molecule in equilibrium position are given by α_0 and q_0 , respectively. The term $(\partial\alpha/\partial q)_0$ denotes the rate of change of polarizability of the molecule with respect to the change in nuclear displacement q [4]. In the equation above, the first term corresponds to Rayleigh scattering in which the oscillating dipole radiates light with a frequency (ν_0) equal to the frequency of the applied electric field. In comparison, the second and third term corresponds to Stokes Raman scattering and anti-Stokes Raman scattering with frequencies $\nu_0 - \nu_m$ and $\nu_0 + \nu_m$, respectively. The derivation for the equation above can be found in reference [4]. From the equation above, it can be derived that Raman scattering is possible only if the polarizability of a molecule (α) changes by the molecular vibration, i.e. $(\partial\alpha/\partial q)_0 \neq 0$. In such cases, the molecule is said to be Raman-active [4]. It is possible to study all Raman-active molecules from a Raman spectrum. A Raman spectrum shows Rayleigh, Stokes Raman and anti-Stokes Raman lines, however, at the normal temperature it is customary to measure only the Stokes side of the spectrum as both (Stokes and anti-Stokes) provide the same information. A typical example of Stokes Raman spectra for four disease stages of inflammatory bowel disease (IBD) is shown in figure 2.1B. Here, the black, green, blue and red spectra correspond with the lowest to the highest stage of IBD. The Raman spectrum in figure 2.1B shows the fingerprint spectral region of $500\text{-}1800\text{ cm}^{-1}$ and $2800\text{-}3020\text{ cm}^{-1}$. Figure 2.1B also indicates prominent peaks with high Raman intensity at wavenumber positions 1002 cm^{-1} for the phenylalanine band, 1440 cm^{-1} for the CH_2 deformation band, $1680\text{-}1620\text{ cm}^{-1}$ for the amide I band, and $3020\text{-}2800\text{ cm}^{-1}$ for the CH stretching intensities [25].

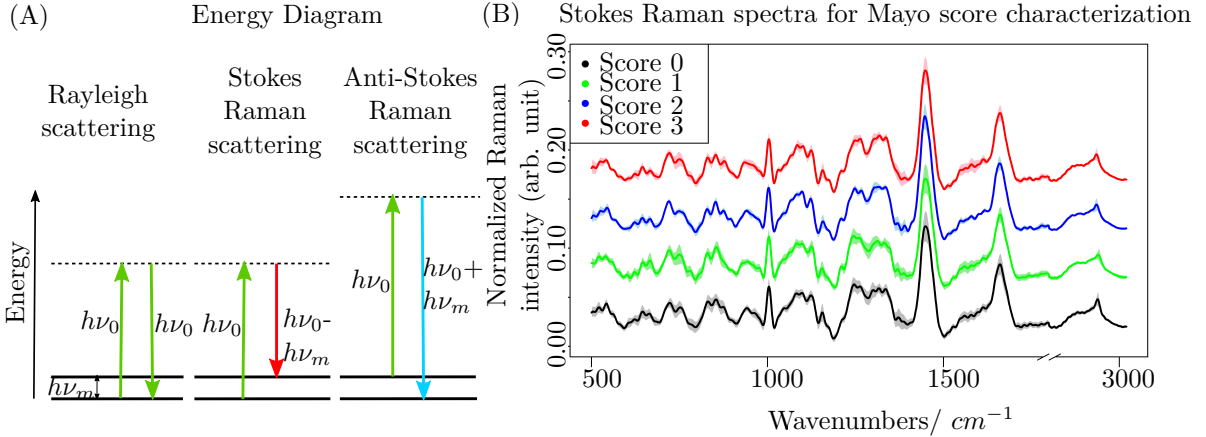


Figure 2.1: In this figure (A) shows an energy diagram for Rayleigh scattering (green), Stokes Raman scattering (red), and anti-Stokes Raman scattering (blue); the black dotted line represents the “virtual” energy level of the molecule. (B) shows the plot of a Stokes Raman spectrum obtained from four different stages of inflammatory bowel disease with the x-axis and y-axis indicating the wavenumber and normalized Raman intensity, respectively.

Until now, this section summarizes the classical theory of Raman scattering. The part following in this section will cover the clinical applications of Raman spectroscopy. Raman spectroscopy is a non-destructive and highly molecularly selective technique. Furthermore, Raman spectroscopy requires minimal sample preparation and is not affected by the presence of water [4]. Due to these properties, Raman spectroscopy is suitable for *in vivo* applications. A few biomedical applications of Raman spectroscopy are listed in references [26, 27, 28, 29]. Applications like the identification of biological cell and tissue types [30], characterization of human skin, detection of various types of diseases [31], detection of premalignant lesions [32, 33], and cancer related studies [14, 34, 35, 36, 37] are just a few of the many clinically relevant examples. This thesis will present an application of Raman spectroscopy for the characterization of stages of inflammatory bowel disease in section 4.5.

2.2 NON-LINEAR MULTIMODAL IMAGING FOR BIOMEDICAL APPLICATIONS

It was previously mentioned in chapter 1 that the NLM imaging used in this thesis is a combination of three non-linear optical imaging techniques, namely: coherent anti-Stokes Raman scattering (CARS) microscopy, two-photon excitation fluorescence (TPEF) microscopy, and second-harmonic generation (SHG) microscopy. As the name suggests, all non-linear imaging techniques are obtained due to the non-linear interaction of light with molecules, i.e. the interaction of light with a molecule mediated by two or more photons. In this section, the mathematical expression of

non-linear optical effects, followed by the properties of non-linear optical processes, is discussed. Further, a detailed explanation of the three non-linear optical modalities CARS, TPEF, and SHG will be provided.

To understand the mathematical description of non-linear optical effects, a molecule can be considered as an aggregation of a positively charged nucleus and negatively charged electrons. In the presence of an external electric field, the bound electrons tend to be slightly displaced from their equilibrium positions. When the electric field \vec{E} arises from the normal light intensities, the induced polarization \vec{P} of the molecule is linearly proportional to the applied electric field. Such linear dependence is the basis of the linear optical phenomena of Rayleigh and Raman scattering explained above [1]. However, when a large electric field arising from high light intensities (e.g., light from high ultrashort lasers) is applied, the induced polarization \vec{P} of the molecule is no longer linearly proportional to the applied electric field. This is because the electron is displaced from its equilibrium position even farther and can no longer be represented by harmonic representation as expressed by Hooke's law [1]. In such cases, anharmonic representation of the induced polarization \vec{P} by the Taylor series expansion in the applied electric field is considered. The non-linearity between polarization \vec{P} and the electric field \vec{E} is given as

$$\begin{aligned}\vec{P} &= \epsilon_0 \cdot (\overleftrightarrow{\chi}^{(1)} \vec{E} + \overleftrightarrow{\chi}^{(2)} \vec{E} : \vec{E} + \overleftrightarrow{\chi}^{(3)} \vec{E} : \vec{E} : \vec{E} + \dots) \\ &= \epsilon_0 \cdot (\overleftrightarrow{\chi}^{(1)} \vec{E} + \underbrace{\overleftrightarrow{\chi}^{(2)} \vec{E}^2}_{\text{SHG}} + \underbrace{\overleftrightarrow{\chi}^{(3)} \vec{E}^3}_{\text{CARS, TPEF}} + \dots).\end{aligned}\quad (2.2)$$

Here $\overleftrightarrow{\chi}^{(n)}$ is n^{th} order susceptibility, which is a tensor of rank $n + 1$, and ϵ_0 refers to the permittivity of vacuum. The complete derivation of the equation above is given in reference [1]. From the equation above, the first term $\overleftrightarrow{\chi}^{(1)} \vec{E}$ is the linear term and $\overleftrightarrow{\chi}^{(2)} \vec{E}^2$, $\overleftrightarrow{\chi}^{(3)} \vec{E}^3$ are non-linear terms. It can be seen from the equation that SHG is a $\overleftrightarrow{\chi}^{(2)}$ process, and CARS and TPEF are $\overleftrightarrow{\chi}^{(3)}$ processes. The three signals, CARS, TPEF, and SHG, can be simultaneously excited and detected by spectral filtering [6]. Thus, it is possible to combine these three techniques in a single multimodal imaging platform. An example of a non-linear multimodal image is given in figure 2.2B. Here CARS forms the red channel, TPEF forms the green channel, and SHG forms the blue channel of the RGB image. Each of the CARS, TPEF, and SHG signals provides morpho-functional information about a specific biomolecule. Hence, the combination of the three techniques will not only provide structural information about the tissue but also assure the visualization of various biomolecules. This is the key advantage of NLM imaging. Additionally, other properties of NLM imaging beneficial for tissue imaging are discussed further.

First, these techniques can be used in label-free manner; therefore, no staining is re-

quired. The label-free nature of non-linear optical techniques makes these techniques non-toxic and capable of *in vivo* measurements. Second, the non-linear interaction allows excitation of molecules in a confined volume around the focal point, thus increasing the spatial resolution of NLM images. Third, these techniques can employ lasers with longer excitation wavelengths, which reduces scattering and provides large depth penetration (as large as 1 mm for TPEF autofluorescence) in biological tissues [38, 39]. Due to these advantages, the NLM technique can serve as an alternative technique to conventional histological, immunohistochemical, and radiological imaging techniques. Few applications of NLM imaging are reported in chapter 4.

So far, the mathematical description and properties of non-linear optical processes have been discussed. Now each of the modalities CARS, TPEF, and SHG will be discussed in detail along with their biomedical applications.

2.2.1 COHERENT ANTI-STOKES RAMAN SCATTERING

Coherent anti-Stokes Raman scattering (CARS) is one of the three modalities of NLM imaging. CARS is a third-order non-linear process which includes a pump photon at frequency ω_p (first green arrow in figure 2.2A) and a Stokes photon at frequency ω_s (red arrow in figure 2.2A). When the frequency difference $\omega_p - \omega_s$ matches the vibrational frequency of a Raman active molecule ω_m (black arrow in figure 2.2A), the molecules begin to vibrate in phase coherently. As the pump beam is tunable, the frequency difference $\omega_p - \omega_s$ can be specifically adjusted to the desired vibrational energy of the relevant molecule. The molecular vibration subsequently inelastically scatters the photons from the pump beam to generate coherent anti-Stokes photons with frequency $\omega_{as} = 2\omega_p - \omega_s$ [40]. In simpler terms, the pump beam raises the electronic system of the molecule into a second virtual state of the energy $\hbar\omega_p + \hbar\omega_m$ (second green arrow in figure 2.2A). From here, the molecule is allowed to relax back to the ground state while the anti-Stokes photons with energy $\hbar\omega_{as}$ are detected and used for imaging (cyan arrow in figure 2.2A).

The first CARS microscopy was developed in 1982 [41], but its more extensive use began after 1999 with numerous applications reported in the life sciences [42, 43, 44]. CARS microscopy can be applied to visualize any molecular vibrations; however, it is usually performed to image the C-H-stretching vibration of methylene groups at 2850 cm^{-1} , which is abundant in lipids. The C-H-stretching resonances are the strongest within the CARS spectrum and enable the highest imaging speeds, i.e. up to video frame rate [6]. CARS imaging was first demonstrated *in vivo* on the skin of a mouse where CH_2 vibrational stretching was tuned to visualize abundant lipid structures in a mouse's ear [45]. CARS imaging can be sensitive to metabolic changes in infected or diseased human tissue [46, 47]. CARS can also provide different views of cellular structures in humans [48, 49], which makes it capable of medical imaging.

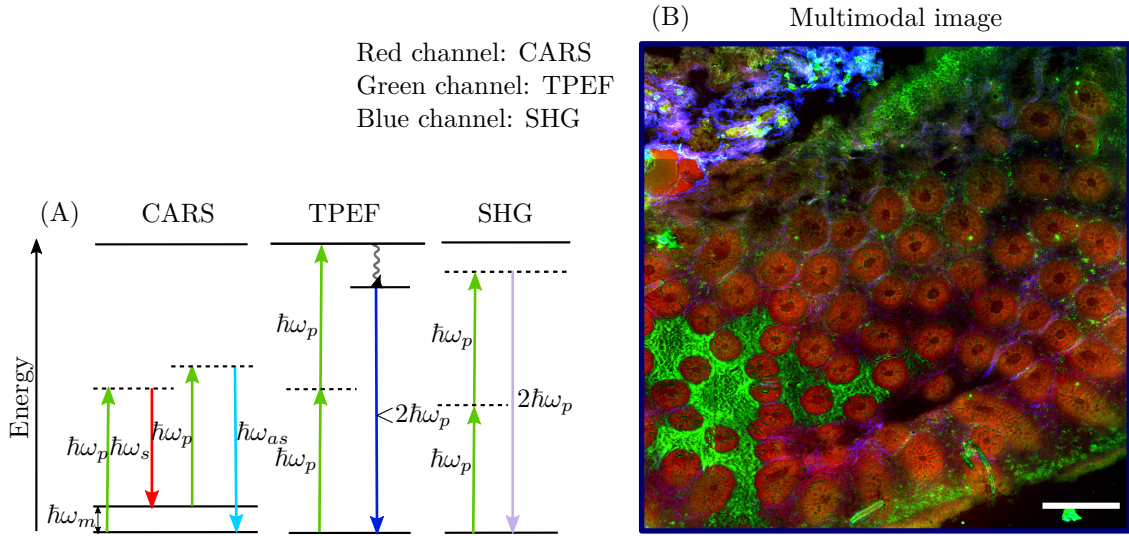


Figure 2.2: In this figure (A) shows an energy diagram for coherent anti-Stokes Raman scattering (CARS), two-photon excitation fluorescence (TPEF), and second harmonic generation (SHG). The explanation of the three modalities is provided in the text. (B) shows a non-linear multimodal image with the three non-linear optical modalities represented in the form of an RGB image. Here the CARS, TPEF, and SHG signals form the red, green, and blue channel of the RGB image, respectively. The scale bar represents 200 μm .

2.2.2 TWO-PHOTON EXCITATION FLUORESCENCE MICROSCOPY

Two-photon excitation fluorescence (TPEF) microscopy is the second non-linear optical modality of the NLM imaging introduced in this thesis. In TPEF microscopy, an autofluorophore is excited by two photons of half the resonance frequency ω_p (two green arrows in figure 2.2A) to achieve an electronic molecular transition $\omega_{TPEF} < 2\omega_p$ (dark blue arrow in figure 2.2A). Due to the non-linearity, the fluorophores are excited only in the proximity of nominal focus; this eventually reduces out-of-focus light [6]. Thus, TPEF imaging assures visualization of tissues at higher depth and better axial resolution. Furthermore, photoinduced damage outside the focal plane, typically observed in conventional fluorescence microscopy, is significantly reduced [50]. These properties of TPEF microscopy makes it suitable for biomedical applications.

TPEF microscopy in a laser scanning microscope was first implemented in 1990 where chromosomes of live cultured pig kidney cells stained with a viable DNA stain were visualized [51]. Since then, a growing number of studies employing TPEF microscopy for tissue imaging have been reported [52, 53, 54]. TPEF microscopy is capable of imaging intrinsic autofluorophores such as nicotinamide adenine dinucleotide hydrogenase (NADH), flavin adenine dinucleotide (FAD), and structural proteins like elastin, keratin, and collagen within epithelial and connective tissue. Applications of TPEF microscopy in various *in vivo* studies have been reported for skin-related stud-

ies in humans [55, 56, 57, 58] and brain imaging in animals [59]. TPEF microscopy can be readily combined with SHG microscopy for *in vivo* studies, which will be discussed in the next section.

2.2.3 SECOND HARMONIC GENERATION MICROSCOPY

The last non-linear optical modality of NLM imaging is second harmonic generation (SHG) microscopy. Contrary to CARS microscopy, SHG is a coherent second-order non-linear process. In SHG microscopy, two photons in the near-infrared region of frequency ω_p (green arrows in figure 2.2A) interact simultaneously with a molecule to generate a photon of double the energy or frequency $\omega_{SHG} = 2\omega_p$ (violet arrow in figure 2.2A). Thus, this process is also called frequency doubling [50]. Due to symmetry reasons, SHG can occur only in bulk noncentrosymmetric structures, for instance some proteins in tissue with a quasi-crystalline structure [6]. SHG microscopy is capable of imaging biomolecules like collagen, muscle, and microtubules. Therefore, collagen-rich tissues like the cornea, tendons, skin, arteries can be efficiently investigated using SHG microscopy [8, 60]. Furthermore, an SHG signal can be an indication of disease progression or alteration of biomolecules in a tissue, which is commonly observed in cancer [61]. Lastly, a combination of SHG and TPEF microscopy has widespread biomedical applications [62], for example the investigation of cells and cellular membranes [63].

With the explanation of Raman spectroscopy and NLM imaging, the stage is set to explain the datasets involved in these two technologies and the need for chemometrics to analyze these datasets. Although this was briefly discussed in chapter 1, the next section provides a detailed explanation.

2.3 THE NEED OF CHEMOMETRICS FOR BIOPHOTONICS

So far, the technical details of Raman spectroscopy and NLM imaging have been discussed. Analyzing the data from Raman spectroscopy and NLM imaging is also essential for a better understanding of biological systems. However, the analysis of both datasets presents several challenges which will be emphasized in this section.

The challenges encountered while analyzing Raman spectroscopic data are a great start. First, Raman spectroscopic data comprises several Raman spectra, which is a plot of the intensity of Stokes Raman scattered light against the wavenumber. A manual interpretation of the intensity and wavenumbers of each Raman spectrum can be tedious and time-consuming. Consequently, manual interpretation causes discrepancies among experts, especially for biological data, due to subtle spectral signatures that are difficult to observe. Second, a Raman spectrum inherently comprises noise due to the measurement setup and external interference. Unwanted noise like fluorescence background, Gaussian noise, CCD background noise, and cosmic spikes affects

the Raman spectrum. Third, Raman spectra of samples prepared at different time points can be significantly different from each other. Likewise, Raman spectroscopic data acquired from biological samples manifest variance within or between experimental and biological replicates. Thus, multivariate statistical methods to remove unwanted noise or other corrupting effects; to perform calibration to remove the spectral contributions dependent on the measurement system; and to extract significant spectral signatures from the high dimensional Raman spectra, are required. In this scenario, a field like “chemometrics” comes into the picture.

Like Raman spectroscopy, NLM imaging presents specific challenges. First, NLM imaging, unlike histopathology, is an untargeted technique. Due to its untargeted nature, specific tissue structures that are highlighted using conventional imaging techniques are not visualized in NLM images. Second, the image contrast associated with a specific biomolecule is unclear. For example, an SHG signal does not show any spectral difference coming from different biomolecules like collagen, tubulin, and cholesterol [6]. This affects the interpretation of non-linear multimodal images to the untrained eye. Third, the extraction of complicated patterns or tissue structures requires annotation by an expert. Finally, noise artefacts visible due to optical systems and mosaic artefacts resulting from stitching of tile scans needs to be removed [64]. Therefore, research on advanced image analysis and pattern recognition methods, which are parts of chemometrics, is crucial.

As a solution to the above-mentioned challenges, the use of chemometric approaches is highly recommended. Chemometrics is an interdisciplinary field of science capable of extracting information from chemical systems using data-driven strategies [65]. This multidisciplinary field encompasses various aspects like multivariate statistics, pattern recognition, image and data analysis, and statistical modelling. Mostly, chemometric approaches for Raman spectroscopic data are used to reduce spikes arising from unwanted noise, correct baseline shifts resulting from spurious background signal or instrument fluctuation, discover different components present in a mixture, and build predictive or descriptive models to analyze the underlying spectra [66, 10]. Similarly, chemometric approaches for non-linear multimodal images are used to filter unwanted noise, correct mosaicking artefacts caused by uneven illumination [67], delineate foreground and different regions in the images, and construct predictive models to identify biomolecular changes [68]. As the complexity of these tasks increases, chemometric methods require complicated algorithms for extracting meaningful patterns from the data. Such complicated algorithms fall under the category of artificial intelligence (AI). Taking advantage of intelligent AI methods, this thesis proposes the use of AI for the reliable analysis of biophotonic data like Raman spectroscopic and non-linear multimodal imaging data. This will be discussed in the next chapters.

3

Artificial intelligence for biophotonics

The previous chapter introduced “chemometrics” as a solution for maximizing the information content from biophotonic data, thereby assuring a better design of experiments and allowing a deeper understanding of the biological system. Chemometric methods like signal and image processing, statistical modelling, descriptive and inference statistics, and pattern recognition are essential for maximizing the information content of Raman spectroscopic and NLM imaging data. The previous section also established grounds for AI-based technologies that are a recent trend in chemometrics. Utilizing AI-based technologies in chemometrics has the following advantages. First, AI-based technologies can provide better knowledge acquisition systems that complement the expert’s knowledge. Second, it can extract patterns from data which are useful for predictive modelling without any manual intervention. Third, AI-based technologies can provide better generalization of unseen data and process incomplete data [69]. Due to these advantages, this thesis uses AI-based technologies for analyzing biophotonic data like Raman spectroscopic and NLM imaging data.

For the smooth understanding of the AI-based technologies used in this thesis, this chapter discusses the fundamentals of AI along with the two major sub-fields of AI – machine learning and deep learning – and its application for data analysis. This chapter will also provide a brief introduction of different models used in machine learning and deep learning.

3.1 FUNDAMENTALS OF AI AND ITS APPLICATIONS

Artificial Intelligence is (mostly) a field of computer science that aims to create intelligent systems which are capable of reasoning, planning, perceiving, processing, solving problems, and learning from experience. AI can be categorized into strong and weak AI [70]. Strong AI, also known as artificial general intelligence, describes programs that can simulate human intelligence and their cognitive abilities. On the other hand, weak AI (a.k.a. narrow AI) are programs designed to perform a specific task [70]. Most researchers today utilize weak AI, for instance Apple’s Siri and Amazon’s Alexa. The weak AI programs form the basis of the most widely used machine learning algorithms which are discussed further.

Machine Learning (ML) (see section 3.2) is a form of AI that generates algorithms which can iteratively learn from the data and make decisions to perform a given task. In other words, an ML algorithm converts an experience achieved from the data into an expertise that is evaluated using a performance metric. The performance of an ML algorithm is heavily dependent on the data used to gain experience. Therefore, it is crucial to provide ML algorithms with concise and meaningful data. The meaningful representation of data is called “features” or “patterns”, and the process of extracting these features is called “feature extraction”. The field of AI is closely related to another broad field used for extracting features or patterns, called “pattern recognition”. Pattern recognition represents the whole dataset in the form of useful patterns or features which can be used as an input to the ML algorithm. It can be considered one of the pre-processing steps while performing data analysis [71]. However, recognizing patterns or features to feed a machine learning algorithm requires much manual effort and is tedious. Therefore, sophisticated learning algorithms that are capable of “feature engineering” are crucial. The “feature engineering” algorithms can automatically extract and learn abstract features like an experienced professional. Such algorithms are also referred to as self-learning algorithms and are a subset of ML algorithms called “deep learning” (DL) algorithms (see section 3.3).

A DL algorithm represents raw data, for instance an image, as a combination of simple patterns like color, edges, and contours. These simple patterns are extracted by a series of layers of a DL model, in contrast to ML models where these patterns are manually extracted. A series of input, hidden, and output layers provide depth to the model and are one of the peculiar characteristics of DL algorithms. ML and DL have numerous applications in image classification, object detection and localization, semantic segmentation, speech recognition, and natural language processing.

To summarize, AI, machine learning, and deep learning are related to each other and can be visualized in the form of concentric circles (see figure 3.1). The idea of AI was the first to become popular, later blossomed machine learning, and “*today’s AI*” is deep learning. The following sections elaborate on the fundamentals of machine learning and deep learning.

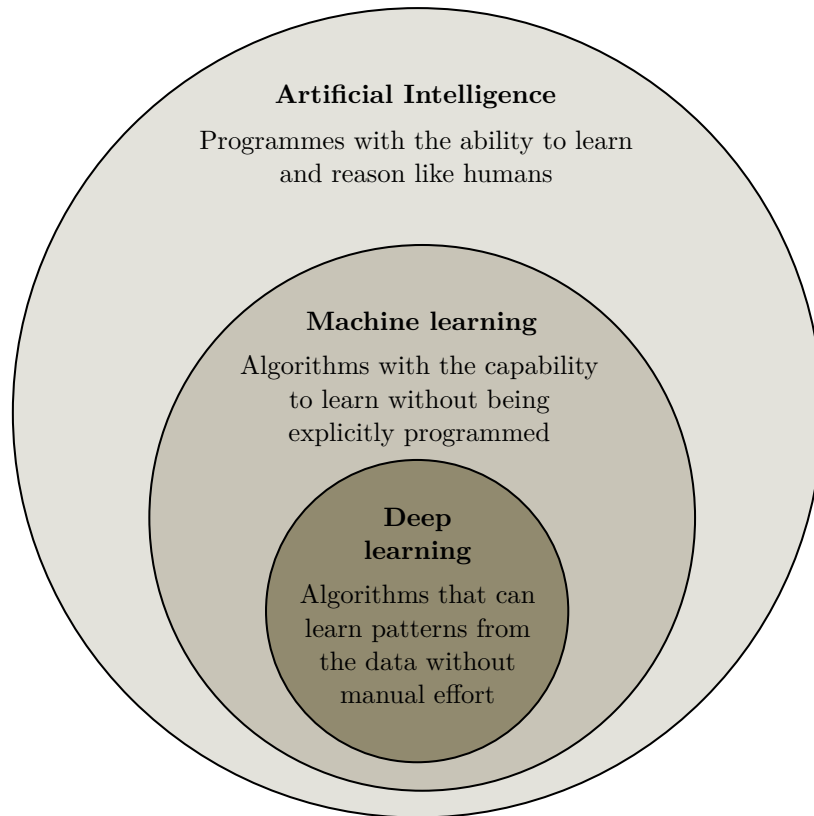


Figure 3.1: This figure shows artificial intelligence (AI), machine learning (ML), and deep learning (DL) in the form of three concentric circles. Deep learning is a subset of machine learning, which is altogether a subset of AI. The figure is modified from reference [72].

3.2 BASICS OF MACHINE LEARNING

In the previous section, it was mentioned that ML is a subset of AI and most researchers today utilize ML programs for performing specific tasks. The definition of ML quoted by Tom M. Mitchell says that: “An ML program is a computer program which is said to learn from experience E concerning some class of tasks T and performance measure P , if its performance at tasks T as measured by P , improves with experience E ” [73]. Briefly, a task T is the type of prediction or inference achieved based on a defined problem, and the process of training on a task T is called learning. An example of a task T is the classification of different types of bacteria using spectroscopic data. In this example, the experience E is the spectroscopic dataset on which the ML program learns to perform the classification task T . Lastly, a performance measure P is a metric that evaluates the accuracy of the ML model. The concepts including dataset, types of learning, performance metric required to construct an ML program are explained in the following sections.

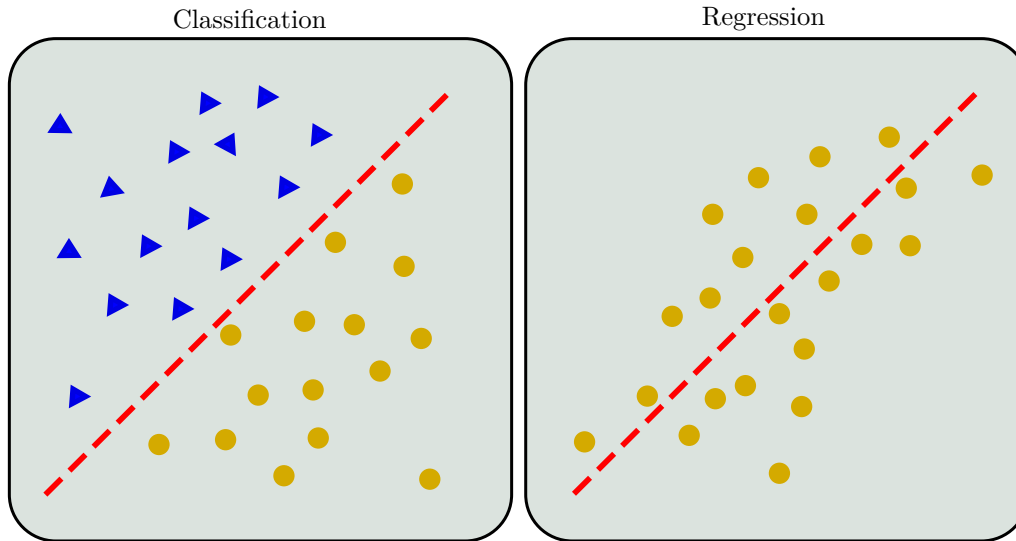


Figure 3.2: This figure shows the difference between two types of supervised learning algorithms, namely: classification and regression. The objective of classification is to assign each data point (blue triangles and yellow circles) into discrete categories or classes, whereas the objective of regression is to find a relationship among the data points (yellow circles) based on their input features.

3.2.1 DATASET

An experience E is a dataset represented by the extracted features or the raw data itself. For instance, pixel values are raw data which represents an image; likewise, color, texture, and shape patterns extracted from that image are features. Mathematically, a dataset can be described as a matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$ with M observations or samples in the form of N -dimensional feature vector $\mathbf{x} = [x_1, x_2, \dots, x_N]^T \in \mathbb{R}^N$. Each observation m is associated with a target value y_m (a.k.a. class or label). For the whole dataset \mathbf{X} , a target vector can be given as $\mathbf{y} = [y_1, y_2, \dots, y_M] \in \mathbb{R}^M$. The target information is mostly given by an expert who helps the ML algorithm to perform a task. The dataset on which the ML algorithm learns to perform a task is called “training dataset”, and the dataset on which the ML algorithm evaluates its performance is called “testing dataset” [74]. In the case of DL algorithms, another set of the dataset is introduced, which is called the “validation dataset”. The validation dataset optimizes many hyperparameters of the DL model [75]. Based on the dataset and the task, an ML algorithm is divided into categories, namely supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning [76] and representation learning [77]. The former two learning categories are discussed in the following sections.

3.2.2 SUPERVISED LEARNING

A supervised learning algorithm learns through a labelled training dataset $\mathcal{D} = ((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_M, y_M))$ which is a finite pair of M observations with each feature vector \mathbf{x}_m associated with its target value y_m , where m is the index for M observations [74]. A supervised learning algorithm learns to predict y from \mathbf{x} by estimating the distribution $p(y|\mathbf{x})$. An example of supervised learning is the prediction of a disease based on the target information given by an expert. Based on the target, a supervised learning algorithm can be divided into two categories: classification and regression, as discussed below.

CLASSIFICATION

Classification is a form of supervised learning where the target variables are discrete or categorical (e.g., $y_m \in \{0, 1\}$ or $y_m \in \{tumor, healthy\}$) [74]. A classification learning algorithm $\mathcal{F} : \mathbb{R}^N \rightarrow \{1, \dots, C\}$ learns to assign each feature vector \mathbf{x}_m to C categories identified by y_m . The function \mathcal{F} can also provide the probability distribution over C categories. An example of a classification task is the identification of normal and tumor tissue or characterization of different disease stages. A few examples of classification are given in chapter 4. These applications utilize commonly used classification algorithms like support vector machine (SVM) and linear discriminant analysis (LDA). Therefore, these two algorithms are discussed below.

Linear discriminant analysis also called Fisher's linear discriminant analysis [78], is a supervised learning algorithm mainly used for classification purposes. The LDA algorithm projects a multi-dimensional dataset onto "discriminant axes". The discriminant axes maximize the ratio of inter-class to intra-class scatter to optimally classify the dataset into two or more classes. The objective of the LDA algorithm $E(\mathbf{W})$ is to maximize the ratio of inter-class to intra-class scatter by finding optimal discriminant axes \mathbf{W} , and is given as

$$E(\mathbf{W}) = \frac{|\mathbf{W}^T \mathbf{S}_b \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_w \mathbf{W}|}, \quad (3.1)$$

where $\mathbf{W} = [w_1|w_2|\dots|w_L]$ with L projections, $\mathbf{S}_b = \sum_i^C (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$ is inter-class scatter matrix, $\boldsymbol{\mu}_i$ as mean of observations in the class i and $\boldsymbol{\mu}$ as mean of all the $\boldsymbol{\mu}_i$, $\mathbf{S}_w = \sum_i^C \mathbf{S}_i$ represents the intra-class scatter matrix with $\mathbf{S}_i = \sum_{j=1}^{M_i} (\mathbf{S}_j - \boldsymbol{\mu}_i)(\mathbf{S}_j - \boldsymbol{\mu}_i)^T$. Here, M_i is the total number of observations or samples in the i^{th} class and \mathbf{S}_j is one such observation [79].

Support vector machine is another supervised learning algorithm used for classification tasks [80]. An SVM algorithm constructs a hyperplane or set of hyperplanes to separate an N -dimensional dataset into different groups or classes. The construction of the hyperplane is based on the "maximum-margin hyperplane" theorem [81].

According to this theorem, a hyperplane is selected to divide M data points such that the distance between the hyperplane and the nearest data point \mathbf{x}_m from either of the groups is maximized. This builds a simple linear classification model. However, based on the complexity of the dataset, a non-linear classification model with a non-linear hyperplane in multidimensional space can be constructed [81].

REGRESSION

Regression is another form of a supervised learning algorithm that learns a function $\mathcal{F} : \mathbb{R}^N \rightarrow \mathbb{R}$ to assign a numeric value to the N -dimensional input data. Unlike the classification learning algorithm, the target variables in regression are continuous. Figure 3.2 shows the difference between classification and regression learning algorithms. An example of regression is predicting the content of iron in an ore from mass spectroscopy measurements [74]. The simplest regression algorithm is the linear regression algorithm, which is explained further.

Linear regression is a linear model which assumes a linear relationship between the N -dimensional input variable \mathbf{x}_m and the output variable y_m [82]. Specifically, this is a multiple or multivariable linear regression model, as the output function is dependent on N predictor variables [82]. Mathematically, the output variable y_m is a linear combination of the N -dimensional vector \mathbf{x}_m , and can be represented as

$$y_m = \beta_0 + \beta_1 x_{m1} + \cdots + \beta_N x_{mN} + \epsilon, \quad (3.2)$$

where $\{\beta_0, \beta_1, \dots, \beta_N\}$ are the regression coefficients and ϵ is the residual term of the linear regression model. A simple linear regression model with just one-dimensional input can be represented by a line, $y_m = \beta_0 + \beta_1 x_{m1}$ (see figure 3.2). With increasing dimensions of the input variable like in equation 3.2, the model can be represented by a hyperplane. Clearly, the linear regression model is used to predict a continuous dependent variable based on predictor (or independent) variables. Likewise, when dependent variables are categorical, the logistic regression model is preferred [82].

The explanations above provide a clear picture of supervised learning algorithms. The next section will focus on unsupervised learning algorithms including clustering and dimension reduction, which are extensively used in this thesis.

3.2.3 UNSUPERVISED LEARNING

Unlike supervised learning algorithms, unsupervised learning algorithms learn through an unlabelled dataset, i.e. the observations are not associated with its targets. These algorithms attempt to unravel specific properties of the underlying dataset [82]. In other words, these algorithms can separate M observations into K groups (a.k.a. clusters) based on its similarities known as “clustering algorithms”. Unsupervised learning algorithms can be used to project a high dimensional feature vector \mathbf{x} to

low dimensional feature vector \mathbf{z} . This is known as “dimension reduction algorithms”. Lastly, unsupervised learning algorithms can learn the probability distribution $p(\mathbf{x})$ of the underlying dataset. This is known as “density estimation algorithms” [71]. The “clustering” and “dimensionality reduction” algorithms are discussed below.

CLUSTERING ALGORITHM

Clustering is an unsupervised learning algorithm used to partition the observations into clusters such that each observation in a cluster is similar to observations from its cluster as compared to the observations in other clusters. The similarity between the observations is calculated using a distance metric (e.g. Euclidean distance) and the clustering algorithm aims to reduce this metric [82]. For instance, an average Euclidean distance can be calculated as a distance metric between the centroid of a cluster and other observations of that cluster. This forms the basis of a prototypical clustering algorithm called the K-means clustering algorithm.

K-means clustering algorithm is most widely used in biophotonic data analysis, especially Raman spectroscopic data. In this method, each observation is labelled as K clusters by calculating (mostly) the Euclidean distance between each observation and the center of each cluster. An observation nearest to the center of a cluster is categorized into a new label of that cluster. The optimization of the clusters is achieved by calculating the cluster centers based on previous and current cluster centers, and iteratively updating the centers until convergence is achieved. Mathematically, a K-means clustering algorithm can be given as, $\arg \min_C \sum_{i=1}^K \sum_{\mathbf{x} \in c_i} \|\mathbf{x} - \mu_i\|^2$, where K is the number of clusters and \mathbf{x} is an observation belonging to the clusters $\{c_1, c_2, \dots, c_K\}$ with centers $\{\mu_1, \mu_2, \dots, \mu_K\}$.

DIMENSION REDUCTION

Dimension reduction is an unsupervised learning algorithm which transforms a high dimensional input dataset into a low dimensional representation. Mathematically, it is given as $\mathcal{F} : \mathbb{R}^N \rightarrow \mathbb{R}^D$, where $D \ll N$. Dimension reduction can be achieved in two different ways. The first way is by retaining the most relevant features from the original set of features; this is known as “feature selection”. The second way is by removing the redundancy of the original feature set and finding a completely different set of relevant features; this is known as “feature extraction” [71]. A commonly used dimension reduction technique based on feature extraction is called principal component analysis (PCA) which is explained below.

Principal component analysis is an unsupervised data transformation technique that projects a high dimensional data matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$ to low dimensional data $\mathbf{Z} \in \mathbb{R}^{M \times D}$. Specifically, the high dimensional feature set (e.g., wavenumbers from the

Raman spectrum or texture features from a non-linear multimodal image) is transformed to a smaller set of orthogonal principal components (or PC's) in the direction of maximum variation. The PC's are obtained by calculating the Eigen values and Eigen vectors obtained from the covariance matrix of the data matrix \mathbf{X} . The highest Eigen value corresponds to the first PC (PC1) and demonstrates the maximum variance in the dataset. Similarly, the second-highest Eigen value corresponds to the second PC (PC2) which represents the largest residual variance and is orthogonal to PC1; therefore, it is uncorrelated and independent [82]. Mathematically, a PCA model to decompose original matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$ can be given as,

$$\mathbf{X} = \mathbf{Z}\mathbf{U}^T + \mathbf{E}, \quad (3.3)$$

where $\mathbf{Z} \in \mathbb{R}^{M \times D}$ is the scores matrix, $\mathbf{U} \in \mathbb{R}^{N \times D}$ is the loadings matrix and \mathbf{E} is a residual matrix. The residual matrix is a measure of lack-of-fit of the PCA model, and a smaller value of \mathbf{E} corresponds to a good PCA model. The scores and the loadings matrices are a characteristic of the M observations and N variables or features, respectively. The combination of PCA and LDA is a useful approach for classification purposes and will be utilized in chapter 4.

3.2.4 CROSS-VALIDATION AND PERFORMANCE MEASURE

In the previous sections, various concepts like dataset and supervised and unsupervised learning algorithms were introduced. This section will introduce the evaluation of these learning algorithms based on a scheme called “cross-validation” and performance metric.

The first concept is cross-validation which is important to evaluate ML algorithms and construct robust models for classification, regression and feature extraction tasks. The simplest cross-validation approach is to divide the dataset into a training dataset, which is used to train the ML model and a testing dataset which is used to evaluate the performance of the ML model. This is called “hold-out” cross-validation [82]. However, this approach causes poor estimation of actual model performance and lacks generalization. Thus, a k -fold cross-validation approach is recommended. A k -fold cross-validation aims to divide the dataset into k parts and train the model for k rounds using a $1/k^{th}$ part as the test dataset in each round. The average test score for the k rounds is used as the performance measure of the model [82].

The next important concept is the performance measure or metric P that interprets the goodness of the ML model. The most used performance metric for classification purposes is a confusion matrix which is a table indicating the number of false positives (FP), true positives (TP), false negatives (FN) and true negatives (TN). Based on the confusion matrix, various performance measures including, but not limited to, accuracy = $(TP+TN)/(TP+TN+FP+FN)$, sensitivity = $TP/(TP+FN)$, specificity = $TN/(TN+FP)$ can be derived. These are the most reported metrics in literature.

Likewise, the mean squared error (MSE) and root mean squared error (RMSE) are commonly used for regression purpose. A machine learning model is optimized to either maximize the accuracy or minimize the RMSE based on the task.

So far, the previous sections have emphasized machine learning algorithms and related concepts. The next section will introduce challenges encountered by traditional machine learning algorithms and the motivation behind using deep learning algorithms.

3.2.5 CHALLENGES MOTIVATING DEEP LEARNING

The conventional ML algorithms discussed above are limited in several ways. Foremost, difficulties arise with ML models while performing tasks like the semantic segmentation (see chapter 4.2) or pseudo-staining (see chapter 4.3) of images, for instance, NLM images with a specific molecular contrast. The shortcomings of conventional ML algorithms motivate the development of “intelligent” algorithms like deep learning algorithms. These algorithms can efficiently learn simple to complex patterns from the images or spectra that could be used to construct reliable models. In this prospect, this section discusses some challenges encountered by ML models.

Foremost, ML models have limitations while learning complicated functions for high dimensional feature space. A high dimensional feature space is a problem when the number of features is much higher than the number of observations ($N \gg M$). This is termed the “curse of dimensionality” [75]. The curse of dimensionality is commonly observed in Raman spectroscopic data, in which the number of wavenumbers is higher than the number of observations in each class. This is one of the reasons to perform preprocessing and feature extraction for Raman spectroscopic data before constructing a conventional ML model. Nevertheless, DL models can alleviate this “curse” due to its inherent nature to ascertain a pattern or extract abstract features from high dimensional data.

Additionally, conventional ML algorithms rely on implicit “priors” like smoothness and local constancy [75]. These priors state that a function learned by an ML algorithm does not change largely within a small region ($\mathcal{F}^*(\mathbf{X}) = \mathcal{F}^*(\mathbf{X} + \epsilon)$). Due to this assumption, test points or unseen dataset return results “near to” similar points in the training dataset. Therefore, conventional ML algorithms like k-nearest neighbors and decision trees that exclusively rely on this assumption tend to generalize on the unseen data poorly [75]. However, DL models include task-specific assumptions that help to generalize the underlying data better. One such assumption is that the underlying data is generated using the composition of features that potentially describe multiple levels of hierarchy in the data. This can be a reason to name these models as “deep” learning models [75].

In prospect, DL has efficiently dealt with a high dimensional dataset in the past. Thus, this thesis emphasizes using DL for biophotonic data analysis. The next section

will provide the fundamentals of DL, and various traditional and advanced DL models.

3.3 BASICS OF DEEP LEARNING

Deep learning is a field of machine learning with advanced capabilities for data analysis (see figure 3.1). A deep learning model emulates the biological structure of the human brain neuron [75]. Briefly, a deep learning model is a computational model that maps input to output by learning patterns in the data through a series of hidden layers. The input of the deep learning model is a vector of features extracted from the data or the raw data itself, and the output of a deep learning model is a single output or vector of values or probabilities. The hidden layer comprises neurons or nodes. A neuron is a “computational unit” with one or more weighted inputs, a transfer function to combine the inputs in a linear or non-linear manner, and lastly, one or more outputs. The neurons of a layer are connected to the other neurons of the same, previous, and consecutive hidden layers. The interconnection of neurons facilitates the deep learning model to learn complex abstractions of the data. Learning complex abstractions of the data is achieved by passing the information of each neuron in the forward direction (a.k.a. *forward pass*) and subsequently optimizing the weights of the inputs of each neuron in the backward direction (a.k.a. *backward pass* or back-propagation) [75]. Due to the property of deep learning models to learn complex features from the data, it has gained overwhelming popularity in image and spectral data analysis. Thus, this thesis introduces various deep learning models to perform analysis on the Raman spectroscopic data and NLM imaging data. Before introducing various deep learning models, it is important to discuss the biological motivation of the most basic type of DL model, which is called artificial neural network.

3.3.1 ARTIFICIAL NEURAL NETWORKS: A BIOLOGICAL MOTIVATION

An artificial neural network (ANN) is a basic type of deep learning model. ANN is loosely inspired by the neural network of the human brain [83]. The primary motivation of ANN is to recreate learning and predictive capabilities, as well as cognitive abilities of the human brain through its fundamental unit called the “neuron”. In biological terms, a neuron comprises a cell body with an axon and many dendrites. From a functional perspective, a neuron processes information received by the dendrites in the form of an electrical signal and propagates it further to other neurons through the axon [84]. Thus, ANNs aim to emulate the process of the neural network of the human brain.

As explained above, an ANN is a computational model that receives an input $\mathbf{x} \in \mathbb{R}^N$ from the preceding neuron and synthesizes it multiplicatively (w_0x_0) with the synaptic strength of the dendrites of that neuron (see figure 3.3). The synaptic strength is the idea behind the weights $\mathbf{w} \in \mathbb{R}^N$ in ANNs, which are learnable and

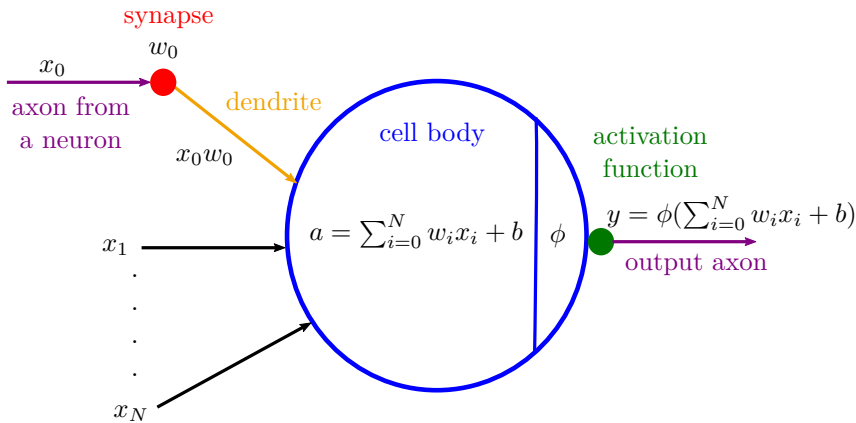


Figure 3.3: This figure shows a computational model of a neuron. The input x_0 arriving from the axon of a previous neuron is multiplied with the synaptic strength w_0 of the current neuron. The signal $w_0 x_0$ is carried by the dendrite to the cell body of the neuron. Many such input signals are carried by the dendrites, which are subsequently added and processed by a linear or non-linear activation function ϕ of the cell body. The non-linear output of this neuron is carried by its axon to the following neuron.

are influenced by the data received from other neurons. In addition to the weights, each neuron has a bias $b \in \mathbb{R}$. The weighted inputs and the bias of each neuron are linearly combined to give output unit activation, $a = \sum_{i=0}^N w_i x_i + b$. The output activation is transformed using a differentiable and non-linear activation function ϕ , such that the final output signal from a neuron is $y = \phi(a) = \phi(\sum_{i=0}^N w_i x_i + b)$. The output signal y is propagated in the forward direction to the following neurons [71]. During training of a DL model, the weights and biases of each neuron, also referred to as “hyperparameters” of the DL model, are optimized. A detailed explanation on optimization of weights will be given in section 3.3.3.

The computational model of the neuron explained so far forms the basis of traditional multilayer perceptrons and other neural network architectures. These architectures will be explained in the next section.

3.3.2 NEURAL NETWORK ARCHITECTURES

Feed-Forward neural network or **multilayer perceptron (MLP)** is also sometimes referred to as an artificial neural network [83]. The earlier section introduced the computational model of a single neuron used in ANNs, and in this section these neurons will be used in layers of MLPs. An MLP network comprises input, output and hidden layers. The MLP network passes the input in the forward direction through a series of L hidden layers. The series of hidden layers not only provide depth to these networks but also generate complex representations of the input. A

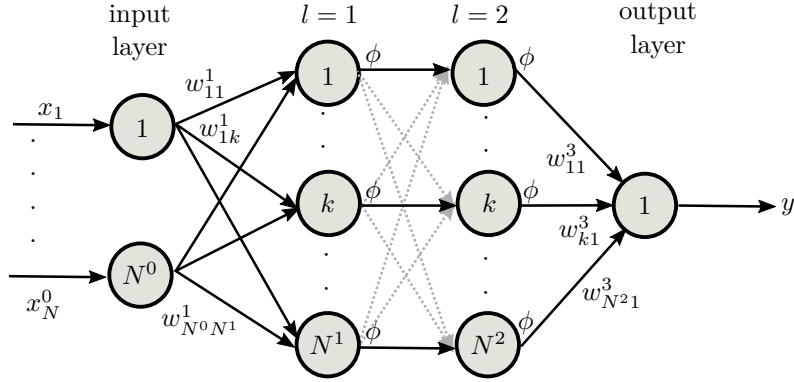


Figure 3.4: This figure shows a multilayer perceptron with one input layer (two neurons), two hidden layers (three neurons) and one output layer (one neuron). The connections between neuron i of layer $l - 1$ and neuron k of layer l is weighted by w_{ik}^l . Each neuron transforms the output using an activation function ϕ . The final output y is a combination of transformed outputs from the intermediate layers.

graphical representation of MLPs with two hidden layers ($L = 2$) is shown in figure 3.4. Mathematically, an output from an arbitrary neuron y_k^l , indexed by $k \in N$ neurons in each layer $l \in \{0, \dots, L\}$ is given by

$$y_k^l = \phi \left(\sum_{i=0}^{N^{l-1}} w_{ik}^l y_i^{l-1} + b_k^l \right), \quad (3.4)$$

where y_i^{l-1} is an output of neuron i in layer $l - 1$, w_{ik}^l is the weight of the connection between neuron i in layer $l - 1$ and neuron k in layer l , b_k^l is the bias value of layer l and ϕ is the non-linear activation function which is assumed to be the same for all layers. In this way, the final output of the MLP network, $y = \mathcal{F}(\mathbf{x}; \Theta)$ is a composition of linear combinations of non-linear outputs from each layer, where $\Theta = \{\mathcal{W}, \mathcal{B}\}$ is the hyperparameter set of weights \mathcal{W} and biases \mathcal{B} .

The non-linearity in the deep neural networks can be introduced with various non-linear activation functions ϕ including sigmoid function $\phi(a) = \frac{1}{1+e^{-a}}$, rectified linear units $\phi(a) = \max(a, 0)$, and the hyperbolic tangent $\phi(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}$. One more type of non-linear activation function commonly used for performing classification task is called the softmax activation function. The softmax function maps the output from the last layer into the probability distribution of classes $c \in C$. Mathematically, a softmax function can be given as

$$P(y_k^l | x_i; \Theta) = \frac{e^{(w_i^l)^T x_i + b_i^l}}{\sum_{c=1}^C e^{(w_c^l)^T x_c + b_c^l}}, \quad (3.5)$$

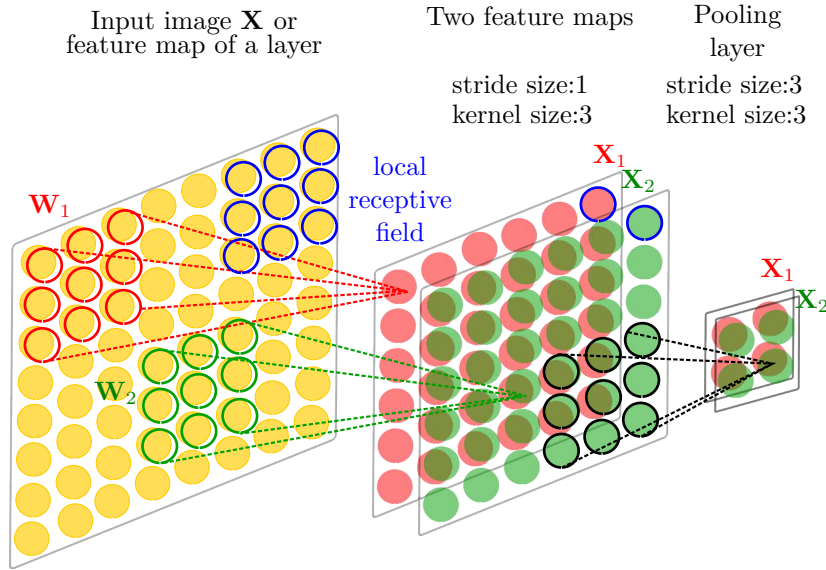


Figure 3.5: This figure shows a convolutional neural network architecture. The input image \mathbf{X} is convolved with two kernels \mathbf{W}_1 and \mathbf{W}_2 to obtain a feature map \mathbf{X}_1 and \mathbf{X}_2 . The pooling layer reduces the spatial dimension of the feature map. The receptive field is shown in blue. This figure is inspired from reference [24].

which satisfies $0 \leq y_k^l \leq 1$ and $\sum_c y_k^l = 1$. The softmax activation function is frequently used as the last layer of the deep neural networks [71].

Up to this point, the architecture of MLPs has been explained. Before introducing architectures of other deep neural networks, it is important to mention a challenge encountered by MLPs while dealing with large input data. As the dimension of input data increases, the number of weights in the first layer of the MLP model also increases. This problem is seen while working with large images. Clearly, the increase in the number of (trainable) weights can cause overfit of the MLP model to the training dataset. Thus, deep learning models like convolutional neural networks to reduce the amount of trainable weights are crucial.

Convolutional neural network (CNN) is similar to a multilayer perceptron with the additional advantage that it works with grid data like a spectrum or an image [85, 86]. The architecture of CNNs was inspired by feed-forward information processing in the early visual cortex of animals [87]. As the name suggests, CNNs employ a mathematical operation called a convolution ($*$) in one of the layers instead of conventional matrix multiplication used in MLPs. Convolution layers exploit the spatial or spectral correlation of input by using shift-invariant trainable kernels or weights. Mathematically, the output of the l^{th} layer, \mathbf{X}_k^l with $k \in N$ neurons is given by

$$\mathbf{X}_k^l = \phi(\mathbf{W}_k^l * \mathbf{X}_k^{l-1} + b_k^l), \quad (3.6)$$

where \mathbf{W}_k^l and b_k^l represent the kernel or weight and bias of neuron k in the l layer, respectively. Each layer comprises a set of K trainable weights $\mathcal{W} = \{\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_K\}$ and biases $\mathcal{B} = \{b_1, b_2, \dots, b_K\}$, which are the hyperparameters of the CNN model. These hyperparameters are optimized while training a CNN model to produce an output \mathbf{X}_k^l , which is also referred to as a “feature map”. In addition to the convolution layer, CNNs differ from conventional MLPs in three more aspects, namely weight sharing, pooling layers and receptive field [24]. The first aspect is weight sharing which reduces the number of trainable parameters by sharing weights of all neurons in a feature map. The second aspect is a pooling layer which reduces the spatial or spectral dimension of the input image or spectrum, thereby decreasing the number of weights in the network. The third aspect is a receptive field which is a hyperparameter that describes the spatial extent of a local region in the input that is affected by a kernel [75]. The figure 3.5 shows a CNN architecture along with a feature map, trainable weights, pooling layer and local receptive field. CNNs can be used in generative models like the autoencoders and generative adversarial networks. **Autoencoder** (AE) is a form of artificial neural network that can be employed in an unsupervised manner [88, 89]. Autoencoders comprise two networks; an encoder and a decoder. An encoder transforms an input $\mathbf{x} \in \mathbb{R}^N$ to a hidden state representation $\mathbf{h} \in \mathbb{R}^D$ and captures the most salient features of the input. The features in the hidden state are called “bottleneck” features and have (mostly) smaller dimensions than the input dimension. On the other hand, a decoder reconstructs the bottleneck features $\mathbf{z} \in \mathbb{R}^N$ with the same dimension as the original input. Traditionally, AEs were used for dimension reduction where the bottleneck features were used as features in the reduced dimension space [75]. Since then, the use of AEs has been investigated for image classification, semantic segmentation and image reconstruction [90]. One such application of AE is to segment non-linear multimodal images which will be discussed in section 4.2. AEs can also be a part of the generative adversarial networks.

Generative adversarial network (GAN) is a generative model comprising two artificial neural networks, namely a generator and a discriminator [91]. As the name suggests, the two networks are adversaries of each other such that the generator is trained against the discriminator to produce images from the original data distribution. Specifically, a generator network takes random numbers as the input and maps them into a visually pleasing image. Instead of using a set of random numbers as the input, a generator can also use an image. This is called image translation. In the image translation task, a generator transforms an image into another image which looks similar to the original image. Because of the ability to translate images, GAN has applications for image deblurring and image super-resolution [92, 93, 94, 95, 96]. The discriminator network is responsible for testing the plausibility of the generated images by differentiating them as ‘real’ or ‘fake’. The output of the discriminator network is a probability that the generated image is acquired from the original data

distribution. The mathematical explanation of AEs and GANs can be found in reference [24].

The different DL models that will be used in chapter 4 have already been discussed. Now, the next important aspect is training and optimizing the large number of hyperparameters of these DL models.

3.3.3 STATISTICAL LEARNING AND NON-LINEAR PARAMETER OPTIMIZATION

So far, it is clear that neural networks are non-linear parametric models which map an input vector \mathbf{x} to an output vector \mathbf{y} . Determining the hyperparameters $\Theta = \{\mathcal{W}, \mathcal{B}\}$ of all the layers in the networks, requires training these networks and optimizing their hyperparameters for minimizing the loss function $E(\Theta)$. Training a deep neural network is analogous to polynomial curve fitting, where the estimated output $\hat{y}(\mathbf{x}; \Theta)$ is close to the target output y , such that the loss function $E(\Theta)$ is minimal [71] (from this point Θ will be referred to as \mathbf{w} including the bias value in w_0^l). The loss function $E(\mathbf{w})$ is given by

$$E(\mathbf{w}) = \frac{1}{2} \sum_{m=1}^M \|\hat{y}(\mathbf{x}_m; \mathbf{w}) - y_m\|^2, \quad (3.7)$$

where $m = 1, \dots, M$, is a set of M labelled training dataset. This is sum-of-square error function which is common for a regression task with a sigmoid or linear activation function as an output layer [71]. However, for a classification task, a probabilistic interpretation of the neural network output is preferred. For this purpose, a dataset $\mathcal{D} = ((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_M, y_M))$ of M independent and identically distributed pairs of input and target labels is considered; the likelihood function is constructed as follows:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{m=1}^M p(y_m|\mathbf{x}_m, \mathbf{w}, \beta). \quad (3.8)$$

Taking the negative logarithm and solving this equation further, an error function can be given as

$$\frac{\beta}{2} \sum_{m=1}^M \{\mathcal{F}(\mathbf{x}_m; \mathbf{w}) - y_m\}^2 - \frac{M}{2} \ln \beta + \frac{M}{2} \ln(2\pi). \quad (3.9)$$

Here, β is the inverse variance of Gaussian noise, assuming a Gaussian distribution of the target labels. Maximizing the likelihood function (equation 3.8) is equivalent to minimizing the sum-of-squared error function (equation 3.7).

It is clear from the previous sections, that the activation function ϕ used in the neural network model introduces non-linearity to the model. Due to the non-linearity of the neural network model $\mathcal{F}(\mathbf{x}; \mathbf{w})$, the error function is a highly non-convex function in nature, as shown in figure 3.6. Due to the non-convexity, converging or finding

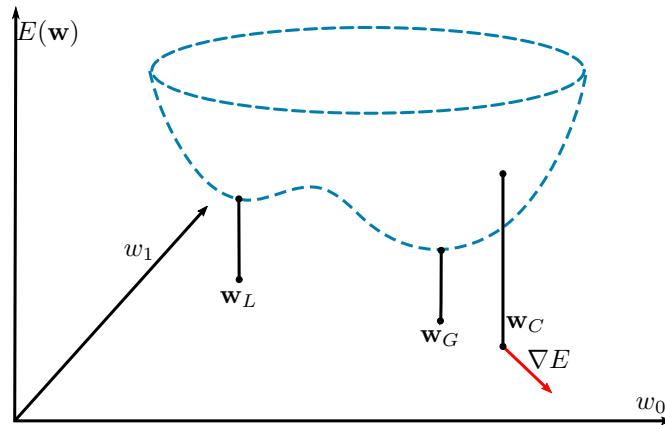


Figure 3.6: The schematic visualization of the error $E(\mathbf{w})$ as a function of weights $\mathbf{w} = [w_0, w_1]$ is shown. For simplicity, the bias parameter is covered in w_0 . The figure shows a local (\mathbf{w}_L) and a global (\mathbf{w}_G) minimum. The gradient of the error function ∇E at a hyperparameter configuration \mathbf{w}_c is highlighted in red. The figure is inspired from the textbook [71].

the global minimum of the error function $E(\mathbf{w})$ during the training process presents difficulties. The convergence of the training process depends on the choice of the initial values of the weights and biases. In simpler words, the training of the neural network begins with an arbitrary configuration of weights and biases, which undergoes non-linear parameter optimization during model training. This means that the parameters are optimized for multiple iterations τ until a global minimum \mathbf{w}_G (for ideal cases) of the error function is obtained. For successful training of neural networks, it is not necessary to obtain a global minimum but rather to compare several local minima to find a sufficiently good solution [71].

Non-linear parameter optimization utilizes the gradient information of the error function $\nabla E(\mathbf{w})$ to converge to a potential global minimum of the highly non-convex error function. One such method of non-linear parameter optimization is the stochastic gradient descent optimization (SGD) method. A stochastic gradient descent optimizer is a first-order method of non-linear parameter optimization. It updates the parameters of the neural network by computing the gradient of the error function $\nabla E(\mathbf{w})$ and using it as a correction term in every iteration τ . The parameter update for a stochastic gradient descent optimizer is given as

$$\mathbf{w}^{(\tau+1)} := \mathbf{w}^{(\tau)} - \eta \nabla E(\mathbf{w})^{(\tau)}, \quad (3.10)$$

where $\eta > 0$ is the learning rate. The learning rate is a hyperparameter which is set before training the neural network. There are other variants of a stochastic gradient descent optimizer including Adam [97], RMSprop [98], Adadelta [99], and Adagrad [100] that have shown success in the training of deep neural networks.

The parameter update is achieved through a well-known technique called the back-propagation technique [101]. This technique in the literal sense means layer-wise propagation of the error in the backwards direction from the last to the first layer of the neural network. A mathematical description of the backpropagation algorithm is given in the next section.

3.3.4 THE BACKPROPAGATION ALGORITHM

Given the non-linear parameter optimization methods in the previous section, the next goal is to use a principled mechanism for evaluating the gradient of the error function $\nabla E(\mathbf{w})$. This goal is achieved through the backpropagation algorithm [101]. The backpropagation algorithm has two distinct stages. In the first stage, the derivatives of the error function $E(\mathbf{w})$ with respect to the weights in the network are calculated. In the second stage, the derivatives are used to compute the adjustments in the weights to reduce the error through an optimization scheme like gradient descent, as shown in equation 3.10. To understand the first stage, which consists of calculating the derivatives of the error function, it is important to consider an MLP network (see figure 3.4) in a supervised regression setup, where the last layer has a linear activation function, the hidden layers have a sigmoid activation function, and the error function is defined by equation 3.7. The derivative of the error function $E(\mathbf{w})$ with respect to the weights in the network begins by applying the chain rule

$$\frac{\partial E}{\partial w_{ik}^l} = \frac{\partial E}{\partial a_k^l} \frac{\partial a_k^l}{\partial w_{ik}^l}, \quad (3.11)$$

where $a_k^l = \sum_i w_{ik}^l y_i^{l-1}$ is the activation of neuron k in layer l before it passes to the non-linear activation function ϕ ; and $y_i = \phi(a_i)$. The first term on RHS of equation 3.11 is referred to as error, $\delta_k^l = \partial E / \partial a_k^l$ and the second term is $\partial a_k^l / \partial w_{ik}^l = y_i^{l-1}$ [71]. Substituting the values above in equation 3.11 results in

$$\frac{\partial E}{\partial w_{ik}^l} = \delta_k^l y_i^{l-1}. \quad (3.12)$$

Intuitively, equation 3.12 makes sense as the weight w_{ik}^l connects the output of neuron i in layer $l-1$ to the input of neuron k in layer l . Similarly, partial derivatives of the error function with respect to all the weights in the networks can be computed beginning from the last layer and propagating backwards to the first layer. The partial derivatives can also be calculated with respect to the input \mathbf{x}_m , although in practice the derivatives are calculated only with respect to the weights of the hidden layers. Nevertheless, the calculation of gradients with respect to the inputs of the network is also a useful approach for visualizing and interpreting the network predictions. Such an application is mentioned in chapter 4.4. Similarly, error backpropagation can be

applied for networks with multiple outputs, networks with many hidden layers and for layers like the convolution layer.

Up to this point, the basic concepts of ML and DL have been discussed. Attention is now drawn to a part of AI which facilitates the transfer of knowledge from an ML or DL algorithm that was already trained to do a particular task [102]. This part is widely known as “transfer learning”, and its fundamentals will be discussed in the next section.

3.4 BASICS OF TRANSFER LEARNING

Transfer learning transfers knowledge learned from one or more source tasks \mathcal{T}_S and uses it to improve the performance on a related target task \mathcal{T}_T . Such knowledge transfer from the source task can affect the performance of the learning algorithm on the target task in three ways. First, the initial performance achieved on the target task using transferred knowledge compared to the initial performance without the transfer of knowledge can be affected. Second, the amount of time required to fully learn the target task given the transferred knowledge compared to the amount of time required to learn a task from scratch can be affected. Lastly, the final performance of an algorithm trained using transferred knowledge compared to the final performance of an algorithm without a transfer of knowledge can be affected. Transferring knowledge can have a negative and positive effect on model performance. Transfer learning that decreases the performance on the target task is referred to as “negative transfer learning”. Therefore, avoiding negative effects of transfer learning and obtaining “positive transfer learning” for the target task is a major challenge. In practical applications, the transfer of knowledge refers to the transfer of features (i.e. weights or parameters in the case of DL models) learned on the source task \mathcal{T}_S using the source dataset \mathcal{D}_S to improve the target task \mathcal{T}_T by optimizing the target function \mathcal{F}_T using the target dataset \mathcal{D}_T . In most cases, $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$.

Transfer learning can be used in two contexts: transfer in inductive learning and transfer in reinforcement learning [102]. This thesis addresses transfer in inductive learning where the transfer of knowledge is used to perform classification tasks using convolutional neural networks (see section 4.5). From here onwards, reference is made to inductive transfer learning as transfer learning. Many examples of transfer learning for biophotonic data have been published in the last few years [103, 104, 105, 106]. In real-world biophotonic applications, transfer learning is desirable due to the lack of a training dataset to construct large DL models. One such example is presented in section 4.5, where two transfer learning strategies (i.e. feature extraction using DL models and the fine-tuning of DL models) are presented. A detailed explanation of the two transfer learning strategies is given in reference [24].

4

Selected work

The previous chapters focused on building the theoretical concepts of Raman spectroscopy, NLM imaging, AI, machine learning and deep learning. Now, this chapter enlightens the readers with various AI-based technologies used for analyzing Raman spectroscopic and NLM imaging data. These applications are based on the publications and manuscripts presented in chapter 8. This chapter will only summarize the AI-based applications; however, a detailed explanation is made in their respective publications (applications also summarized in figure 4.1). An outline of the applications of AI-based technologies is as follows:

- The first application uses machine learning for the identification of sepsis in a peritonitis mouse model using NLM images. The machine learning approach utilizes manually extracted statistical features to construct a linear classification model for sepsis identification.
- The second application uses machine learning and deep learning to perform an image semantic segmentation task. An image semantic segmentation task classifies pixels of an NLM image into four tissue regions using a PCA-LDA model and an autoencoder model. The tissue regions, especially the crypt region, are essential for assessing the activity of inflammatory bowel disease.
- The third application uses deep learning to generate a pseudo-stain H&E model that can computationally stain an NLM image into an H&E stained image. The deep learning approach uses generative adversarial networks in a supervised and an unsupervised manner for pseudo-staining NLM images.

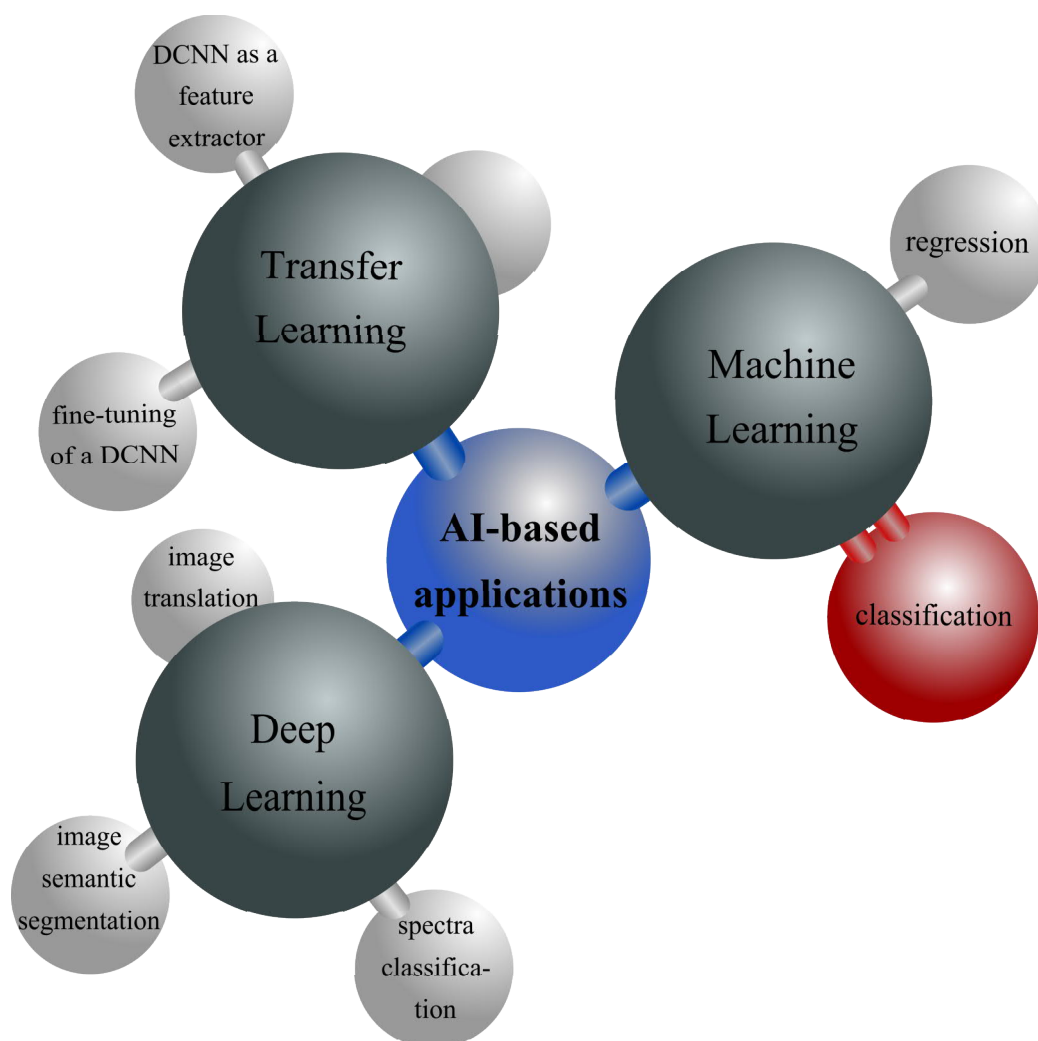


Figure 4.1: This figure shows applications of AI for Raman spectroscopic and non-linear multimodal imaging data depicted in the form of a dimethylformamide (DMF) solvent molecule. The major division of AI applications is shown by the carbon atoms (gray) and the applications in the form of hydrogen (white) and oxygen (red) atoms.

- The next application is spectral classification using deep learning models. The deep learning model is a one-dimensional deep convolutional neural network (1D-CNN) which is used for the classification of the Raman spectrum into four disease stages of ulcerative colitis. The 1D-CNN model is also used to interpret the contribution of each Raman band for classification of the four disease stages.
- Finally, the last section shows inductive transfer learning of deep learning models to work with small datasets. This work combines histologically and immunohistochemically stained images to identify breast cancer.

4.1 IDENTIFICATION OF SEPSIS USING NON-LINEAR MULTIMODAL IMAGES

This chapter presents a machine learning application for the identification of early sepsis liver injury in a mouse model using NLM images. Sepsis is life-threatening organ failure resulting from dysregulated host responses to infection. Identification of sepsis using biomolecular information provided by NLM images can assure early sepsis detection. However, identification of biomolecular information is a challenging task. The subtle variations in biomolecular information caused by sepsis infection are difficult to detect by pathologists. For this purpose, a machine learning algorithm is used. Machine learning algorithms not only provide better accuracy but also have the potential to automate the whole sepsis identification process in clinics (for details of this work refer [107]).

As discussed in section 3.2, machine learning approaches require the pre-processing of data and extraction of features to build a model. Therefore, in this work, as a pre-processing step, a region which shows the most biomolecular alterations in a mouse liver infected by sepsis was extracted. This region was 20 μm around the periportal and pericentral veins visible in the NLM images (see left panel in figure 4.2). The 20 μm region around both veins was extracted using the TPEF signal and performing Otsu's thresholding and regional morphological operations [108]. In the 20 μm region, nine first-order statistical features based on the image histogram were extracted. The statistical features were extracted for each pixel using a 5×5 window around that pixel. This led to nine texture feature images for each channel (CARS, TPEF, SHG) of the NLM image. Statistical feature images like mean, minimum, skewness, and standard deviation are shown in the right panel of figure 4.2. Subsequently, from the nine texture feature images a median value was estimated in the 20 μm region. Thus, 27 median values (9 texture feature images \times 3 channels) were obtained from the texture feature images. Based on the median values of the texture feature images, the PCA-LDA model (see section 3.2.2) was constructed to classify NLM images into two categories, namely: sepsis and control. To interpret the contribution of each modality (i.e., the CARS, TPEF, and SHG signal), the median values from the texture feature images of the individual modalities were used for constructing a PCA-LDA model. The PCA-LDA model used a two-step cross-validation procedure [109] with a leave-one-mouse-out cross-validation strategy. The internal cross-validation step optimized the number of principal components, and the external cross-validation step predicted an independent test set. The average of the prediction on the independent test set was used to measure performance metrics like sensitivity, specificity, and the receiver operating characteristic (ROC) curve (see right panel in figure 4.2).

The results in the form of the ROC curve obtained from the PCA-LDA classification model using features from all the modalities showed 85% sensitivity. The individual modalities, namely CARS, TPEF, and SHG, achieved a sensitivity of 93%, 83%, and 49% respectively [107]. A pathologist confirmed these results for their biological inter-

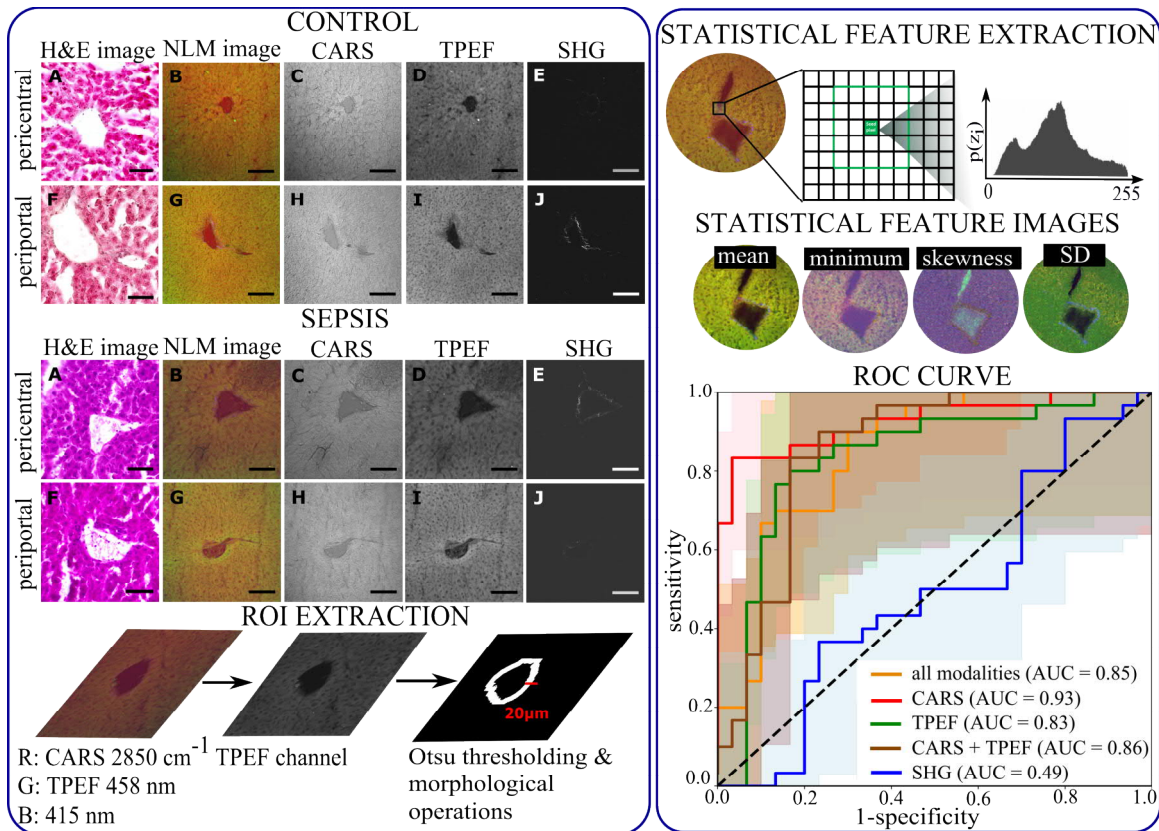


Figure 4.2: An application of machine learning for identifying sepsis using non-linear multimodal images is shown. The left panel shows two types of veins: periportal and pericentral veins for control and sepsis individuals, respectively. The left panel also shows image pre-processing to obtain a region of interest around the veins, followed by feature extraction in the right panel to build a PCA-LDA model. The ROC results of the PCA-LDA model for a binary classification task are also shown in the right panel. The scale bar is $100 \mu\text{m}$.

pretation. According to the biological interpretation, the CARS signal at 2850 cm^{-1} maps the CH_2 group, and the TPEF signal at $426\text{-}490 \text{ nm}$ maps NAD(P)H levels in a tissue region. These two biomolecules showed alterations in the mice infected with sepsis, due to the inflammation and metabolic activity in the $20 \mu\text{m}$ region around both veins. Furthermore, SHG at 415 nm maps collagen in the tissue section. The collagen metabolism was poor in early septic liver injury. Thus, the PCA-LDA model built on SHG texture features achieved poor classification performance (49%). Overall, the results of this pre-clinical study proved that a machine learning approach with non-linear multimodal imaging could endeavor to detect sepsis in the early stages.

4.2 SEMANTIC SEGMENTATION OF NON-LINEAR MULTIMODAL IMAGES

This chapter presents an application of deep learning for the semantic segmentation of NLM images. Semantic segmentation of NLM images into regions can be beneficial for the assessment of inflammatory bowel disease (IBD). Assessment of IBD requires observing certain regions, particularly the crypt and mucosa region, in the tissue section. These regions are traditionally annotated by pathologists, which is a tedious process. Therefore, annotating these regions through an automatic semantic segmentation pipeline is desired. In this prospect, this section proposes the semantic segmentation of crypts and the mucosa region using a ML and DL approach (for detailed information of this work, see [110]).

For the ML approach, a PCA-LDA model was trained using statistical features extracted for pixels. Eleven statistical features including mean, standard deviation, skewness, kurtosis, median, energy, entropy, RMS, variance, maximum, and minimum were extracted for each channel using a 5×5 window around a pixel. The definition of all statistical features can be found in [107]. Thus, 33 texture features were extracted per pixel, and PCA was used for dimension reduction. The reduced feature dimension space was used to train an LDA model to classify pixels into three regions: non-mucosa, mucosa without crypt, and crypt (see upper panel in figure 4.3). The fourth region, i.e. the background region, was obtained by using K-means clustering for the NLM image (see section 3.2.3).

For the DL approach, an autoencoder like the SegNet model [111] with an encoder and a decoder structure was used. The encoder had 13 blocks of a convolutional layer, a ReLU activation layer, and a max-pooling layer. Similarly, the decoder had 13 blocks of an upsampling layer, a convolutional layer, and a batch normalization layer. The last layer of the decoder is a softmax activation layer, which maps the features of the decoder to probability values associated with the four regions. The input of the SegNet model was a pair of NLM patches and pathologically annotated patches. The weights of the SegNet model were optimized using an SGD optimizer with a learning rate of 10^{-4} and minimizing a categorical cross-entropy loss function. The output of the SegNet model was a segmented map of an NLM image with four distinct regions. The regions like background, non-mucosa, mucosa without crypt, and crypt are indicated by black, blue, green, and red, respectively (see lower panel in figure 4.3). Subsequently, a qualitative and quantitative evaluation of the segmented map was done for the PCA-LDA and the SegNet model.

The qualitative evaluation of the segmented map for the test dataset using statistical features and the PCA-LDA model showed an overall poor performance. The poor performance can be attributed to the difficulties encountered by the PCA-LDA model to classify the pixels based on the statistical features. The statistical features used for the PCA-LDA model calculate values based on the intensity at the pixel level and thus could barely retain structural information of the crypt and the mu-

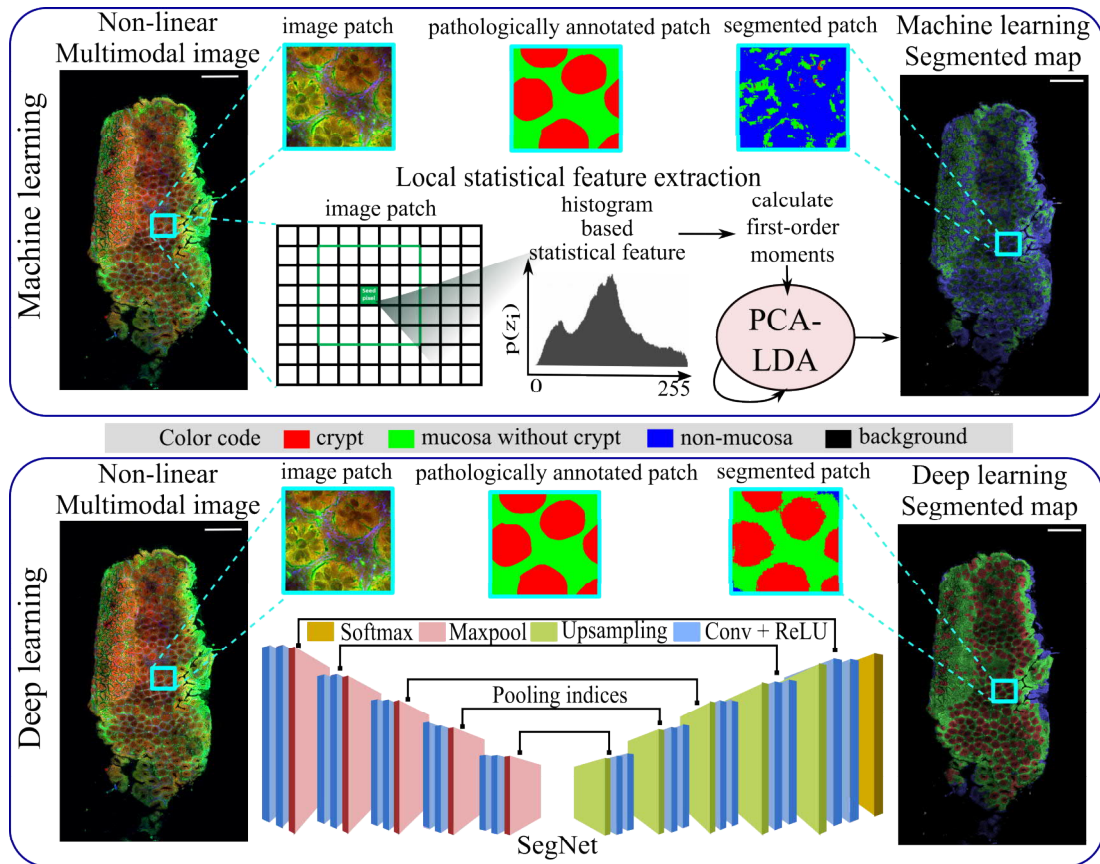


Figure 4.3: An application of AI for semantic segmentation of NLM images is shown. The upper panel shows the ML approach by extracting statistical features for each pixel and using a PCA-LDA model for classifying these pixels. The lower panel shows an autoencoder architecture (the SegNet model) to classify the pixels into four tissue regions. The four regions are background (black), non-mucosa (blue), mucosa without crypt (green), and crypt (red). The scale bar indicates $200 \mu\text{m}$.

cosa region. However, the qualitative evaluation of the segmented map for the test dataset using the SegNet model showed an acceptable performance for NLM images with regularly shaped crypts and a distinct mucosa region. The NLM images with distorted and overlapping crypts were difficult to segment using the SegNet model. The quantitative evaluation was performed by an F1 score. The F1 score for the crypt region, which is essential for IBD characterization using the SegNet model and the PCA-LDA model, is $\sim 63\%$ and $\sim 18\%$, respectively. Likewise, the F1 score for the mucosa region using the SegNet model and the PCA-LDA model is $\sim 55\%$ and $\sim 27\%$, respectively [110]. In summary, this work achieved semantic segmentation of NLM images using the machine learning and deep learning approaches.

4.3 PSEUDO H&E STAINING OF NON-LINEAR MULTIMODAL IMAGES

This chapter presents an application of deep learning for the pseudo-staining of an NLM image into an H&E stained image without following the conventional staining protocol in laboratories. Pseudo-staining of NLM images can be beneficial for several reasons. Firstly, pseudo-stained H&E images obtained from NLM images do not require staining of the same or parallel tissue sections. Secondly, it does not require image registration of pathologically stained H&E images to the coordinate space of NLM images (if pseudo-staining performed in unsupervised manner). Thirdly, the pseudo-staining of NLM images can be used to virtually stain NLM images into any type of stained images.

The pseudo-staining of NLM images in this section proposes two models using generative adversarial networks. The first model is a CGAN model to perform supervised learning (see section 3.2.2). The CGAN model requires a pair of NLM images and its corresponding pathologically stained H&E image. The second model is the cycle CGAN model, which uses unsupervised learning (see section 3.2.3). Contrary to the CGAN model, the cycle CGAN model does not require a corresponding set of an NLM image and a pathologically stained H&E image. The CGAN and cycle CGAN models are summarized below; however, the architecture and training details of both models can be found in chapter 8.

The CGAN model comprises two deep learning models called a generator and a discriminator. The generator is an autoencoder, while the discriminator is a deep convolutional neural network (see section 3.3.2). The input to the generator model is an NLM image, and the output is a pseudo-stained H&E image. The quality of the pseudo-stained H&E image is evaluated by the discriminator. The discriminator classifies the pseudo-stained H&E image as “real” or “fake” based on small regions in that image (see upper panel in figure 4.4). The generator creates more “real” looking pseudo-stained H&E images by optimizing the mean absolute error between the pathologically (or target) stained H&E image and the pseudo-stained H&E image. The generator model is also optimized by an adversarial loss, which is updated via the discriminator. The discriminator identifies the pseudo-stained H&E image as “fake” by optimizing its weights through the mean squared error loss function.

The cycle CGAN model is an extension of the CGAN model with two generators and two discriminators. The first generator model uses an NLM image as an input and creates a pseudo-stained H&E image at the output. The second generator uses the pseudo-stained H&E image (output of generator 1) and reconstructs it into an NLM image. The two discriminators determine the realness of the pseudo-stained H&E image and the reconstructed NLM image obtained from their respective generator models (see lower panel in figure 4.4). Each generator-discriminator pair is trained similarly to a CGAN model. Additionally, the two generator models use a cycle consistency loss function, which is the mean absolute error between the original NLM

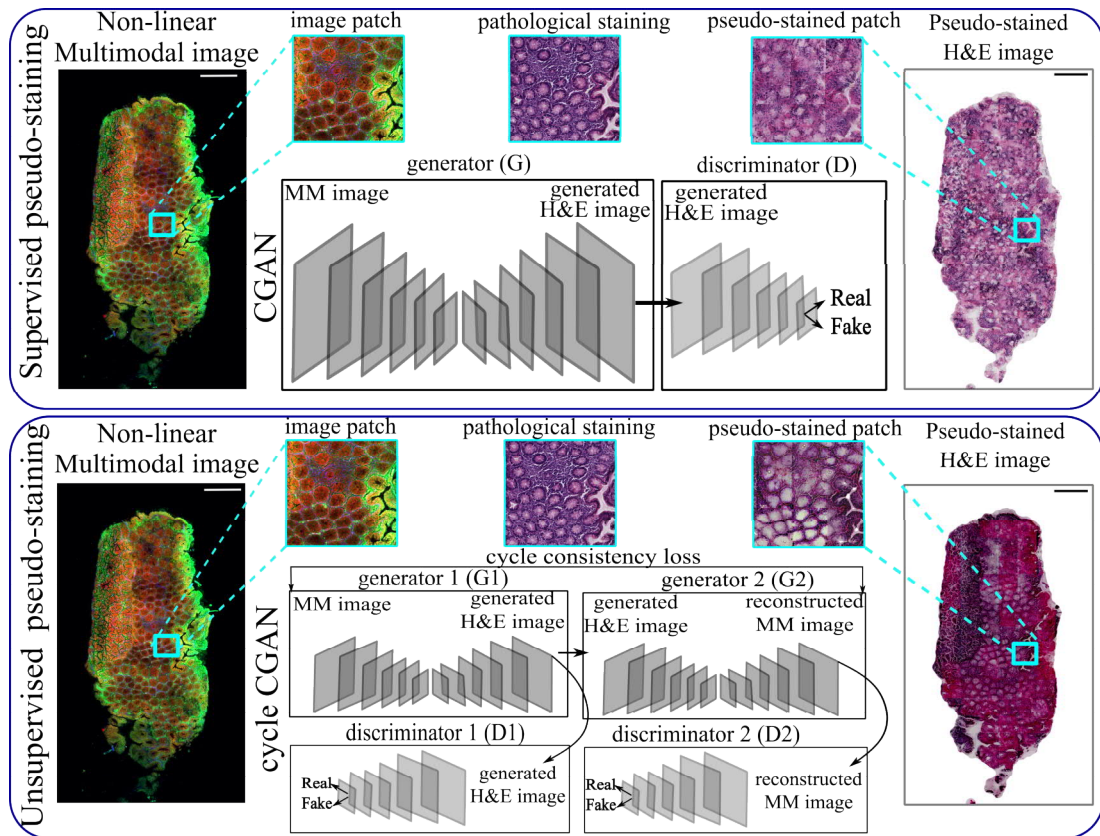


Figure 4.4: An application of deep learning for image translation of non-linear multimodal images is shown. The upper panel shows supervised approach by using the conditional generative adversarial (CGAN) and lower panel shows an unsupervised approach using the cycle conditional generative adversarial (cycle CGAN) to perform pseudo H&E staining of multimodal images. The scale bar indicates $200 \mu\text{m}$.

image (input of generator 1) and the reconstructed NLM image (output of generator 2).

The H&E images generated from the CGAN and cycle CGAN model were visually and quantitatively evaluated. On visual inspection, it was seen that the pseudo-stained H&E images generated by the CGAN and the cycle CGAN model were similar to the pathologically stained H&E image. The quantitative analysis based on the structure similarity index and color shading similarity showed values of >0.50 and >0.90 , respectively. Overall, it was observed that the results of the pseudo-stained H&E image could be improved before accepting in clinics.

4.4 INTERPRETATION OF DEEP LEARNING MODELS

Traditionally, deep learning models have been considered “*black-box*” models, which means that it is difficult to interpret the predictions made by the deep learning models. Thus, using deep learning models for clinical applications is limited, as the interpretation of disease associated biomarkers is a major concern. Similarly, understanding the predictions made by the deep learning model trained on Raman spectroscopic data is important for understanding important Raman bands associated with a disease. In conventional machine learning models like PCA or LDA, the interpretation of Raman bands is acquired by the loading’s matrix. However, in deep learning models where the function is highly non-linear, the interpretation of Raman bands is not straightforward. Therefore, this section presents the interpretation of the deep learning models by approximating the non-linear function using first-order Taylor expansion.

In this work, a one-dimensional convolutional neural network (1D-CNN) was trained for classifying the Raman spectroscopic data of UC patients into four Mayo endoscopic scores. Along with a 1D-CNN classification model, a 1D-CNN regression model was constructed for detecting border-line patients and quantifying the extent of misclassification obtained by the 1D-CNN classification model (see upper panel in figure 4.5). As mentioned in section 3.3.3, the 1D-CNN classification and regression models differ in the activation functions of the last layer and the loss function. The details of the architecture of the 1D-CNN classification and 1D-CNN regression models and their training process are discussed in [25]. After the training of the 1D-CNN classification model, it was evaluated using performance metrics like mean sensitivity, mean specificity, and risk factor. Subsequently, the trained 1D-CNN classification model was used for interpretation purposes. For interpretation purposes, the softmax activation function of the last layer in the 1D-CNN classification model was replaced by a linear activation function to achieve the unnormalized score. Further, the non-linear function of the 1D-CNN classification model was approximated using first-order Taylor expansion to obtain a variable weighting of the Raman bands. The variable weighting is represented by a heat map (see lower panel in figure 4.5). The heat map is yellow for more relevant Raman bands and violet for less relevant Raman bands when predicting a Mayo endoscopic score. For each Mayo endoscopic score, salient Raman maps are marked with red dotted lines in the lower panel of figure 4.5.

The results of the 1D-CNN classification model achieved a mean sensitivity of $\sim 78\%$, a mean specificity of $\sim 93\%$, and a risk factor of ~ 1.4 . The individual mean sensitivities for a Mayo endoscopic score of 0, 1, 2, and 3 were 100.00%, 81.82%, 75.00%, and 55.56%, respectively. The critical part was the interpretation of the 1D-CNN classification model, which showed influential Raman bands via the heat map. The important Raman bands predicted by the 1D-CNN model were assigned to a biomolecule by a pathologist. According to the pathologist, Raman bands associated with proteins, lipids, cholesterol, amino acids, saccharides, and DNA were marked

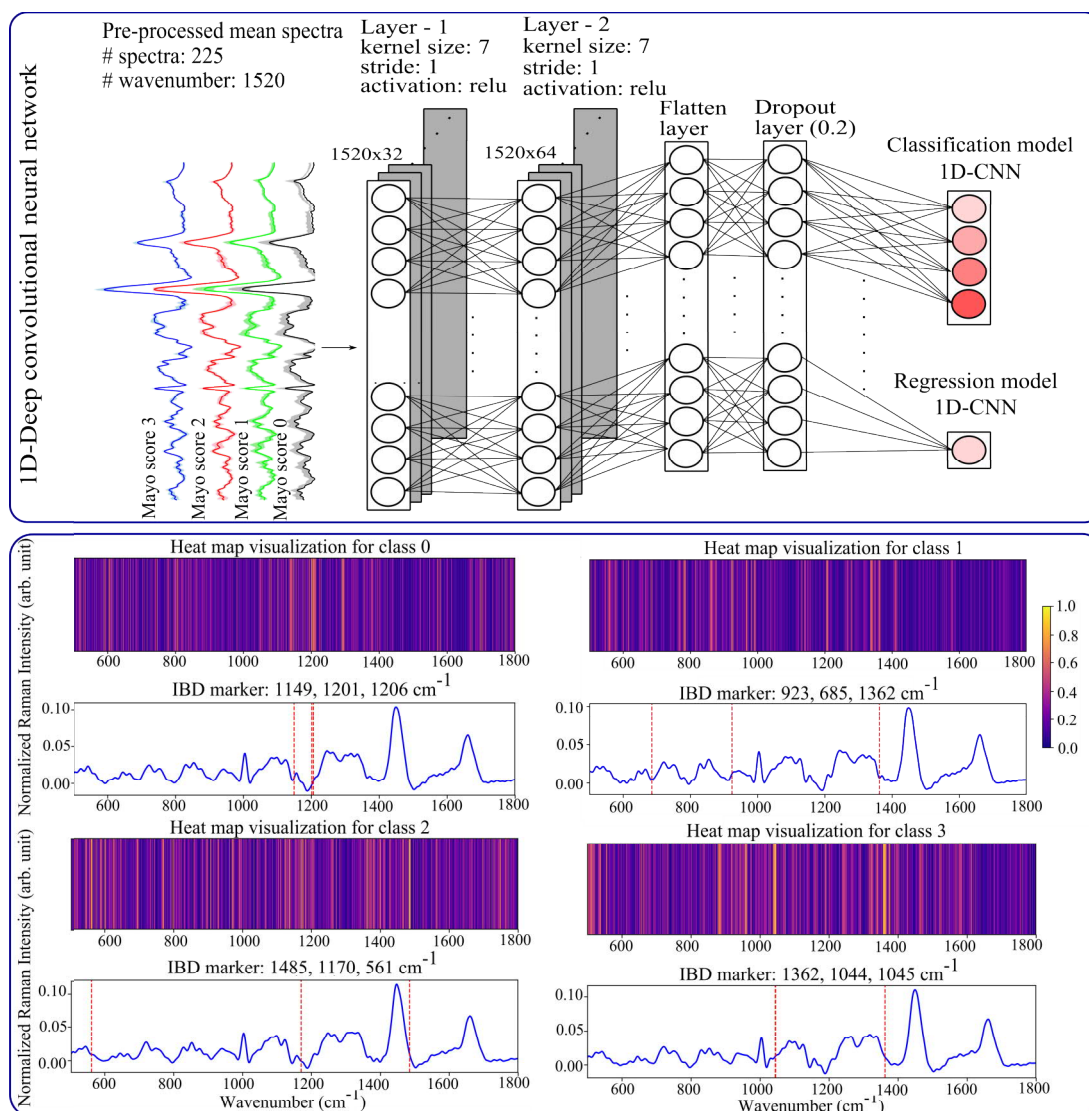


Figure 4.5: An application of deep learning for spectra classification is shown. The upper panel shows a 1D-deep convolutional neural network for the classification of Raman spectroscopic data into four Mayo endoscopic scores. It also shows a 1D-CNN regression model. The lower panel visualizes a heat map for the four Mayo endoscopic scores obtained after approximating the non-linear function of the 1D-CNN classification model.

salient by the 1D-CNN classification model.

Overall, this work utilized first-order Taylor expansion for interpretation of a highly non-linear function learned by a DCNN model. This work provides information about Raman bands relevant for the prediction of four Mayo endoscopic scores. However, the first-order Taylor approximation of non-linear function cannot retain the relationship between the Raman bands for a Mayo score prediction. The relationship between the

Raman bands is essential when a group of biomolecules are altered during disease progression. Thus, the interpretation of deep learning models using higher-order systems requires further investigation.

4.5 TRANSFER LEARNING FOR BREAST CANCER DIAGNOSIS USING DATA FUSION

Breast cancer is the world's most commonly occurring type of cancer among women. Diagnosing breast cancer utilizes a histology (H&E) staining technique and is performed via visual evaluation of the H&E stain tissue section by a pathologist. Visual analysis of the H&E stain tissue can lead to subjective interpretation. Thus, this work motivates the use of deep learning for automatic breast cancer detection. Furthermore, this work motivates the fusion of histology with immunohistochemistry (IHC) imaging data for reliable and early breast cancer diagnosis. Finally, due to the small dataset size this work also proposes the use of transfer learning for deep learning models to diagnose breast cancer based on H&E and IHC imaging data.

In this work, H&E imaging data along with four types of IHC imaging data (i.e., progesterone receptor (PR), estrogen receptor (ER), human epidermal growth factor-2 (Her2), and Ki-67 nuclear protein) were obtained from the biopsies of 23 women. Each biopsy was a combination of these five stain type images (see upper panel in figure 4.6). Based on the five stain type images, the biopsies were classified as normal or tumor using pre-trained deep convolutional neural networks (DCNN). As the dataset size was limited, a full training of DCNN was avoided; instead, pre-trained DCNNs based on two transfer learning strategies were used. The two transfer learning strategies were performed using three pre-trained DCNNs, namely: VGG16, Inceptionv3, and ResNet50. The two transfer learning strategies, DCNN as a feature extractor and the fine-tuning of DCNN, are explained further.

In the first transfer learning strategy, each pre-trained DCNN was used to extract off-the-shelf features from the five stain type images. To use the pre-trained DCNN as a feature extractor, patches of 1024×1024 in size were extracted from the five stain type images and down sampled to the input size requirement of the pre-trained DCNN. Subsequently, features from the pre-trained DCNN were extracted from the down sampled patches. The extracted features from the patches of all stain types were combined, and a PCA-LDA model was trained. The PCA model was used to reduce the dimensions of the features, and LDA was used to classify the patches as a tumor or normal (see lower panel of figure 4.6). The model training used a two-step leave-one-patient-out cross-validation strategy [109]. In this two-step cross-validation strategy, the internal step used 10-fold cross-validation to optimize the number of PCs, whereas the external cross-validation step evaluated the model performance on an independent test set. The predictions on the independent test set were used to calculate the confusion matrix, mean sensitivity, and mean F1 score.

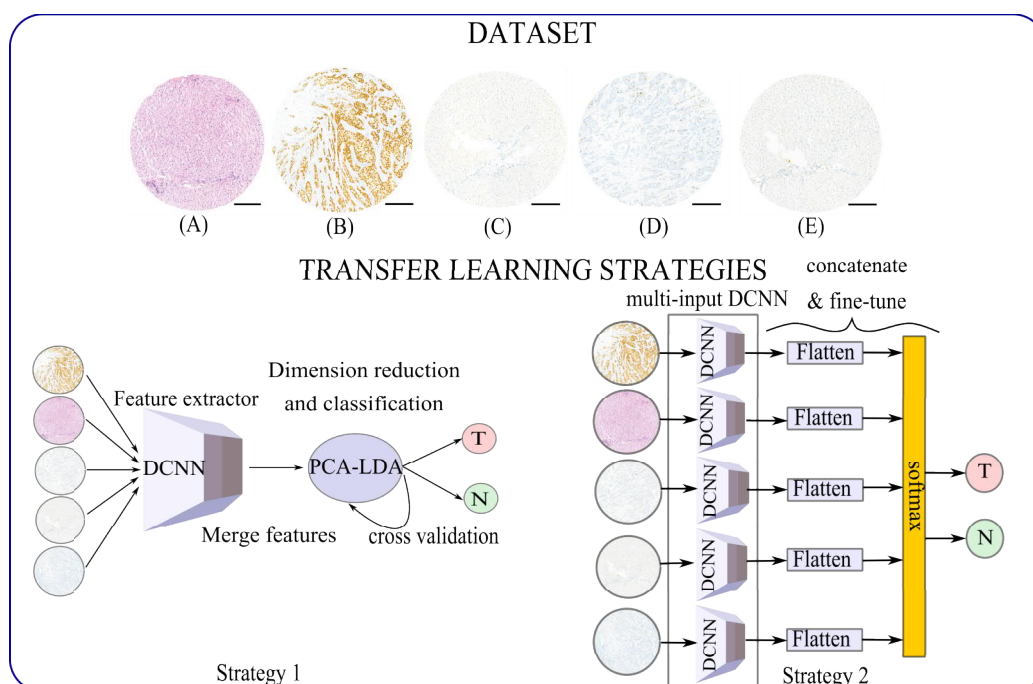


Figure 4.6: An application of transfer learning using deep learning models for breast cancer diagnosis is shown. The upper panel shows five stain type images: (A) H&E stain image, (B) estrogen receptor (ER), (C) human epidermal growth factor-2 (Her2), (D) Ki-67 nuclear, and (E) progesterone receptor (PR). The scale bar is $200 \mu\text{m}$. The lower panel shows two transfer learning strategies: DCNN as a feature extractor (left) and DCNN for fine-tuning (right).

In the second transfer learning strategy, a multi-input pre-trained DCNN was constructed such that the five stain type images were the input, and two neurons representing tumor or normal were the output (see lower panel in figure 4.6). The weights of the last two layers of the three pre-trained multi-input DCNN were fine-tuned to perform the binary classification task. The fine-tuning of the models generated in strategy 2 were evaluated using leave-one-patient-out cross-validation, similar to strategy 1. In the two-step cross-validation procedure, the internal step optimized the hyperparameters of the DCNN based on the training and validation dataset, and the external step evaluated the model performance based on an independent test set. A detailed explanation of the two transfer learning strategies is given in chapter 8.

The results showed that transfer learning, in cases of small datasets, is beneficial. It was seen that for small datasets, DCNNs as feature extractors (i.e., transfer learning strategy 1) produced promising results for all three DCNNs explored in this work. Furthermore, it was seen that the data fusion of histopathology and immunohistochemistry slightly improves breast cancer diagnosis; therefore, it should be encouraged in clinics.

5

Summary

This section provides an entire summary of the thesis. Chapter 1 shortly introduced light-matter interaction and biophotonic technologies that are concerned with the use of light to study biological objects. Chapter 1 also introduced various biophotonic technologies including optical coherence tomography, infrared spectroscopy, and Raman spectroscopy. Out of all the biophotonic technologies, this thesis used only two biophotonic technologies, namely: Raman spectroscopy and non-linear multimodal imaging. It was mentioned in chapter 2 that both biophotonic technologies provide multivariate datasets, and analysis of these datasets are crucial for better understanding the biochemical system. For data analysis purposes, this chapter also introduced the field of chemistry called “chemometrics”. The recent trend in chemometrics is to use AI, machine learning, and deep learning, all of which were elaborated on in chapter 3. It was mentioned in chapter 3 that the recent AI models based on deep learning models can efficiently perform complex data analysis tasks. Chapter 3 also presented various machine learning and deep learning models; their applications were discussed in chapter 4. These applications are related to image and spectra classification, semantic segmentation, and image translation. Furthermore, chapter 4 also presented a way to interpret the deep learning models. This chapter also presented the “transfer learning” of deep learning models when working with small datasets. Before summarizing each application, each topic presented in this thesis, including AI, machine learning, and deep learning, will be briefly discussed.

Artificial intelligence (AI) is a science which simulates human intelligence by programming machines and manifesting traits of human learning and problem solving. It was already mentioned in chapter 1 that AI has rapidly growing applications in the

field of natural language processing, finance, agriculture, banking, autonomous vehicles, chat bot systems, gaming, and healthcare. Similarly, AI has paved its way by analyzing chemical systems way back. The first analytical chemistry-related AI program was capable of generating molecular structures from the molecular formula and predict its mass spectrum [112]. Since then, a lot of technical advancements like using sub-fields of AI, machine learning, and deep learning for analyzing chemistry-related data has been observed.

Machine learning is a sub-field of AI. Machine learning algorithms can automatically learn and improve from experience without being explicitly programmed. As mentioned in chapter 3, the aim of a machine learning algorithm is to become better in performing a task by gaining experience from the dataset. The goodness of the machine learning algorithm is evaluated by performance measure. In this thesis, two categories of machine learning were introduced, namely: supervised and unsupervised learning. Further, for supervised learning, classification models like SVM and LDA, as well as a regression model like multivariate regression, were introduced. Similarly, for unsupervised learning, K-means clustering and PCA were explained. Finally, in chapter 3, various limitations of machine learning were addressed. One of the limitations encountered by machine learning algorithms is the need for a “feature extraction” step before constructing a classification or regression model. As the feature extraction step is subjective and depends on an expert, the performance of the machine learning algorithm is dependent on the extracted features. For this purpose, algorithms that can extract features automatically from the data are needed. Such algorithms are categorized into a sub-field of machine learning called “deep learning”.

Deep learning algorithms use artificial neural networks (ANNs) motivated by the human neural system. Briefly, an ANN comprises input, output, and hidden layers. These layers are composed of basic computational units called “neurons” or “nodes”. Each neuron receives input signals from the neurons of the preceding layer and passes the output signal to the neurons of the following layers. The output signal of a neuron is non-linearly transformed by an activation function. Chapter 3 provides mathematical expression of ANNs and other deep learning models like a convolutional neural network, an autoencoder, and generative adversarial networks. Knowing the two sub-fields of AI, the applications of machine learning and deep learning presented in this thesis are further summarized below.

- **Image classification** of non-linear multimodal images was investigated using machine learning. This study was performed to identify early sepsis in mouse liver sections. In the machine learning approach, statistical features based on the histogram of the non-linear multimodal image were extracted. Based on the statistical features, a linear classification model like the PCA-LDA model was trained. The PCA model was used to reduce the dimension size of the statistical features, and the LDA model was used to classify these features into two

groups: sepsis and control. Furthermore, to understand the contribution of the individual modalities (CARS, TPEF, and SHG) of the non-linear multimodal image, the statistical features were extracted from each modality. Based on the statistical features extracted from the individual modalities, the PCA-LDA model was used for the classification task. The results of this study showed that features extracted from CARS and TPEF achieved better classification performance as compared to features extracted from SHG modality (see section 4.1).

- **Semantic Image segmentation** using machine learning and deep learning approaches was investigated in this thesis to identify regions of non-linear multimodal images that are important for the characterization of inflammatory bowel disease. The basic idea of semantic image segmentation is to classify pixels instead of images. The classification of pixels was done using machine learning and deep learning approaches. In the machine learning approach, statistical features for pixels were extracted, and a linear classification (PCA-LDA) model based on these features was constructed. The PCA model was used to reduce the dimension size and the LDA model was used to classify the pixels into four regions: background, mucosa without crypt, crypt, and non-mucosa. In the machine learning approach, the background region was extracted using an unsupervised learning algorithm like the K-means clustering algorithm. The deep learning approach, on the other hand, used an autoencoder model (the SegNet model [111]) for segmentation of non-linear multimodal images into four regions. Here, a feature extraction step was not performed. This is the advantage of deep learning models: the features were self-learned from the dataset itself. Finally, on comparison of the results from the machine learning and deep learning models, it was seen that the the SegNet model performed better than the PCA-LDA model (see section 4.2).
- **Image translation** as used in this thesis transforms a non-linear multimodal image into a histologically stained H&E image. Such a transformation is beneficial for virtual staining of the non-linear multimodal images without performing conventional H&E staining in laboratories, thus reducing the effort of pathologists. Here, image transformation is performed using a deep learning model like a generative adversarial network (GAN) (see section 3.3.2). The GAN was used in a supervised and unsupervised approach. The supervised GAN model (the Pix2Pix model) required a corresponding pair of a non-linear multimodal image and an H&E stained image. However, the unsupervised deep learning approach (the cycle CGAN model) did not require paired images. The unsupervised approach reduced the effort of image registration, which is a vital step in the supervised approach. Moreover, the unsupervised method was advantageous due to its ability to reconstruct original non-linear multimodal images

(see section 4.3).

- **Spectra classification** of Raman spectroscopic data using the deep learning model was explored in this thesis. Here, a one-dimensional deep convolutional neural network (1D-CNN) was presented. The 1D-CNN model was used for the classification task to characterize four Mayo scores in ulcerative colitis patients. Similarly, the 1D-CNN model was used in a regression task for detecting border-line patients and quantifying the extent of misclassification. In both classification and regression task, the input was a Raman spectrum which was passed through two hidden layers. The output was either probabilities for the classification task or a real-valued number for the regression task. The results were promising to characterize disease stages of ulcerative colitis (see section 4.4).
- **Interpreting AI** models or understanding the predictions made by deep learning models was another contribution in this thesis. For this purpose, the 1D-CNN model used for the spectra classification task was used. The non-linear function of the trained 1D-CNN model was interpreted using a first-order Taylor expansion. The first-order Taylor expansion was used as the variable weighting of the Raman spectral bands to investigate biologically relevant Raman bands for Mayo score classification. A Raman band with high variable weighting was interpreted to highly influence the classification of a Mayo score, while a Raman band with low variable weighting was less likely to influence that Mayo score (see section 4.4).
- **Transfer learning of deep learning models** using histology and immunohistochemistry imaging data was investigated in this thesis. The combination of histology and immunohistochemistry was performed to gain different insights into breast cancer diagnosis. The transfer learning of deep learning models was explored due to the small dataset size of histology and immunohistochemistry images. Here, two transfer learning strategies were developed. In the first strategy, the deep learning models were used to extract features for the two imaging datasets. In the second strategy, the weights of the deep learning models were optimized to obtain the best classification results. The results showed that the first transfer learning strategy (i.e., utilizing deep learning models as a feature extractor) works efficiently for the breast cancer dataset presented here (see section 4.5).

It was impractical to present all applications of deep learning models; therefore, these models, along with other potential applications, are presented in chapter 7.

6

Zusammenfassung

Dieser Abschnitt stellt eine Zusammenfassung der Doktorarbeit dar. Kapitel 1 führte kurz in die Licht-Materie-Wechselwirkung und biophotonische Technologien, welche sich mit der Nutzung von Licht zur Untersuchung biologischer Objekte befassen, ein. In Kapitel 1 werden auch verschiedene biophotonische Technologien vorgestellt, darunter die optische Kohärenztomographie, die Infrarotspektroskopie und die Raman-Spektroskopie. Von allen biophotonischen Technologien werden in dieser Arbeit nur zwei biophotonische Technologien verwendet, nämlich die Raman-Spektroskopie und die nichtlineare multimodale Bildgebung. In Kapitel 2 wird beschrieben, dass beide biophotonischen Technologien multivariate Datensätze liefern und dass die Analyse dieser Datensätze für ein besseres Verständnis des biochemischen Systems von entscheidender Bedeutung ist. In diesem Kapitel wird auch der Begriff “Chemometrie” eingeführt, welche die Analyse von chemischen Daten bezeichnet. Der jüngste Trend in der Chemometrie ist die Verwendung von Künstlichen Intelligenz (KI) basierenden Verfahren, maschinellem Lernen und tiefem Lernen, wobei alle diese Techniken in Kapitel 3 näher erläutert werden. In Kapitel 3 wird auch darauf eingegangen, dass KI-Modelle, die auf tiefem Lernen basieren, komplexe Datenanalyseaufgaben effizient durchführen können. In Kapitel 3 werden auch verschiedene Modelle für maschinelles Lernen und tiefes Lernen vorgestellt. Die Anwendung dieser Verfahren wird in Kapitel 4 diskutiert. Diese Anwendungen sind Bild- und Spektrenklassifikation, semantische Segmentierung und Bildübersetzung. Darüber hinaus wird in Kapitel 4 auch eine Möglichkeit zur Interpretation von tiefen Lernmethoden vorgestellt. In diesem Kapitel wurde auch das Transfer-Lernen von tiefen Lernmodellen bei der Arbeit mit kleinen Datensätzen vorgestellt. Bevor die einzelnen Arbeiten zusam-

mengefasst werden, wird im Folgenden jedes in dieser Arbeit vorgestellte Thema, einschließlich KI, maschinelles Lernen und tiefes Lernen, kurz diskutiert.

Künstliche Intelligenz (KI) ist eine Wissenschaft, die menschliche Intelligenz simuliert, indem sie Maschinen programmiert und Eigenschaften des menschlichen Lernens und Problemlösens adaptiert. Es wurde bereits in Kapitel 1 erwähnt, dass KI viele und wachsende Anwendungen in verschiedenen Bereichen wie der Verarbeitung natürlicher Sprache, Finanzen, Landwirtschaft, Bankwesen, autonome Fahrzeuge, Chat-Bot-Systeme, Spiele und Gesundheitswesen hat. In ähnlicher Weise wurde KI auch zur Analyse chemischer Systeme angewandt. Das erste chemische KI-Programm war in der Lage, molekulare Strukturen aus der Molekülformel zu generieren und ihr Massenspektrum vorherzusagen [112]. Seitdem wurden viele technische Fortschritte wie die Verwendung von Teilgebieten der KI, maschinelles Lernen und tiefes Lernen für die Analyse chemiebezogener Daten erreicht.

Maschinelles Lernen ist ein Teilgebiet der KI. Algorithmen des maschinellen Lernens können automatisch aus Erfahrungen lernen und sich so verbessern, ohne explizit programmiert zu werden. Wie in Kapitel 3 erwähnt, besteht das Ziel eines maschinellen Lernalgorithmus darin, bei der Ausführung einer Aufgabe besser zu werden, indem Erfahrungen aus einem (Training-)Datensatz gesammelt werden. Die Güte des Algorithmus für maschinelles Lernen wird durch Leistungsmessung bewertet. In dieser Arbeit wurden zwei Kategorien des maschinellen Lernens eingeführt, nämlich: überwachtes und unüberwachtes Lernen. Ferner wurden für das überwachte Lernen Klassifikationsmodelle wie SVM und LDA sowie ein Regressionsmodelle wie die multivariate Regression vorgestellt. In ähnlicher Weise wurden für das unüberwachte Lernen das K-Means-Clustering und die Hauptkomponententransformation (PCA) erläutert. Schließlich wurden in Kapitel 3 verschiedene Limitierungen des maschinellen Lernens angesprochen. Eine der Limitierungen von klassischen maschinellen Lernverfahren ist die Notwendigkeit eine Merkmalsextraktion vor der Konstruktion eines Klassifikations- oder Regressionsmodells durchzuführen. Da der Schritt der Merkmalsextraktion subjektiv ist und von einem Experten abhängt, ist die Leistung des maschinellen Lernens von den extrahierten Merkmalen abhängig. Zu diesem Zweck werden Algorithmen benötigt, die Merkmale automatisch aus den Daten extrahieren können. Solche Algorithmen werden in einen Unterbereich des maschinellen Lernens kategorisiert, der als tiefes Lernen bezeichnet wird.

Algorithmen des tiefen Lernens verwenden zum Beispiel künstliche neuronale Netze (ANNs), die durch das menschliche neuronale System motiviert sind. Kurz gesagt, ein ANN umfasst Eingabe-, Ausgabe- und verborgene Schichten. Diese Schichten setzen sich aus grundlegenden Recheneinheiten zusammen, die Neuronen oder Knoten genannt werden. Jedes Neuron empfängt Eingangssignale von den Neuronen der vorhergehenden Schicht und leitet das Ausgangssignal an die Neuronen der folgenden Schichten weiter. Das Ausgangssignal eines Neurons wird durch eine Aktivierungs-

funktion nichtlinear transformiert. Kapitel 3 bietet eine mathematische Einführung in ANNs und weitere tiefe Lernmodelle wie neuronalen Faltungsnetzwerke, Autoencoder und Generative Adversarial Networks (GANs). Unter Kenntnis der vorgestellten KI Teilgebiete werden die in dieser Arbeit vorgestellten Anwendungen des maschinellen Lernens und des tiefen Lernens im Folgenden zusammengefasst.

- Die Bildklassifikation von nichtlinearen multimodalen Bildern wurde mit Hilfe des maschinellen Lernens untersucht. Diese Studie wurde durchgeführt, um eine frühe Sepsis in Leberabschnitten von Mäusen zu identifizieren. Im Rahmen des maschinellen Lernens wurden statistische Merkmale basierend auf dem Histogramm der nichtlinearen multimodalen Bilder extrahiert. Basierend auf den statistischen Merkmalen wurde ein lineares Klassifikationsmodell (PCA-LDA-Modell) trainiert. Das PCA-Modell wurde verwendet, um die Dimensionsgröße der statistischen Merkmale zu reduzieren, und das LDA-Modell wurde verwendet, um diese Merkmale in zwei Gruppen zu klassifizieren: Sepsis und Kontrolle. Um den Beitrag der einzelnen Modalitäten (CARS, TPEF und SHG) des nichtlinearen multimodalen Bildes zu verstehen, wurden außerdem die statistischen Merkmale aus jeder Modalität separat analysiert. Basierend auf den statistischen Merkmalen, die aus den einzelnen Modalitäten extrahiert wurden, konnte wieder ein PCA-LDA-Modell für die Klassifikationsaufgabe konstruiert werden. Die Ergebnisse dieser Studie zeigten, dass die aus CARS und TPEF extrahierten Merkmale im Vergleich zu den aus der SHG-Modalität extrahierten Merkmalen eine bessere Klassifizierungsleistung erzielten (siehe Abschnitt 4.1).
- In einer weiteren Arbeit wurde eine semantische Bildsegmentierung mit Hilfe von maschinellen Lernverfahren und tiefen Lernansätzen untersucht, um Regionen in nicht-linearen multimodalen Bildern zu identifizieren, welche für die Charakterisierung von entzündlichen Darmerkrankungen wichtig sind. Die Grundidee der semantischen Bildsegmentierung besteht darin, Pixel anstelle von Bildern zu klassifizieren. Die Klassifizierung von Pixeln wurde mit Hilfe von maschinellen Lernverfahren und tiefen Lernansätzen durchgeführt. Beim maschinellen Lernen wurden statistische Merkmale für die Pixel extrahiert, und ein auf diesen Merkmalen basierendes lineares Klassifikationsmodell (PCA-LDA) wurde konstruiert. Das PCA-Modell wurde verwendet, um die Größe der Dimension zu reduzieren, und das LDA-Modell wurde verwendet, um die Pixel in vier Regionen zu klassifizieren: Hintergrund, Schleimhaut ohne Krypten, Krypten und Nicht-Schleimhaut-Gewebe. Beim maschinellen Lernen wurde die Hintergrundregion mit einem unüberwachten Lernalgorithmus (K-Means-Clustering) extrahiert. Beim Ansatz des tiefen Lernens wurde dagegen ein Autoencoder-Modell (SegNet-Modell [111]) zur Segmentierung von nicht-linearen multimodalen Bildern in die beschriebenen vier Regionen verwendet. Ein Merkmalsextraktionsschritt wurde hier nicht durchgeführt, sondern das tiefe Ler-

nen konstruierte die Merkmale basierend auf dem Datensatz selbst. Schließlich wurde beim Vergleich der Ergebnisse aus den Modellen des maschinellen Lernens und des tiefen Lernens festgestellt, dass das SegNet-Modell besser abschneidet als das PCA-LDA-Modell (siehe Abschnitt 4.2).

- Die Bildübersetzung, wie sie in dieser Arbeit verwendet wird, wandelt ein nichtlineares multimodales Bild in ein histologisch gefärbtes H&E-Bild um. Eine solche Transformation ist vorteilhaft, da für die virtuelle Färbung der nichtlinearen multimodalen Bilder keine konventionelle H&E-Färbung im Labor durchgeführt werden muss, wodurch der Aufwand für den Pathologen reduziert wird. Hier wird die Bildtransformation mit Hilfe eines tiefen Lernmodells (GANs) durchgeführt (siehe Abschnitt 3.3.2). Das GAN wurde in einem beaufsichtigten und einem unbeaufsichtigten Ansatz verwendet. Das überwachte GAN-Modell (das Pix2Pix-Modell) erforderte ein entsprechendes Paar aus einem nicht-linearen multimodalen Bild und einem H&E-gefärbten Bild. Der Ansatz des unüberwachten Lernens (cycle-GAN-Modell) erforderte jedoch keine gepaarten Bilder. Der unüberwachte Ansatz reduzierte den Aufwand für die Bildregistrierung, die ein entscheidender Schritt beim überwachten Ansatz ist. Darüber hinaus war die unüberwachte Methode aufgrund ihrer Fähigkeit zur Rekonstruktion der ursprünglichen nicht-linearen multimodalen Bilder vorteilhaft (siehe Abschnitt 4.3).
- Die Spektrenklassifizierung von Raman-spektroskopischen Daten mit Hilfe eines tiefen Lernmodells wurde in dieser Arbeit untersucht. Hier wurde ein eindimensionales tiefes Faltungsnetzwerk (1D-CNN) vorgestellt. Das 1D-CNN-Modell wurde für eine Klassifikationsaufgabe verwendet, bei der vier Mayo-Scores von Patienten mit Colitis Ulcerosa separiert werden sollten. In ähnlicher Weise wurde das 1D-CNN-Modell für eine Regressionsaufgabe zur Erkennung von Patienten zwischen zwei Scores und zur Quantifizierung des Ausmaßes der Fehlklassifikation verwendet. Sowohl bei der Klassifikations- als auch bei der Regressionsaufgabe war die Eingabe ein Raman-Spektrum, das durch zwei verborgene Schichten geführt wurde. Das Ergebnis waren entweder Wahrscheinlichkeiten für die Klassifikationsaufgabe oder eine reell-wertige Zahl für die Regressionsaufgabe. Die Ergebnisse waren vielversprechend und die Krankheitsstadien der Colitis Ulcerosa konnten gut charakterisiert werden (siehe Abschnitt 4.4).
- Die Interpretation von KI-Modellen und das Verständnis der KI-Vorhersage war ein weiterer Beitrag in dieser Arbeit. Zu diesem Zweck wurde das für die Spektrenklassifikationsaufgabe verwendete 1D-CNN-Modell verwendet. Die nicht-lineare Funktion des trainierten 1D-CNN-Modells wurde mit Hilfe einer Taylor-expansion erster Ordnung interpretiert. Die Taylorentwicklung erster Ordnung wurde als Gewichtung der Raman-Spektralbänder verwendet, um biologisch relevante Raman-Banden für die Mayo-Score-Klassifizierung zu untersuchen. Eine

Raman-Bande mit hoher Gewichtung wurde so interpretiert, dass sie die Klassifikation eines Mayo-Scores stark beeinflusst, während eine Raman-Bande mit niedriger Gewichtung die Vorhersage des entsprechenden Mayo-Scores weniger wahrscheinlich beeinflusst (siehe Abschnitt 4.4).

- In dieser Arbeit wurde das Transferlernen tiefer Lernmodelle unter Verwendung histologischer und immunhistochemischer Bilddaten untersucht. Die Kombination von Histologie und Immunhistochemie wurde durchgeführt, um eine besser Brustkrebsdiagnose zu erreichen. Das Transferlernen von Modellen des tiefen Lernens wurde aufgrund der geringen Datensatzgröße von Histologie- und Immunhistochemie-Bildern untersucht. Dabei wurden zwei Transferlernstrategien untersucht. Bei der ersten Strategie wurden die tiefen Lernmodelle verwendet, um Merkmale für die beiden Bildgebungsdatensätze zu extrahieren. In der zweiten Strategie wurden die Gewichte der tiefen Lernmodelle optimiert, um die besten Klassifikationsergebnisse zu erhalten. Die Ergebnisse zeigten, dass die erste Transferlernstrategie (d.h. die Verwendung von tiefen Lernmodellen als Merkmalsextraktor) für den hier vorgestellten Brustkrebs-Datensatz effizient funktioniert (siehe Abschnitt 4.5).

Da es unmöglich ist alle Anwendungen des tiefen Lernens zu präsentieren, werden diese Modelle zusammen mit anderen möglichen Anwendungen in Kapitel 7 vorgestellt.

7

Future research directions

Up to this point, this thesis has presented numerous applications of AI models, particularly deep learning models like a convolutional neural network (1D and 2D), autoencoders, and generative adversarial networks for analyzing biophotonic data. It is ambitious to explore all variations of deep learning models for the data analysis purpose. Thus, this chapter presents other applications of deep learning models that were not explored in this thesis. This section will answer the question, namely: “*What’s next in biophotonic data analysis?*”. The potential applications will be explained for Raman spectroscopic and NLM imaging data; however, these applications can be valid for analogous biophotonic data.

The first issue is potential applications of deep learning models for Raman spectroscopic data, such as deep convolutional neural networks or autoencoders that have been used thus far in this thesis for image or spectra classification and image segmentation (see chapter 4). However, the deep convolutional neural networks can also be used to pre-process Raman spectroscopic data, including despiking, baseline correction, and calibration. Similarly, one of the advanced deep neural networks like the “recurrent neural network” (RNN) can also be investigated for the pre-processing of Raman spectra. This can be possible due to the inherent nature of RNNs (i.e., efficiently analyzing a sequential dataset) [24]. Using AEs, CNNs, or RNNs for spectra pre-processing requires further investigation. In addition to spectra pre-processing, DL models like AEs that were traditionally known for feature extraction can also be used for extracting important spectral features from a Raman spectral dataset instead of using a conventional PCA model. The effects of feature extraction by AEs and PCA can be comparatively studied in the future. Furthermore, adversarial

networks used for image translation in chapter 4 can also be beneficial for a spectral dataset. For instance, its application for generating more samples of Raman spectra (i.e., “data augmentation”) is worth investigating. Data augmentation has not been extensively used for a spectroscopic dataset until now, as its capability to fully represent the original dataset is debatable. However, a systematic investigation in the direction of data augmentation is still required. Lastly, transfer learning for analyzing small spectroscopic datasets based on models trained on a large dataset also requires investigation. Transfer learning for a spectroscopic dataset can be beneficial when the two datasets have (almost) similar spectral signatures; thus, the effect of transfer learning using the spectroscopic datasets should be investigated. The reference [24] provides more applications of DL models for analyzing spectroscopic data.

In the previous paragraph, applications for Raman spectroscopic data were mentioned. This paragraph addresses the use of various DL models for NLM images. It was already mentioned in chapter 4: the use of DL models like CNNs, AEs, and GANs for tasks like image classification, image segmentation, and image translation. However, the DL models were never used for pre-processing the NLM images. Therefore, the pre-processing of NLM images like removing unwanted noise, correcting stitching artefacts in a CARS signal, or increasing spatial resolution using the DL models can be explored in the future [24]. For performing pre-processing tasks, generative adversarial networks can play an important role. Furthermore, GANs can also be a part of the data augmentation of NLM images similar to a spectral dataset. Additionally, RNNs mentioned for spectral data analysis can also be used for the image classification and segmentation of NLM images. The use of RNNs in classification or segmentation tasks can be beneficial due to the property of RNNs to work efficiently with sequential data. Other DL models like conditional random fields (CRFs) [113] have applications in structure prediction. Therefore, CRFs for the post processing of NLM images also require systematic investigation [24].

After mentioning the potential applications of DL models for Raman spectroscopic and NLM imaging data, this section will be concluded by addressing the future direction of AI for biophotonic data analysis from a clinical perspective. One of the future applications is utilizing these models in clinics and hospitals, which is also one of the primary motivations of this thesis. This requires incorporating the DL models directly into medical (e.g., endoscopic) units or devices for obtaining real-time decision systems. Lastly, developing a plugin for openly available image processing tools which can preprocess, analyze, translate, and augment biophotonic data using AI models is also one of the future research directions.

Bibliography

- [1] J. Popp, V. Tuchin, A. Chiou, and S.H. Heinemann. *Handbook of biophotonics: Photonics for health care*, volume 2. John Wiley & Sons, USA, 2011.
- [2] R. R. Jones, D. C. Hooper, L. Zhang, D. Wolverson, and V. K. Valev. Raman techniques: fundamentals and frontiers. *Nanoscale research letters*, 14(1):1–34, 2019.
- [3] R. L. McCreery. *Raman spectroscopy for chemical analysis*, volume 225. John Wiley & Sons, USA, 2005.
- [4] J.R. Ferraro. *Introductory raman spectroscopy*. Elsevier, USA, 2003.
- [5] A. Zumbusch, G. R. Holtom, and X. S. Xie. Three-dimensional vibrational imaging by coherent anti-stokes raman scattering. *Physical review letters*, 82(20):4142, 1999.
- [6] T. Meyer. *Towards multimodal nonlinear microscopy in clinics*. PhD thesis, Friedrich-Schiller-Universität Jena, 2013.
- [7] G. Weber. Enumeration of components in complex systems by fluorescence spectrophotometry. *Nature*, 190(4770):27–29, 1961.
- [8] J. Popp. *Ex-vivo and In-vivo Optical Molecular Pathology*. John Wiley & Sons, USA, 2014.
- [9] T. W. Bocklitz, S. Guo, O. Ryabchykov, N. Vogler, and J. Popp. Raman based molecular imaging and analytics: a magic bullet for biomedical applications!? *Analytical chemistry*, 88(1):133–151, 2016.
- [10] S. Guo. *Chemometrics and Statistical Analysis in Raman Spectroscopy-based Biological Investigations*. PhD thesis, Friedrich-Schiller-Universität Jena, 2018.
- [11] R. G. Brereton. *Chemometrics for pattern recognition*. John Wiley & Sons, USA, 2009.

- [12] G. Quintás, N. Portillo, C. Juan García-Cañaveras, J. V. Castell, A. Ferrer, and A. Lahoz. Chemometric approaches to improve plsda model outcome for predicting human non-alcoholic fatty liver disease using uplc-ms as a metabolic profiling tool. *Metabolomics*, 8(1):86–98, 2012.
- [13] M. Fan, X. Liu, X. Yu, X. Cui, W. Cai, and X. Shao. Near-infrared spectroscopy and chemometric modelling for rapid diagnosis of kidney disease. *Science China Chemistry*, 60(2):299–304, 2017.
- [14] U. Neugebauer, T. W. Bocklitz, JH. Clement, C. Krafft, and J. Popp. Towards detection and identification of circulating tumour cells using raman spectroscopy. *Analyst*, 135(12):3178–3182, 2010.
- [15] A. Taleb, J. Diamond, J. J. McGarvey, J. R. Beattie, C. Toland, and P.W. Hamilton. Raman microscopy for the chemometric analysis of tumor cells. *The Journal of Physical Chemistry B*, 110(39):19625–19631, 2006.
- [16] C. Bielecki, C. Marquardt, A. Stallmach, T.W. Bocklitz, M. Schmitt, J. Popp, C. Krafft, A. Gharbi, and T. Knosel. Classification of inflammatory bowel diseases by means of raman spectroscopic imaging of epithelium cells. *Journal of biomedical optics*, 17(7):076030, 2012.
- [17] J. R. Baena and B. Lendl. Raman spectroscopy in chemical bioanalysis. *Current opinion in chemical biology*, 8(5):534–539, 2004.
- [18] I.M. Cockburn, R. Henderson, and S. Stern. The impact of artificial intelligence on innovation. Technical report, National bureau of economic research, 2018.
- [19] S. AD. Popenici and S. Kerr. Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 12(1):22, 2017.
- [20] M. Skilton and F. Hovsepian. *The 4th industrial revolution: Responding to the impact of artificial intelligence on business*. Palgrave Macmillan, London, UK, 2017.
- [21] R. Lindsay, B. Buchanan, E. Feigenbaum, and J. Lederberg. Applications of artificial intelligence for organic chemistry. In *McGraw-Hill advanced computer science series*, 1980.
- [22] Z. Hippe. *Artificial intelligence in chemistry: structure elucidation and simulation of organic reactions*. Elsevier, USA, 2013.

- [23] K. C. Chu. Applications of artificial intelligence to chemistry. use of pattern recognition and cluster analysis to determine the pharmacological activity of some organic compounds. *Analytical chemistry*, 46(9):1181–1187, 1974.
- [24] P. Pradhan, S. Guo, O. Ryabchykov, J. Popp, and T. W. Bocklitz. Deep learning a boon for biophotonics? *Journal of Biophotonics*, page e201960186, 2020.
- [25] T. Kirchberger-Tolstik, P. Pradhan, M. Vieth, P. Grunert, J. Popp, T. W. Bocklitz, and A. Stallmach. Towards an interpretable classifier for characterization of endoscopic mayo scores in ulcerative colitis using raman spectroscopy. *Analytical Chemistry*, 0(0):null, 0. PMID: 32965101.
- [26] C. Krafft and S. Vergo. Biomedical applications of raman and infrared spectroscopy to diagnose tissues. *Journal of Spectroscopy*, 20(5-6):195–218, 2006.
- [27] E.E. Lawson, B.W. Barry, A.C. Williams, and HGM. Edwards. Biomedical applications of raman spectroscopy. *Journal of Raman Spectroscopy*, 28(2-3):111–117, 1997.
- [28] D. I. Ellis and R. Goodacre. Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and raman spectroscopy. *Analyst*, 131(8):875–885, 2006.
- [29] I. Nabiev, I. Chourpar, and M. Manfait. Applications of raman and surface-enhanced raman scattering spectroscopy in medicine. *Journal of Raman Spectroscopy*, 25(1):13–23, 1994.
- [30] D. Cialla-May, X.S. Zheng, K. Weber, and J. Popp. Recent progress in surface-enhanced raman spectroscopy for biological and biomedical applications: from cells to clinics. *Chemical Society Reviews*, 46(13):3945–3961, 2017.
- [31] E.E. Lawson, H.G.M. Edwards, A.C. Williams, and B.W. Barry. Applications of raman spectroscopy to skin research: Reviewarticle. *Skin Research and Technology*, 3(3):147–153, 1997.
- [32] S. Luo, C. Chen, H. Mao, and S. Jin. Discrimination of premalignant lesions and cancer tissues from normal gastric tissues using raman spectroscopy. *Journal of biomedical optics*, 18(6):067004, 2013.
- [33] S.P. Singh, A. Deshmukh, P. Chaturvedi, and M. C. Krishna. In vivo raman spectroscopic identification of premalignant lesions in oral buccal mucosa. *Journal of biomedical optics*, 17(10):105002, 2012.

- [34] M. Jermyn, K. Mok, J. Mercier, J. Desroches, J. Pichette, K. Saint-Arnaud, L. Bernstein, M-C. Guiot, K. Petrecca, and F. Leblond. Intraoperative brain cancer detection with raman spectroscopy in humans. *Science translational medicine*, 7(274):274ra19–274ra19, 2015.
- [35] K. Kong, C. Kendall, N. Stone, and I. Notinger. Raman spectroscopy for medical diagnostics—from in-vitro biofluid assays to in-vivo cancer detection. *Advanced drug delivery reviews*, 89:121–134, 2015.
- [36] S. Feng, R. Chen, J. Lin, J. Pan, G. Chen, Y. Li, M. Cheng, Z. Huang, J. Chen, and H. Zeng. Nasopharyngeal cancer detection based on blood plasma surface-enhanced raman spectroscopy and multivariate analysis. *Biosensors and Bioelectronics*, 25(11):2414–2419, 2010.
- [37] D. Lin, S. Feng, J. Pan, Y. Chen, J. Lin, G. Chen, S. Xie, H. Zeng, and R. Chen. Colorectal cancer detection by gold nanoparticle based surface-enhanced raman spectroscopy of blood serum and statistical analysis. *Optics express*, 19(14):13565–13577, 2011.
- [38] R. Cicchi, S. Sestini, V. De Giorgi, D. Massi, T. Lotti, and F. S. Pavone. Nonlinear laser imaging of skin lesions. *Journal of Biophotonics*, 1(1):62–73, 2008.
- [39] R. Cicchi, N. Vogler, D. Kapsokalyvas, B. Dietzek, J. Popp, and F. S. Pavone. From molecular structure to tissue architecture: collagen organization probed by shg microscopy. *Journal of biophotonics*, 6(2):129–142, 2013.
- [40] G. Bautista and M. Kauranen. Vector-field nonlinear microscopy of nanostructures. *ACS Photonics*, 3(8):1351–1370, 2016.
- [41] M. D. Duncan, J. Reintjes, and T.J. Manuccia. Scanning coherent anti-stokes raman microscope. *Optics letters*, 7(8):350–352, 1982.
- [42] C. Krafft, B. Dietzek, J. Popp, and M. Schmitt. Raman and coherent anti-stokes raman scattering microspectroscopy for biomedical applications. *Journal of biomedical optics*, 17(4):040801, 2012.
- [43] T. Gottschall, T. Meyer, M. Baumgartl, C. Jauregui, M. Schmitt, J. Popp, J. Limpert, and A. Tünnermann. Fiber-based light sources for biomedical applications of coherent anti-stokes raman scattering microscopy. *Laser & Photonics Reviews*, 9(5):435–451, 2015.
- [44] C. L. Evans and X. S. Xie. Coherent anti-stokes raman scattering microscopy: chemical imaging for biology and medicine. *Annu. Rev. Anal. Chem.*, 1:883–909, 2008.

- [45] L.C. Evans, E.O. Potma, M. Puoris' haag, D. Côté, C.P. Lin, and X. S. Xie. Chemical imaging of tissue in vivo with video-rate coherent anti-stokes raman scattering microscopy. *Proceedings of the national academy of sciences*, 102(46):16807–16812, 2005.
- [46] B. Rakic, S. M. Sagan, M. Noestheden, S. Bélanger, X. Nan, C. L. Evans, X. S. Xie, and J. P. Pezacki. Peroxisome proliferator-activated receptor α antagonism inhibits hepatitis c virus replication. *Chemistry & biology*, 13(1):23–30, 2006.
- [47] X. Nan, A. M. Tonary, A. Stolow, X. S. Xie, and J.P. Pezacki. Intracellular imaging of hcv rna and cellular lipids by using simultaneous two-photon fluorescence and coherent anti-stokes raman scattering microscopies. *ChemBioChem*, 7(12):1895–1897, 2006.
- [48] M. Schliwa and G. Woehlke. Molecular motors. *Nature*, 422(6933):759–765, 2003.
- [49] R. D. Vale. The molecular motor toolbox for intracellular transport. *Cell*, 112(4):467–480, 2003.
- [50] N. Vogler, S. Heuke, T. W. Bocklitz, M. Schmittl, and J. Popp. Multimodal imaging spectroscopy of tissue. *Annual Review of Analytical Chemistry*, 8:359–387, 2015.
- [51] W. Denk, J. H. Strickler, and W. W. Webb. Two-photon laser scanning fluorescence microscopy. *Science*, 248(4951):73–76, 1990.
- [52] C. Xu, W. Zipfel, J. B. Shear, R. M. Williams, and W. W. Webb. Multiphoton fluorescence excitation: new spectral windows for biological nonlinear microscopy. *Proceedings of the National Academy of Sciences*, 93(20):10763–10768, 1996.
- [53] A. Volkmer, V. Subramaniam, D. J. S. Birch, and T. M. Jovin. One-and two-photon excited fluorescence lifetimes and anisotropy decays of green fluorescent proteins. *Biophysical journal*, 78(3):1589–1598, 2000.
- [54] S. Huang, A. A. Heikal, and W. W. Webb. Two-photon fluorescence spectroscopy and microscopy of nad (p) h and flavoprotein. *Biophysical journal*, 82(5):2811–2825, 2002.
- [55] L. L. Hsu, S. B. Pelet, T. M. Hancewicz, P.D. Kaplan, and P. TC. So. Two-photon 3-d mapping of ex vivo human skin endogenous fluorescence species based on fluorescence emission spectra. *Journal of biomedical optics*, 10(2):024016, 2005.

- [56] B. AI. van den Bergh, J. Vroom, H. Gerritsen, H. E. Junginger, and J. A. Bouwstra. Interactions of elastic and rigid vesicles with human skin in vitro: electron microscopy and two-photon excitation microscopy. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1461(1):155–173, 1999.
- [57] B. R. Masters and P. TC. So. Confocal microscopy and multi-photon excitation microscopy of human skin in vivo. *Optics express*, 8(1):2–10, 2001.
- [58] H. Zeng, C. MacAulay, B. Palcic, and D. I. McLean. Spectroscopic and microscopic characteristics of human skin autofluorescence emission. *Photochemistry and photobiology*, 61(6):639–645, 1995.
- [59] F. Pan and W-B. Gan. Two-photon imaging of dendritic spine development in the mouse cortex. *Developmental neurobiology*, 68(6):771–778, 2008.
- [60] F. S. Pavone and P. J. Campagnola. *Second harmonic generation imaging*. CRC Press, UK, 2016.
- [61] A. Keikhosravi, J. S. Bredfeldt, A. K. Sagar, and K. W. Eliceiri. Second-harmonic generation imaging of cancer. In *Methods in cell biology*, volume 123, pages 531–546. Elsevier, 2014.
- [62] W. R. Zipfel, R. M. Williams, R. Christie, A. Y Nikitin, B. T. Hyman, and W. W. Webb. Live tissue intrinsic emission microscopy using multiphoton-excited native fluorescence and second harmonic generation. *Proceedings of the National Academy of Sciences*, 100(12):7075–7080, 2003.
- [63] A. Zoumi, A. Yeh, and B. J. Tromberg. Imaging cells and extracellular matrix in vivo by using second-harmonic generation and two-photon excited fluorescence. *Proceedings of the National Academy of Sciences*, 99(17):11014–11019, 2002.
- [64] F. B. Legesse, O. Chernavskaja, S. Heuke, T. W. Bocklitz, T. Meyer, J. Popp, and R. Heintzmann. Seamless stitching of tile scan microscope images. *Journal of microscopy*, 258(3):223–232, 2015.
- [65] S. D. Brown, T. B. Blank, S. T. Sum, and L. G. Weyer. Chemometrics. *Analytical chemistry*, 66(12):315–359, 1994.
- [66] T. W. Bocklitz, A. Walter, K. Hartmann, P. Rösch, and J. Popp. How to preprocess raman spectra for reliable and stable models? *Analytica chimica acta*, 704(1-2):47–56, 2011.
- [67] O. Chernavskaja, S. Guo, T. Meyer, N. Vogler, D. Akimov, S. Heuke, R. Heintzmann, T. W. Bocklitz, and J. Popp. Correction of mosaicking artifacts in

- multimodal images caused by uneven illumination. *Journal of Chemometrics*, 31(6):e2901, 2017.
- [68] T. W. Bocklitz, F. S. Salah, N. Vogler, S. Heuke, O. Chernavskaia, C. Schmidt, M. J. Waldner, F. R. Greten, R. Bräuer, M. Schmitt, et al. Pseudo-he images derived from cars/tpef/shg multimodal imaging in combination with raman-spectroscopy as a pathological screening tool. *BMC cancer*, 16(1):534, 2016.
- [69] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.
- [70] S. Russel, P. Norvig, et al. *Artificial intelligence: a modern approach*. Pearson Education Limited, UK, 2013.
- [71] C. M. Bishop. *Pattern recognition and machine learning*. Springer, USA, 2006.
- [72] J. Moore and N. Raghavachari. Artificial intelligence based approaches to identify molecular determinants of exceptional health and life span-an interdisciplinary workshop at the national institute on aging. *Frontiers in Artificial Intelligence*, 2, 08 2019.
- [73] T. M. Mitchell et al. *Machine learning*, 1997.
- [74] A. Smola and SVN. Vishwanathan. *Introduction to machine learning*. Cambridge University, UK, 32:34, 2008.
- [75] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, Cambridge, UK, 2016.
- [76] R. S. Sutton, A. G. Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press, Cambridge, UK, 1998.
- [77] B. Yoshua, C. Aaron, and V. Pascal. *Representation learning: A review and new perspectives*, 2012.
- [78] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.
- [79] D. T. Tran, M. Gabbouj, and A. Iosifidis. Multilinear class-specific discriminant analysis. *Pattern Recognition Letters*, 100:131–136, 2017.
- [80] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

- [81] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92*, page 144–152, New York, NY, USA, 1992. Association for Computing Machinery.
- [82] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, Germany, 2009.
- [83] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [84] P. Mason. *Medical neurobiology*. Oxford University Press, England, UK, 2017.
- [85] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.
- [86] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [87] D. H. Hubel and TN. Wiesel. Shape and arrangement of columns in cat’s striate cortex. *The Journal of physiology*, 165(3):559–568, 1963.
- [88] D. H. Ballard. Modular learning in neural networks. In *Proceedings of the Sixth National Conference on Artificial Intelligence - Volume 1, AAAI'87*, page 279–284. AAAI Press, 1987.
- [89] H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59:291–4, 02 1988.
- [90] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
- [91] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.
- [92] C. Fu, S. Lee, H. D. Joon, S. Han, P. Salama, K. W. Dunn, and E. J. Delp. Three dimensional fluorescence microscopy image synthesis and segmentation.

- In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2221–2229, 2018.
- [93] A. Osokin, A. Chessel, R. Carazo Salas, and F. Vaggi. Gans for biological image synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2233–2242, 2017.
- [94] Y. Rivenson, Zoltán Göröcs, H. Günaydin, Y. Zhang, H. Wang, and A. Ozcan. Deep learning microscopy. *Optica*, 4(11):1437–1443, Nov 2017.
- [95] H. Wang, Y. Rivenson, Y. Jin, Z. Wei, R. Gao, H. Günaydin, L. A. Bentolila, C. Kural, and A. Ozcan. Deep learning enables cross-modality super-resolution in fluorescence microscopy. *Nature Methods*, 16(1):103–110, 2019.
- [96] Y. Rivenson, H. Wang, Z. Wei, K. Haan, Y. Zhang, Y. Wu, H. Gunaydin, J. Zuckerman, T. Chong, A. Sisk, L. Westbrook, W. Wallace, and A. Ozcan. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nature Biomedical Engineering*, 3:466, 06 2019.
- [97] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [98] T. Tieleman and G. Hinton. Rmsprop gradient optimization. URL http://www.cs.toronto.edu/tijmen/csc321/slides/lecture_slides_lec6.pdf, 2014.
- [99] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.
- [100] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.
- [101] D. E. Rumelhart, J. L. McClelland, PDP Research Group, et al. Explorations in the microstructure of cognition. *Foundations*, 1:318–362, 1986.
- [102] L. Torrey and J. Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global, 2010.
- [103] N. Ali, E. Quansah, K. Köhler, T. Meyer, M. Schmitt, J. Popp, A. Niendorf, and T. W. Bocklitz. Automatic label-free detection of breast cancer using nonlinear multimodal imaging and the convolutional neural network resnet50. *Translational Biophotonics*, 1(1-2):e201900003, 2019.

- [104] H. Lin, C. Wei, G. Wang, H. Chen, L. Lin, M. Ni, J. Chen, and S. Zhuo. Automated classification of hepatocellular carcinoma differentiation using multiphoton microscopy and deep learning. *Journal of biophotonics*, page e201800435, 2019.
- [105] N. Singla, K. Dubey, and V. Srivastava. Automated assessment of breast cancer margin in optical coherence tomography images via pretrained convolutional neural network. *Journal of biophotonics*, 12(3):e201800255, 2019.
- [106] S. Weng, X. Xu, J. Li, and S. TC. Wong. Combining deep learning and coherent anti-stokes raman scattering imaging for automated differential diagnosis of lung cancer. *Journal of biomedical optics*, 22(10):106017, 2017.
- [107] M. Yarbakht, Pranita P. Pradhan, N. Köse-Vogey, H. Bae, S. Stengel, T. Meyer, M. Schmitt, A. Stallmach, J. Poppn, T. W. Bocklitz, et al. Nonlinear multimodal imaging characteristics of early septic liver injury in a mouse model of peritonitis. *Analytical chemistry*, 91(17):11116–11121, 2019.
- [108] R. C. Gonzalez, R. E. Woods, and S. L. Eddins. *Digital image processing using MATLAB*. Pearson Education India, 2004.
- [109] S. Guo, T. W. Bocklitz, U. Neugebauer, and J. Popp. Common mistakes in cross-validating classification models. *Analytical methods*, 9(30):4410–4417, 2017.
- [110] P. Pradhan, T. Meyer, M. Vieth, A. Stallmach, M. Waldner, M. Schmitt, J. Popp, and T. W. Bocklitz. Semantic segmentation of non-linear multimodal images for disease grading of inflammatory bowel disease: A segnet-based application. In *International Conference on Pattern Recognition Applications and Methods 2019*, 2019.
- [111] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.
- [112] B. Buchanan and G. Sutherland. Heuristic dendral: A program for generating explanatory hypotheses in organic chemistry. Technical report, Stanford University California Department of Computer Science, 1968.
- [113] C. Sutton, A. McCallum, et al. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, 2012.

8

Publications, manuscripts and conference
proceeding

P1 TOWARDS AN INTERPRETABLE CLASSIFIER FOR CHARACTERIZATION OF EN-
 DOSCOPIC MAYO SCORES IN ULCERATIVE COLITIS USING RAMAN SPEC-
 TROSCOPY

Reprinted with permission from [T. Kirchberger-Tolstik, P. Pradhan, M. Vieth, P. Grunert, J. Popp, T. W. Bocklitz, A. Stallmach, Towards an interpretable classifier for characterization of endoscopic Mayo scores in ulcerative colitis using Raman Spectroscopy, 2020, *Analytical Chemistry*, ACS Publications, Washington]. Copyright 2020 American Chemical Society.

The declared individual contributions of the doctoral candidate and the other doctoral candidates participate as co-authors in the publications are listed below.

T. Kirchberger-Tolstik ¹ , P. Pradhan ² , M. Vieth ³ , P. Grunert ⁴ , J. Popp ⁵ , T. W. Bocklitz ⁶ , A. Stallmach ⁷ , Towards an interpretable classifier for characterization of endoscopic Mayo scores in ulcerative colitis using Raman Spectroscopy, 2020, <i>Analytical Chemistry</i> , ACS Publications, Washington.							
Involved in (Please tick the boxes that apply.)							
	1	2	3	4	5	6	7
Conceptual research design			X		X	X	X
Planning of research activities	X	X			X	X	X
Data collection	X			X			X
Data analysis and interpretation	X	X				X	X
Manuscript writing	X	X	X	X	X	X	X
Suggested publication equivalence value		1.0					

Towards an Interpretable Classifier for Characterization of Endoscopic Mayo Scores in Ulcerative Colitis Using Raman Spectroscopy

Tatiana Kirchberger-Tolstik, Pranita Pradhan, Michael Vieth, Philip Grunert, Juergen Popp,*
Thomas Wilhelm Bocklitz,* and Andreas Stallmach*

Cite This: <https://dx.doi.org/10.1021/acs.analchem.0c02163>

Read Online

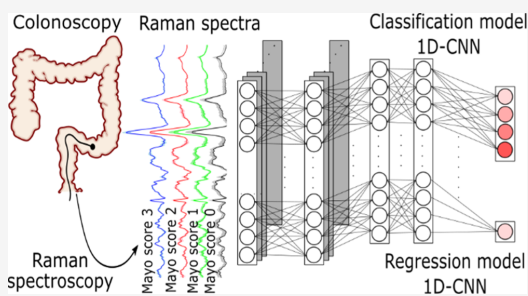
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: Ulcerative colitis (UC) is one of the main types of chronic inflammatory diseases that affect the bowel, but its pathogenesis is yet to be completely defined. Assessing the disease activity of UC is vital for developing a personalized treatment. Conventionally, the assessment of UC is performed by colonoscopy and histopathology. However, conventional methods fail to retain biomolecular information associated to the severity of UC and are solely based on morphological characteristics of the inflamed colon. Furthermore, assessing endoscopic disease severity is limited by the requirement for experienced human reviewers. Therefore, this work presents a nondestructive biospectroscopic technique, for example, Raman spectroscopy, for assessing endoscopic disease severity according to the four-level Mayo subscore. This contribution utilizes multidimensional Raman spectroscopic data to generate a predictive model for identifying colonic inflammation. The predictive modeling of the Raman spectroscopic data is performed using a one-dimensional deep convolutional neural network (1D-CNN). The classification results of 1D-CNN achieved a mean sensitivity of 78% and a mean specificity of 93% for the four Mayo endoscopic scores. Furthermore, the results of the 1D-CNN are interpreted by a first-order Taylor expansion, which extracts the Raman bands important for classification. Additionally, a regression model of the 1D-CNN model is constructed to study the extent of misclassification and border-line patients. The overall results of Raman spectroscopy with 1D-CNN as a classification and regression model show a good performance, and such a method can serve as a complementary method for UC analysis.



Ulcerative colitis (UC) is one of the main types of chronic inflammatory diseases of the gastrointestinal tract.¹ Its occurrence is most common in adults aged 30–40 years old with rising incidence worldwide.² Presently, there is no cure for UC; moreover, it causes patient disability with high relapse rates.³ The aim of UC management is to induce and maintain clinical and endoscopic remission, where careful medical attention is crucial to monitor and control the disease.² The characterization of inflammation is crucial to determine the severity of the disease and design a personalized treatment for UC patients. Commonly, the evaluation of UC severity during endoscopy can be made by different scoring systems. The most commonly used is the activity index which is based on the Mayo endoscopic score.⁴ Here, colon tissue can be classified during endoscopy into four categories: 0-normal, 1-mild, 2-moderate, and 3-severe.⁵ Nevertheless, a significant disagreement among endoscopists is seen for scoring the disease severity in UC patients which eventually affects the patient management.^{6,7} The disagreement among endoscopists is influenced by various factors such as inter- and inpatient

variability, the degree of expertise in inflammatory bowel disease (IBD) endoscopy, and other clinical factors. Therefore, a proper consensus to mitigate the disagreements among observers is crucial and efforts in this direction are currently performed. Because of the above-mentioned reasons, an automatic and interpretable algorithm for classification of the disease severity in UC patients based on biomolecular information is desired.

Over the past years, Raman spectroscopy has proven to be a promising technique for nondestructive, label-free characterization of IBD.^{8–13} The first statistical model for classification and prediction of patient tissue samples based on Raman

Received: May 20, 2020

Accepted: September 23, 2020

Published: September 23, 2020

spectroscopy was published by Bielecki *et al.* in 2012.⁸ This publication contained patients with UC ($n = 13$), Crohn's Disease (CD) ($n = 14$) and healthy controls ($n = 11$). These three groups showed significantly different molecular specific Raman signatures that allowed classification of both diseases against healthy controls with an accuracy of more than 98%. Moreover, the development of Raman-based fiber optic probes and its application for the diagnosis of IBD was published by Bi *et al.* in 2011.⁹ Here, a fiber probe system was used for fast investigation of UC and CD in tissue samples with detection of significant differences in phenylalanine, lipids, and nucleic acids.⁹ Furthermore, fiber optic probe-based Raman spectroscopy coupled to a colonoscope, as a minimally invasive diagnostics tool, was used for the first time *in vivo* for characterization of IBD in the colon by Pence *et al.* in 2017.¹⁰ However, none of the abovementioned work has characterized the inflammation stages in UC using Raman spectroscopy.

In this work, characterization of inflammation in UC patients based on four Mayo endoscopic subscores was achieved. For this purpose, Raman spectroscopic data and deep convolutional neural networks (DCNNs)¹⁴ were utilized. A predictive modeling was performed using a two-layered one-dimensional deep convolutional neural network (1D-CNN) such that the input to the 1D-CNN was a Raman spectrum and the output was a probability of the spectrum belonging to one of the four Mayo endoscopic subscores. This classification model was evaluated by performance metrics, including sensitivity, specificity, and risk factor. Furthermore, the same 1D-CNN was used as a regression model to investigate the extent of misclassification obtained using the 1D-CNN classification model. The 1D-CNN regression model was also used to investigate the border-line patients, for instance, a patient transforming from Mayo endoscopic subscore 2 to Mayo endoscopic subscore 3. Finally, the predictions obtained using the 1D-CNN classification model were interpreted using a first-order Taylor expansion. The interpretation of the 1D-CNN classification model was carried out to obtain important Raman bands in the Raman spectrum associated with each Mayo endoscopic subscore. To the author's knowledge, this is the first application of DCNN for Mayo score classification in UC patients and first-order Taylor expansion for interpretation of the Raman spectroscopic data.

In summary, the aim of our study was the improvement of UC diagnostics by applying Raman spectroscopy and DCNN for nondestructive and label-free characterization and classification of colon biopsy samples. We are able to demonstrate that Raman spectroscopy along with data analysis can serve as a promising method for label-free and nondestructive characterization of colon inflammation in UC.

MATERIALS AND METHODS

Tissue Collection and Selection. This research project was approved by the local medical ethics committee, and written consent was obtained from all patients. Our ethical approval number is 2158-11/07. In the Department of Internal Medicine IV, Division of Interdisciplinary Endoscopy at Jena University Hospital, colon biopsies were obtained from 140 patients diagnosed with UC during colonoscopy. The biopsy samples were taken, immediately shock frozen in liquid nitrogen and stored at $-80\text{ }^{\circ}\text{C}$. For the diagnosis, the biopsies were investigated during endoscopy and evaluated by experienced doctors from Department of Endoscopy from Jena University Hospital. Each tissue biopsy taken during the

endoscopy procedure was classified based on Mayo score classification: 0-normal or inactive disease, 1-mild disease with erythema, decreased vascular patterns, and mild friability, 2-moderate disease with marked erythema, the absence of vascular patterns, friability, and erosions, and 3-severe disease with spontaneous bleeding and ulceration.⁵ Afterward, the classification was confirmed by the head of the Endoscopy department. Based on time and resource limitations, around 10 biopsies for each of the four Mayo endoscopic scores were selected for Raman spectroscopic investigations. Table 1 shows the patient characteristics of all samples undergoing investigations together with classification scores.

Table 1. Patient Information Including Mayo Endoscopic Subscore, Gender, and Age of the Samples Chosen for This Study

Mayo score	0	1	2	3	total
		Gender			
female	7	4	8	4	23
male	3	7	4	5	19
		Age			
<20		1	2		3
20–29	2	3	3	2	10
30–39		2	3		5
40–49	2	5	2	6	15
50–59	2		1	1	4
60–69	1		1		2
>70	3				3

On the day of planned Raman measurements, a cryostat (Cryostat Leica 3050 S, Leica Biosystems, Germany) was used for tissue sectioning. One section was prepared with $20\text{ }\mu\text{m}$ thickness and placed onto a calcium fluoride slide (CaF₂; Vacuum-UV quality, Crystal GmbH, Berlin) for Raman measurements. No pretreatment of the tissue sections with any fixation solutions was performed in order to keep our measurement as close to the *in vivo* conditions as possible. Subsequently, a parallel tissue section for finding regions of interest (ROIs) and pathological classification was cut ($10\text{ }\mu\text{m}$ thickness) and placed on a glass slide. After cutting the parallel section, it was stained by Ab-Pas staining and investigated under a light microscope in order to find ROIs of the intestinal epithelium. According to the predefined regions, Raman spectroscopic imaging was performed on the $20\text{ }\mu\text{m}$ -thick section on a WITec Raman microscope (WITec, Ulm, Germany, Model CRM 2000). Although the tissue sections on CaF₂ slides remained intact, they were also stained with Ab-Pas and afterward ROIs were detected under the light microscope and used for correlation and further pathological diagnostics.

Tissue Sample Preparations. A confocal Raman microscope (WITec, Ulm, Germany, Model CRM 2000) with 300 lines/mm grating (blaze wavelength 750 nm) and a 785 nm diode laser as excitation was used for the collection of various Raman maps from each tissue sample. The laser light was focused with a 50 \times NA 0.95 objective (EC Epiplan-Apochromat, Zeiss, Germany) coupled to the microscope using a single-mode optical fiber. The scattered Raman signal was detected using a back-illuminated deep-depletion charged-couple device (CCD) camera operating at $-65\text{ }^{\circ}\text{C}$. Before each measurement, the Raman system was calibrated to 520.7 cm^{-1} spectral line of silicon. Moreover, a reference spectrum of

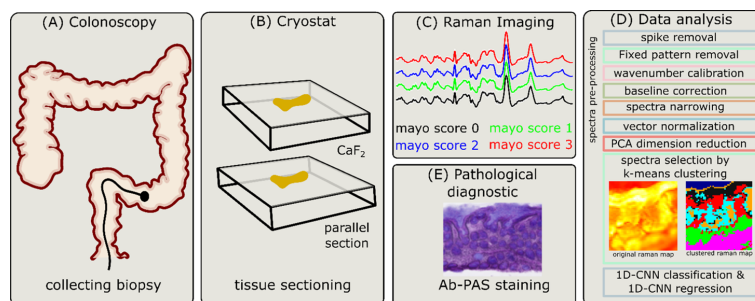


Figure 1. Experimental workflow including (A) collection of biopsies during colonoscopy, (B) tissue sectioning using a cryostat for spectroscopic measurements and Ab-Pas staining, (C) Raman spectroscopic mapping experiments showing four mean spectra associated to four Mayo endoscopic scores, (D) data analysis including spectral preprocessing, clustering, and classification, and (E) pathological diagnostics with the following correlation of Ab-Pas-stained tissue, Raman maps, and cluster analysis.

paracetamol (4-acetamidophenol, A3035, Sigma-Aldrich, Germany) was taken every day for wavenumber calibration during data analysis, in order to avoid artifacts from different measurement days. Raman spectral images of the preselected ROIs in the intestinal epithelium were acquired with a measurement area of $50 \times 50 \mu\text{m}^2$. This measurement area was selected based on the average size of the crypts in the epithelial layer in the human colon samples and kept constant during all measurements. The step size of $1 \mu\text{m}$ as a lateral resolution for the Raman maps and the spectral region of $200\text{--}3200 \text{ cm}^{-1}$ were set in the mapping mode of WITec Control. Prebleaching of the autofluorescence of 2 s was performed at each measured point to avoid the high autofluorescence of the tissue samples. The integration time for each Raman spectrum was kept 5 s. In summary, 227 Raman maps with 567,500 spectra were acquired and used for further data preprocessing and analysis.

Data Acquisition and Preprocessing. Prior to the construction of a classification model, the Raman spectral data were preprocessed using a standard procedure explained elsewhere.¹⁵ The preprocessing of Raman spectra was performed using R on a commercially available PC system Intel Core™ i5-7500 CPU, 3.40 GHz, 16 GB RAM.

Briefly, the spectra were despiked by removing unwanted spikes arising because of high-energy particles hitting the CCD. Furthermore, wavenumber calibration was performed using a standard spectrum of 4-acetamidophenol¹⁶ and the spectral baseline was corrected using an asymmetric least squares method.¹⁷ Thereafter, the spectral range was narrowed to $500\text{--}3020 \text{ cm}^{-1}$, and the silent region from 1800 to 2800 cm^{-1} was removed. Finally, the Raman spectra were normalized by vector normalization and the spectral dimension was reduced by principal component analysis¹⁸ using 50 principal components (PCs). Based on the reduced spectral dimension, a *k*-means clustering algorithm¹⁸ was applied to cluster the spectra using a distance-based similarity metric. Every cluster was color-coded to visualize the Raman map as a false-color plot and the mean spectrum for every cluster was calculated (Figure 1). To remove the cluster representing the spatial background or unwanted noise, the Euclidean distance, d , was calculated between the mean spectrum of each cluster and the mean spectrum of the whole Raman map. If the Euclidean distance, d , of the cluster mean spectra to the overall mean spectra was less than a threshold value (in this case, 0.37), the cluster mean spectra were used for further analysis. The

threshold value was optimized by a manual quality check of the cluster mean spectra. Mathematically, the Euclidean distance used for spectra selection can be given as

$$d(\bar{\mathbf{s}}, \mathbf{s}) = \sqrt{\sum_{i=1}^{\text{wn}} (\bar{s}_i - s_i)^2}$$

where $\bar{\mathbf{s}}$ is the mean spectrum of the whole Raman map, \mathbf{s} is the mean spectrum of a cluster, and i is an index indicating the wn wavenumber positions in the spectra. Subsequently, a mean spectrum was calculated from the mean spectra of the selected clusters such that each Raman map was represented by solely one mean Raman spectrum. In this way, 227 mean spectra were obtained, as the dataset comprised 227 Raman maps from 42 patients (Mayo score 0: 10 patients, Mayo score 1: 11 patients, Mayo score 2: 12 patients, and Mayo score 3: 9 patients) as mentioned earlier (Table 2). The dimension of Raman maps was $50 \times 50 \mu\text{m}^2$.

Table 2. Overview of the Dataset for All Four Mayo Endoscopic Subscores Is Given^a

Mayo score	# patients	# Raman maps	# raw spectra	# mean map spectra
0	10	76	190×10^3	76
1	11	56	140×10^3	56
2	12	44	110×10^3	44
3	9	51	127.5×10^3	51

^a# denotes number.

Classification and Regression Using 1D-CNN. The preprocessed spectra were classified into four Mayo endoscopic scores using a 1D-CNN (Figure 2). The 1D-CNN classification model comprised two convolution layers with a rectified linear unit activation layer¹⁹ and batch normalization layer.²⁰ The input dimension of the two convolution layers is 1520 (corresponding to 1520 wavenumbers) with 32 and 64 kernels of size 7, respectively. The convolution layers were followed by a flattened and a dropout layer with a 20% dropout rate.²¹ The number of kernels, kernel size, and the dropout rate for the 1D-CNN classification model were optimized by monitoring validation sensitivity during training. The last layer of the 1D-CNN was a dense layer of four neurons (corresponding to four Mayo endoscopic scores) with a softmax activation layer²² (Figure 2). The 1D-CNN

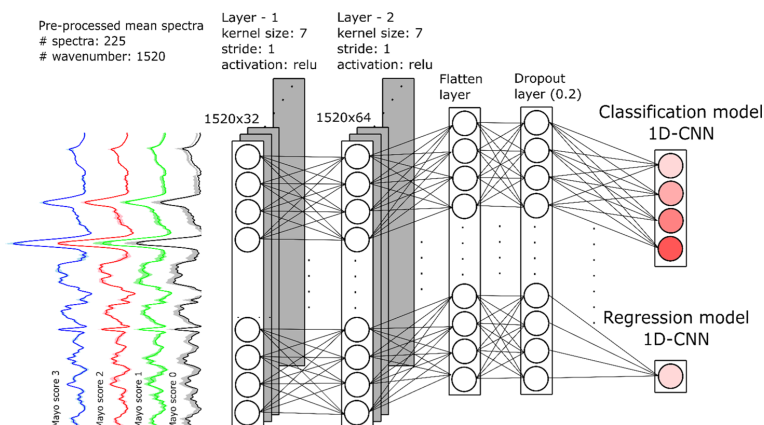


Figure 2. Combined 1D-CNN classification and 1D-CNN regression model is shown in this figure. The major difference between these two models is the last layer. For classification purpose, the 1D-CNN comprises four output neurons with softmax activation function and was trained using the categorical cross-entropy loss function. On the other hand, for regression purpose, the 1D-CNN has one output neuron with linear activation function and was trained using mean absolute error loss function. These two models also differed in their respective hyperparameter setting explained in the text.

Table 3. Evaluation of the 1D-CNN Classification Model Based on the Three Evaluation Metrics, Namely, Sensitivity, Specificity, and Risk Factor^a

pathological diagnosis	model prediction—1D CNN				evaluation metrics		
	Mayo score 0	Mayo score 1	Mayo score 2	Mayo score 3	sensitivity (%)	specificity (%)	risk-factor
Mayo score 0	10	0	0	0	100.00	87.50	1.44
Mayo score 1	2	9	0	0	81.82	100.00	
Mayo score 2	0	0	9	3	75.00	93.33	
Mayo score 3	2	0	2	5	55.56	90.90	

^aThe mean sensitivity, mean specificity, and risk factor of the 1D-CNN classification model are 78.09%, 92.93%, and 1.44, respectively.

classification model was trained using a leave-one-patient-out cross-validation strategy such that for every iteration, a patient was used as an independent test set and the remaining patients were used as the train set for the 1D-CNN classification model. For the model training, the hyperparameters including optimizer (stochastic gradient descent²³), learning rate (10^{-5}), batch size (5), and the number of epochs (300) were kept constant in all iterations. A categorical cross-entropy loss function was used to optimize the trainable parameters of the 1D-CNN classification model. The prediction of the 1D-CNN classification model on the independent test set was quantitatively evaluated using a confusion matrix, mean sensitivity, and mean specificity. The predictions of each of the 227 Raman maps were utilized for a voting scheme to obtain one Mayo subscore prediction per patient. The reported values in the confusion table (Table 3) are generated by majority voting of the Raman maps evaluated for 42 patients. Additionally, the 1D-CNN classification model was evaluated using a new metric called the risk factor.

The risk factor was used to assess the severity of the misclassification. The risk factor is higher if a patient with Mayo score 3 is predicted as Mayo score 0, whereas the risk factor is lower if a patient with Mayo score 3 is predicted as Mayo score 2. The risk factor is a weighted mean absolute error which largely penalizes the model for severe misclassifications. Mathematically, it can be given as

$$wMAE = \frac{\sum_{i=0}^T \sum_{j \neq i}^P w_{ij} |i - j|}{\sum_{i=0}^T \sum_{j \neq i}^P w_{ij}}$$

where $T, P \in [0, 3]$ is the total number of Mayo scores and w_{ij} is the number of patients with i and j as the true and predicted Mayo score. The average of the mean sensitivities, mean specificities, and risk factor computed over all the 42 patients is reported in Table 3.

The misclassified patients obtained using the 1D-CNN classification model were further analyzed by constructing a 1D-CNN regression model. The regression analysis was used to study the extent of misclassification generated using the 1D-CNN classification model and to characterize the border-line patients. For this purpose, the last layer of the 1D-CNN classification model was modified using one output neuron with a linear activation function²⁴ instead of the softmax activation function. The training of the 1D-CNN regression model was achieved using a mean-squared error loss function. Like the 1D-CNN classification model, the 1D-CNN regression model was validated using a leave-one-patient-out cross-validation strategy. The hyperparameters for the 1D-CNN regression model were Adam optimizer²⁵ with learning rate 10^{-5} , batch size of 5 spectra, and 300 epochs.

The training of both the 1D-CNN classification and regression model was performed using Python with packages including sklearn,¹⁷ Numpy,²⁶ Rpy2, Tensorflow,²⁷ and Keras²⁸ on a commercially available PC system with NVIDIA

GeForce GTX 1060, 6 GB. The total model training time was approximately 2 h for each classification and regression model within the cross-validation loop. The prediction of one single spectrum required to extract a prediction for a patient is 3–5 s on the utilized hardware.

Interpretation of the 1D-CNN Classification Model.

Interpretation of DCNNs is a challenging task. In this work, saliency maps²⁹ were used to interpret the predictions of a particular Mayo endoscopic score based on the wavenumber of the finger-print region (500–1800 cm^{-1}). The wavenumbers were ranked for a mean spectrum \bar{s} of each Mayo score m by approximating the nonlinear function $Y_m(\bar{s})$ of the 1D-CNN classification model using the first-order Taylor expansion.³⁰ This expansion can be written as

$$Y_m(\bar{s}) \approx w_m^T \bar{s} + b_m$$

where w_m is the derivative of Y_m w.r.t the mean spectrum \bar{s} and can be used as a variable weighting of the wavenumbers.³⁰ The weights or magnitude of the derivative w_m indicates which wavenumbers influence the classification of Mayo score m .²⁹ The magnitude of the derivative was plotted as a heat map for the correctly classified mean spectrum \bar{s} for all the four Mayo scores. The yellow color in the heat map shows higher derivative values, thus indicating higher impact of a Raman band on classification of a Mayo score. The Raman bands with high impact (yellow color) were analyzed by an expert for its biological implications.

The saliency map visualization was performed using a Keras-visualization toolkit³¹ for the correctly predicted Mayo endoscopic score using the last dense layer of the 1D-CNN classification model. For the visualization purpose, the last dense layer with the softmax activation function was replaced by a linear activation function to achieve an unnormalized prediction score.

RESULTS AND DISCUSSION

Characterization of Raman Spectra in UC with its Mayo Subscore. Raman imaging of 42 biopsy samples from the UC patients was performed in the colon mucosa area. In summary, 227 Raman maps with 567500 spectra were used for characterization of mucosa inflammation in UC. Already during colonoscopy, each area of the colon was characterized based on the Mayo endoscopic score (with stages from 0 to 3) by an experienced endoscopist. Following the biopsy, the specimens were shock frozen and later cut for the Raman spectral investigations. In order to analyze molecular variations taking place during mucosal inflammation in UC, tissue sections were measured by Raman imaging and annotated into one of the four Mayo endoscopic scores. During the analysis of the Raman datasets, spectral differences for each Mayo score were found. The mean spectra in the fingerprint spectral region of 500–1800 cm^{-1} and in the region 2800–3020 cm^{-1} are presented in Figure 3. These Raman spectra showed typical molecular characteristics of biological tissue, like the phenylalanine band (1002 cm^{-1}), the CH_2 deformation band (1440 cm^{-1}), the amide I band (1680–1620 cm^{-1}), and the CH stretching intensities (3020–2800 cm^{-1}). However, the spectral differences shown in Figure 3 obtained by the mean spectra of each Mayo score exhibit only subtle changes, barely visible by the naked eye. Therefore, a sophisticated visualization method is needed. For better visualization of the spectral variations, the difference spectra were calculated

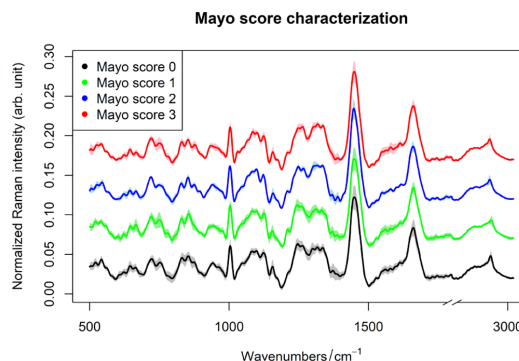


Figure 3. Mean spectra and its standard deviation of each Mayo endoscopic score with an offset. The mean spectra for the four Mayo endoscopic scores 0, 1, 2, and 3 are shown in black, green, blue, and red, respectively. The mean spectra of the four Mayo scores show subtle variations.

between each of the Mayo scores and the plot is included in the Supporting Information (Figure S1). The difference spectra show very interesting tendency which is as follows: for normal (Mayo score 0) against mild inflamed colon (Mayo score 1), protein bands were identified as the main difference; for mild (Mayo score 1) and moderate inflamed colon (Mayo score 2), mostly lipids and some protein were found different; and for moderate (Mayo score 2) against severe inflammation (Mayo score 3), bands of DNA as well as proteins and lipids were identified. The analysis of normal colon spectra (Mayo score 0) against severe inflamed colon spectra (Mayo score 3) showed variations in proteins and lipids; nevertheless, bands in the range of 1136–1140 cm^{-1} could not be identified based on the reference literature.³² A list of five Raman bands with a minimum and maximum difference between the Mayo endoscopic scores is given in the Supporting Information (Table S1).

Modeling of the Spectroscopic Inflammation Cascade. In order to characterize the inflammation and detect significant spectral signatures, a 1D-CNN shown in Figure 2 was applied to the Raman spectra of 227 Raman maps. The results in Table 3 show the classification of the inflammation cascade based on Mayo endoscopic scores. The modeling of the stages of the Mayo endoscopic score is difficult because of subtle spectroscopic changes between the Mayo scores. Furthermore, the variance between and within the patients of a specific Mayo score makes the classification even more difficult. However, the 1D-CNN used as a feature extractor and as a classifier showed an acceptable performance. The 1D-CNN classification model achieved a mean sensitivity of ~78% and ~93%, respectively. The risk factor of the model is ~1.4 which is an acceptable value for a good model (as the maximum risk factor in this case is 4). The confusion matrix and individual mean sensitivities for each Mayo score can be seen in Table 3. The individual mean sensitivities for Mayo score 0 (100.00%) and Mayo score 1 (81.82%) are much better than the mean sensitivities for Mayo score 2 (75.00%) and 3 (55.56%). From the mean sensitivity, it can be interpreted that lower Mayo scores (0 and 1) can be correctly predicted; however, higher Mayo scores (2 and 3) are misclassified among each other possibly because of subtle spectral variations and biological variations within and between

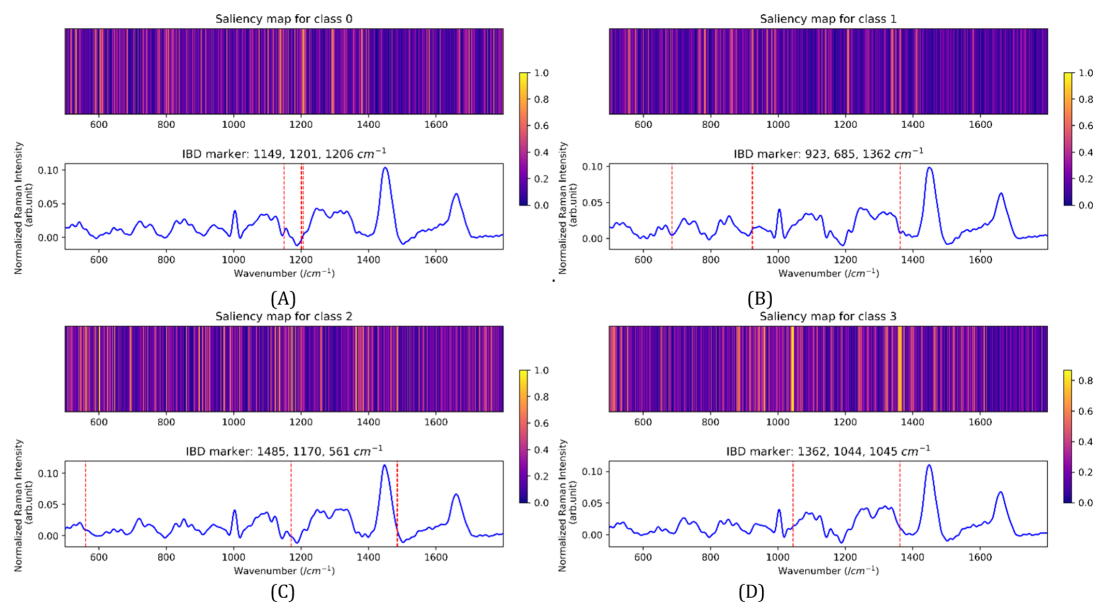


Figure 4. Interpretation of important wavenumber regions is achieved by saliency map visualizations of the finger-print region for all the four Mayo scores (A–D). A high saliency score (yellow) indicates important contribution of wavenumber regions to each Mayo score classification. The upper and lower panel in each subfigure shows saliency maps and mean spectra for each Mayo score 0, 1, 2, and 3, respectively. The 10 most important wavenumber positions for all the Mayo score are given in Table 4.

Table 4. Biological Annotation for the 10 Most Significant Raman Bands for the Classification Model Is Presented in the Table^a

Mayo score	wavenumbers (cm ⁻¹)									
0	1149	1201	1206	1290	806	1209	1381	1699	805	984
biological annotation		P	P	DNA		P		P		
1	923	685	1362	922	766	982	815	967	1340	781
biological annotation		DNA	DNA		P		P, DNA		DNA	DNA
2	1485	1170	561	602	1487	646	1137	1364	1486	990
biological annotation	DNA	P		DNA	DNA	P			P	
3	1362	1044	1045	1042	1359	1375	1046	1358	960	1361
biological annotation	DNA				P				C	DNA

^aAnnotations were combined into classes: proteins (P), cholesterol (C), and DNA based on reference 32.

the patients. It is also seen that 4 patients with inflammation (*i.e.*, Mayo score 1 and 3) were predicted as the lowest Mayo score 0. This was a severe misclassification and was penalized by the metric risk factor. Furthermore, there were total 7 patients that were misclassified as its adjacent lower or higher Mayo score. This misclassification between the adjacent Mayo scores can be attributed to the subtle changes between the adjacent Mayo scores, yet, this can be clinically accepted. Furthermore, the misclassification between adjacent Mayo scores was analyzed using a 1D-CNN regression model (Figure 2). The misclassification of higher Mayo score (1 and 3) to a lower Mayo score (0 or 1) can be a risk and was also investigated by an expert. With this regard, the pathological data of the misclassified patients were investigated to confirm the cause of the misclassification because of other illness, medication, age, or gender. Unfortunately, no significant correlation for misclassifications was found based on above-described parameters. Additionally, the 1D-CNN classification model was also evaluated for a binary classification task

considering Mayo score 0 as the noninflamed group and Mayo score 1, 2, and 3 as the inflamed group. The results of the binary classification task achieved a mean sensitivity of 71.42%. The results show that the inflamed group can be efficiently predicted by the model to eventually receive a treatment. However, a noninflamed patient is also predicted as inflamed and we suspect similar reasons for misclassification as mentioned above. Subsequently, the interpretation of the 1D-CNN classification model was obtained by first-order Taylor expansion which is discussed further.

Model Interpretation Using Saliency Maps. In order to analyze molecular information important for the Mayo endoscopic score prediction of the UC inflammation, saliency maps were calculated. The results of this calculation for each Mayo endoscopic score are presented in Figure 4. The weight or importance of each Raman band in the finger-print region (500–1800 cm⁻¹) is shown by the heat map for each Mayo score.

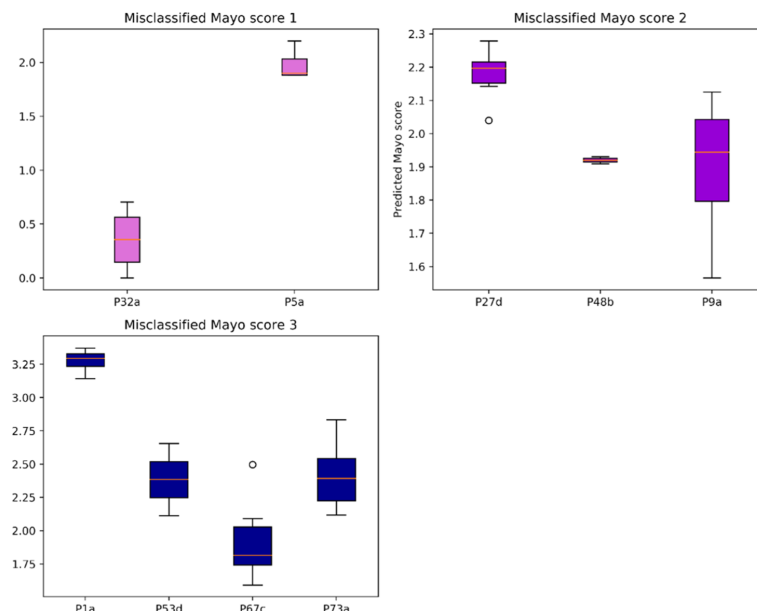


Figure 5. Box plots of 9 misclassified patients using the 1D-CNN classification model for all the four Mayo scores. The box plots show the extent of the misclassification from the true Mayo score. Out of the 9 misclassified patients, P27d, P9a, P1a, P53d, and P73a are close to the true Mayo score.

Additionally, we detected 10 most significant wavenumber positions in the finger-print region ($500\text{--}1800\text{ cm}^{-1}$) for every Mayo score (0, 1, 2, and 3) and assigned them to a biomolecule based on the reference literature.³² The assignment of these biomolecules is summarized in Table 4 and Figure S3. For the prediction of normal tissue (Mayo score 0), 3 out of 5 identified significant Raman bands correspond to proteins at amide III and CH_2 wagging vibrations ($1201\text{--}1207\text{ cm}^{-1}$) as well as proteins at 1699 cm^{-1} and DNA band at 1290 cm^{-1} . For prediction of mild, moderate, and severe inflammation in colon (Mayo score 1, 2, and 3, respectively), the most important Raman bands were DNA bands (10 of 16 bands) rather than bands of proteins ($n = 5$) and cholesterol ($n = 1$). Additionally, cholesterol was important in classification of severe inflammation. Many of the Raman bands were not possible to assign to any type of molecular vibration based on references, therefore requires further investigation. Here, we admit that the selective classification of normal colon based on amide III and CH_2 wagging vibrations of proteins, typical for any biological tissue, can be one of the reasons of misclassification of inflamed colon against the healthy one (Table 4).

Analysis of Misclassification Using the 1D-CNN Regression Model. As mentioned earlier, regression analysis was performed for two main reasons. First, to quantify the extent of misclassification achieved using the 1D-CNN classification model and second, to study the border-line patients. The regression analysis for Mayo score 1 shows two misclassified patients, out of which patient P5a has Mayo score lying between 1.5 and 2 and patient P32a lies between 0 and 0.75 (Figure 5, top left). Likewise, the regression analysis of Mayo score 2 shows 3 misclassified patients (*i.e.*, P48b, P27d, and P9a) predicted closer to the true Mayo score (*i.e.*, Mayo score 2). Thus, here, we can interpret that the extent of

misclassification of Mayo score 2 is not very severe (Figure 5, top right). Finally, the regression analysis of Mayo score 3 shows 3 patients (*i.e.*, P1a, P73a, and P53d) out of 4 predicted close to Mayo score 3 (Figure 5, bottom left). The 2 patients, namely, P1a and P73a, were predicted as Mayo score 0 using the 1D-CNN classification model; however, by 1D-CNN regression analysis, they were predicted in the clinically acceptable range. Similarly, patients P32a and P5a were predicted as Mayo score 0 using the 1D-CNN classification model; however, these patients were estimated in an acceptable range using the 1D-CNN regression model. Thus, a 1D-CNN regression model is needed to complement the 1D-CNN classification model for assessing the Mayo score in UC patients. Furthermore, the box plot acquired from the regression analysis of all the patients and all the Mayo scores is given in the Supporting Information (Figure S2). The plot of regression analysis shows a significantly better prediction for Mayo score 1 and Mayo score 2 than the prediction for Mayo score 0 and 3. Finally, the misclassified patients were studied by a medical expert for other factors such as age, sex, other diseases, and treatments, which can possibly cause the misclassification; however, no significant correlations of these factors with the misclassifications were found. We suspect that interpatient and inpatient variations and probably selective classification of normal colon against inflamed based on common protein bands (Amide III vibration) in the range of $1200\text{--}1300\text{ cm}^{-1}$ can cause this problem. Moreover, existing disagreements for Mayo scoring of the UC should be stated here. Nevertheless, we utilized the Mayo score for our study because it is the most common classification score for IBD inflammation in clinics. Further classification scores such as Riley score, Nancy score, and endoscopic images and videos will be utilized in further systematic investigations.

CONCLUSIONS

The major challenge in diagnostics of diverse diseases such as UC is that a wide spectrum of pathologies is involved in the development of UC, which have similar clinical, endoscopic, and histological manifestations.³³ Therefore, novel technologies are needed for independent identification of UC, its inflammation stages, and moreover, biopsy-free monitoring of this untreatable disease. In our study, we focused on the investigation of molecular changes occurring during inflammation of UC by nondestructive and label-free Raman spectroscopy. The Raman spectra were utilized to predict the Mayo endoscopic score in patient biopsy sections using a 1D-CNN. The prediction of the Mayo endoscopic score achieved a mean sensitivity of 78% and mean specificity of 93%. The low sensitivity can be attributed to limited dataset, interpatient variance, and local changes in different inflammation stages. Moreover, the Raman bands important for the prediction of the Mayo endoscopic scores were interpreted using a first-order Taylor expansion of the 1D-CNN. The results show that important molecular changes occur during inflammation and molecules such as proteins, lipids, cholesterol, amino acids, DNA, and saccharides were involved. This work has broader implications and can be used for similar Raman datasets or other spectroscopic datasets, which includes the classification of fresh biopsy samples or the diagnostics of other kinds of diseases. Although the general study workflow can be followed for other studies, all parts of the study such as the type of the sample and Raman spectroscopic alterations because of the external factor will influence the spectral data. To tackle this proper calibration, model transfer methods or fine-tuning of the DCNN needs to be performed.³⁴ Overall, utilizing Raman spectroscopy along with DCNNs for characterizing inflammation stages in UC can provide not only a minimal-risk diagnostic procedure but also a real-time decision-making system.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.analchem.0c02163>.

Box plot of regression analysis for all Mayo endoscopic scores; difference spectrum for Mayo score 0–1, 1–2, 2–3, and 0–3; five Raman bands showing maximal and minimal differences in the difference spectrum of Mayo score 0–1, 1–2, 2–3, and 0–3; and saliency map visualizations with threshold set at 0.75 (PDF)

AUTHOR INFORMATION

Corresponding Authors

Juergen Popp – Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller University, 07743 Jena, Germany; Leibniz Institute of Photonic Technology, Member of Leibniz Health Technology, 07745 Jena, Germany; orcid.org/0000-0003-4257-593X; Phone: +49 (0) 3641 948300; Email: juergen.popp@uni-jena.de

Thomas Wilhelm Bocklitz – Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller University, 07743 Jena, Germany; Leibniz Institute of Photonic Technology, Member of Leibniz Health Technology, 07745 Jena, Germany; orcid.org/0000-0003-2778-6624; Phone: +49 (0) 3641 948328; Email: thomas.bocklitz@uni-jena.de

Andreas Stallmach – Department of Internal Medicine IV (Gastroenterology, Hepatology, Infectious Disease), Jena University Hospital, 07747 Jena, Germany; Phone: +49 (0) 3641 9-32 4221; Email: andreas.stallmach@med.uni-jena.de

Authors

Tatiana Kirchberger-Tolstik – Leibniz Institute of Photonic Technology, Member of Leibniz Health Technology, 07745 Jena, Germany; Department of Internal Medicine IV (Gastroenterology, Hepatology, Infectious Disease), Jena University Hospital, 07747 Jena, Germany

Pranita Pradhan – Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller University, 07743 Jena, Germany; Leibniz Institute of Photonic Technology, Member of Leibniz Health Technology, 07745 Jena, Germany; orcid.org/0000-0002-0558-2914

Michael Vieth – Klinikum Bayreuth GmbH, 95445 Bayreuth, Germany

Philip Grunert – Department of Internal Medicine IV (Gastroenterology, Hepatology, Infectious Disease), Jena University Hospital, 07747 Jena, Germany

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.analchem.0c02163>

Author Contributions

T.K.-T. and P.P. contributed equally to this work. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was funded by the Federal Ministry of Education and Research (BMBF) Germany (FKZ: 01 E0 1002). The German Research Foundation (DFG) provided research funding to TB (BR 4182/3-1), to TWB (BO 4700/1-1), to JP (PO 563/30-1), and to STA (295/11-1). Additionally, the funding of the CRC 1278 Poly-Target by the DFG is highly acknowledged. This work received financial support by the Ministry for Economics, Sciences and Digital Society of Thuringia (TMWWDG), under the framework of the Landesprogramm ProDigital (DigLeben-5575/10-9). The project on which these results are based was supported by the Free State of Thuringia under the number 2019 FGR 0083 and cofinanced by European Union funds within the framework of the European Social Fund (ESF).

REFERENCES

- (1) Xie, T.; Zhang, T.; Ding, C.; Dai, X.; Li, Y.; Guo, Z.; Wei, Y.; Gong, J.; Zhu, W.; Li, J. *Gastroenterol. Rep.* **2018**, *6*, 38–44.
- (2) Ungaro, R.; Mehandru, S.; Allen, P. B.; Peyrin-Biroulet, L.; Colombel, J.-F. *Lancet* **2017**, *389*, 1756–1770.
- (3) Actis, G. C.; Pellicano, R.; Rosina, F. *World J. Gastrointest. Pharmacol. Therapeut* **2014**, *5*, 169–174.
- (4) Bewtra, M.; Brensinger, C. M.; Tomov, V. T.; Hoang, T. B.; Sokach, C. E.; Siegel, C. A.; Lewis, J. D. *Inflamm. Bowel Dis.* **2014**, *20*, 1070–1078.
- (5) Ikeya, K.; Hanai, H.; Sugimoto, K.; Osawa, S.; Kawasaki, S.; Iida, T.; Maruyama, Y.; Watanabe, F. *J. Crohn's and Colitis* **2015**, *10*, 286–295.
- (6) Daperno, M.; Comberlato, M.; Bossa, F.; Biancone, L.; Bonanomi, A. G.; Cassinotti, A.; Cosentino, R.; Lombardi, G.;

- Mangiarotti, R.; Papa, A.; Pica, R.; Rizzello, F.; D'Incà, R.; Orlando, A. *Dig. Liver Dis.* **2014**, *46*, 969–973.
- (7) Fernandes, S. R.; Pinto, J. S. L. D.; Marques da Costa, P.; Correia, L.; GEDII, M. D. *Inflamm. Bowel Dis.* **2018**, *24*, 254–260.
- (8) Bielecki, C.; Marquardt, C.; Stallmach, A.; Bocklitz, T.; Schmitt, M.; Popp, J.; Krafft, C.; Gharbi, A.; Knosel, T. *J. Biomed. Optic.* **2012**, *17*, 076030.
- (9) Bi, X.; Walsh, A.; Mahadevan-Jansen, A.; Herline, A. *Dis. Colon Rectum* **2011**, *54*, 48–53.
- (10) Pence, I. J.; Beaulieu, D. B.; Horst, S. N.; Bi, X.; Herline, A. J.; Schwartz, D. A.; Mahadevan-Jansen, A. *Biomed. Opt. Express* **2017**, *8*, 524–535.
- (11) Haifler, M.; Pence, I.; Sun, Y.; Kutikov, A.; Uzzo, R. G.; Mahadevan-Jansen, A.; Patil, C. A. *J. Biophot.* **2018**, *11*, No. e201700188.
- (12) Pandey, R.; Zhang, C.; Kang, J. W.; Desai, P. M.; Dasari, R. R.; Barman, I.; Valdez, T. A. *J. Biophot.* **2018**, *11*, No. e201700259.
- (13) Managò, S.; Mirabelli, P.; Napolitano, M.; Zito, G.; De Luca, A. *J. Biophot.* **2018**, *11*, No. e201700265.
- (14) Krauß, S. D.; Roy, R.; Yosef, H. K.; Lechtonen, T.; El-Mashtoly, S. F.; Gerwert, K.; Mosig, A. *J. Biophot.* **2018**, *11*, No. e201800022.
- (15) Bocklitz, T.; Walter, A.; Hartmann, K.; Rösch, P.; Popp, J. *Anal. Chim. Acta* **2011**, *704*, 47–56.
- (16) Dörfer, T.; Bocklitz, T.; Tarcea, N.; Schmitt, M.; Popp, J. *J. Phys. Chem.* **2011**, *225*, 753–764.
- (17) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (18) Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer Series in Statistics: New York, USA, 2001.
- (19) Agarap, A. F. Deep Learning Using Rectified Linear Units (Relu). **2018**, ArXiv, abs/1803.08375.
- (20) Ioffe, S.; Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on International Conference on Machine Learning JMLR.Org: Lille, France, 2015*; Vol. 37, pp 448–456.
- (21) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
- (22) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, Massachusetts, United States, 2016.
- (23) Saad, D. *On-line Learning in Neural Networks*; Cambridge University Press: New York, USA, 2009.
- (24) Sibi, P.; Jones, S. A.; Siddarth, P. *J. Theor. Appl. Inf. Technol.* **2013**, *47*, 1264–1268.
- (25) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. **2015**, abs/1412.6980. CoRR.
- (26) Oliphant, T. E. *Guide to NumPy*; CreateSpace Independent Publishing Platform: North Charleston: South Carolina, USA, 2015.
- (27) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. Tensorflow: A system for large-scale machine learning. *Proceedings of the 12th Symposium on Operating Systems Design and Implementation*, Savannah: GA, USA, 2016; pp 265–283.
- (28) Chollet, F. *Keras*; Github, 2015.
- (29) Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. **2014**, abs/1312.6034. CoRR.
- (30) Bocklitz, T. Understanding of Non-linear Parametric Regression and Classification Models: A Taylor Series based Approach. *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods Scitepress—Science and Technology Publications: Prague, Czech Republic, 2019*; pp 874–880.
- (31) Kotikalapudi, R. a. c. *Keras-Vis*; GitHub, 2017.
- (32) Movasaghi, Z.; Rehman, S.; Rehman, I. U. *Appl. Spectrosc. Rev.* **2007**, *42*, 493–541.
- (33) Otero Regino, W.; González, A.; Gómez Zuleta, M. *Colomb. J. Gastroenterol.* **2009**, *24*, 272–278.
- (34) Guo, S.; Ryabchikov, O.; Ali, N.; Houhou, R.; Bocklitz, T. Comprehensive Chemometrics. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier: Oxford, United Kingdom, 2020.

Supporting Information

Towards an interpretable classifier for characterization of endoscopic Mayo scores in ulcerative colitis using Raman Spectroscopy

Tatiana Kirchberger-Tolstik ^{†,‡,§}, Pranita Pradhan ^{‡,‡,§}, Michael Vieth [¶], Philip Grunert [†],
Juergen Popp ^{‡,‡,*}, Thomas Wilhelm Bocklitz ^{‡,‡,*}, Andreas Stallmach ^{†,*}

[‡]Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller University, 07743 Jena, Germany, [‡]Leibniz Institute of Photonic Technology, Member of Leibniz Health Technology, 07745 Jena, Germany, [†]Department of Internal Medicine IV (Gastroenterology, Hepatology, Infectious Disease), Jena University Hospital, 07747 Jena, Germany, [¶]Klinikum Bayreuth GmbH, Preuschwitzer Str. 101, 95445 Bayreuth, Germany.

* Corresponding authors: andreas.stallmach@med.uni-jena.de, juergen.popp@uni-jena.de and thomas.bocklitz@uni-jena.de
§TT and PP contributed equally to this work.

ABSTRACT: Ulcerative colitis (UC) is one of the main types of chronic inflammatory diseases that affect the bowel, but its pathogenesis has yet to be completely defined. Assessing the disease activity of UC is vital for developing a personalized treatment. Conventionally, the assessment of UC is performed by colonoscopy and histopathology. However, conventional methods fail to retain biomolecular information associated to the severity of UC and are solely based on morphological characteristics of the inflamed colon. Further, assessing endoscopic disease severity is limited by the requirement for experienced human reviewers. Therefore, this work presents a non-destructive bio-spectroscopic technique, e.g. Raman spectroscopy, for assessing endoscopic disease severity according to the four-level Mayo subscore. This contribution utilizes the multi-dimensional Raman spectroscopic data to generate a predictive model for identifying colonic inflammation. The predictive modelling of the Raman spectroscopic data is performed using a one-dimensional deep convolutional neural network (1D-CNN). The classification results of 1D-CNN achieved a mean sensitivity of 78% and a mean specificity of 93% for the four Mayo endoscopic scores. Furthermore, the results of the 1D-CNN are interpreted by a first-order Taylor expansion, which extracts the Raman bands important for classification. Additionally, a regression model of the 1D-CNN model is constructed to study the extent of misclassification and border-line patients. The overall results of Raman spectroscopy with 1D-CNN as classification and regression model show a good performance, and such method can serve as a complementary method for UC analysis.

Table of contents

Figure S1 Difference spectrum of Mayo endoscopic score 0-1, 1-2, 2-3 and 0-3.

Figure S2 Results of regression analysis for all Mayo endoscopic scores.

Figure S3 Saliency map showing important Raman bands with threshold set at 0.75.

Table S1 A list of five Raman bands with minimum and maximum difference between the Mayo endoscopic scores 0-1, 1-2, 2-3, 0-3.

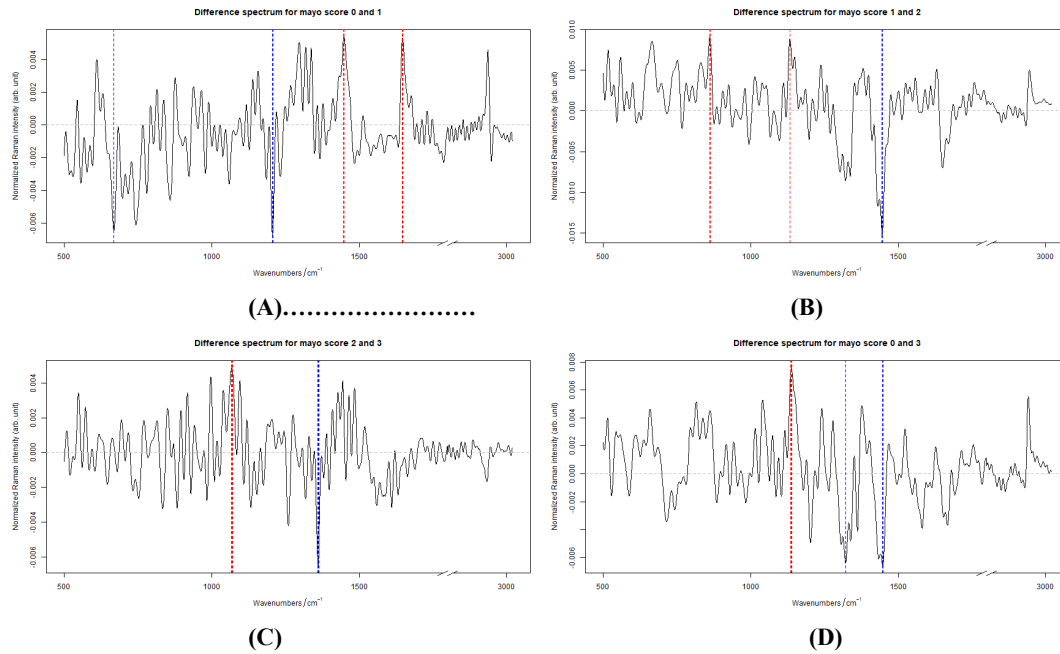


Figure S1. (A-D) shows difference spectrum of Mayo endoscopic score 0-1, 1-2, 2-3 and 0-3. The red and blue dotted vertical lines show five Raman band with maximum and minimum difference between the two Mayo scores. The information about the five Raman bands with maximum and minimum difference is given in the table S1 along with its biological annotation.

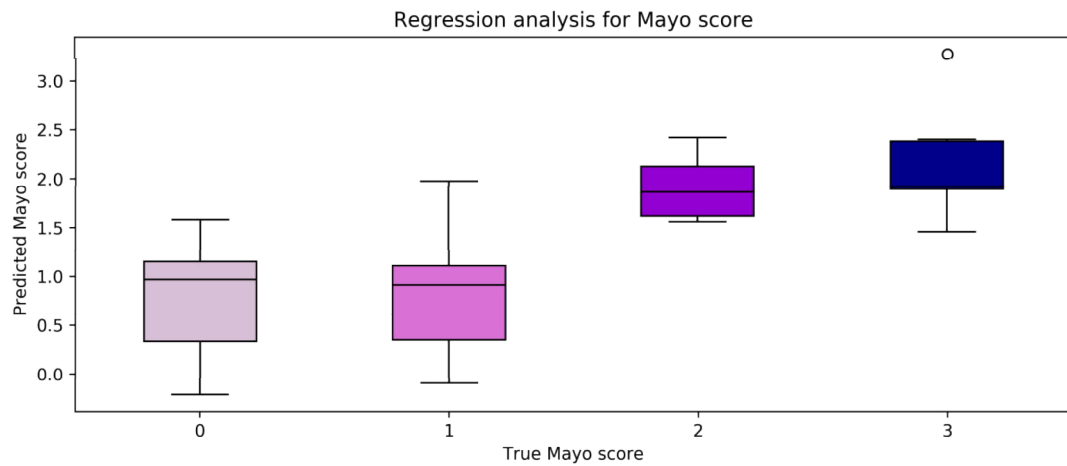


Figure S2. Results of regression analysis for all the Mayo endoscopic scores shows the extent of misclassification between the Mayo scores. The average RMSE for 42 patients is 0.54. Regression analysis shows that the predictions of Mayo score 1 and 2 lies within an acceptable range, however, predictions of Mayo score 0 and 3 lie between 0 to 1.5 and 2 to 2.5, respectively.

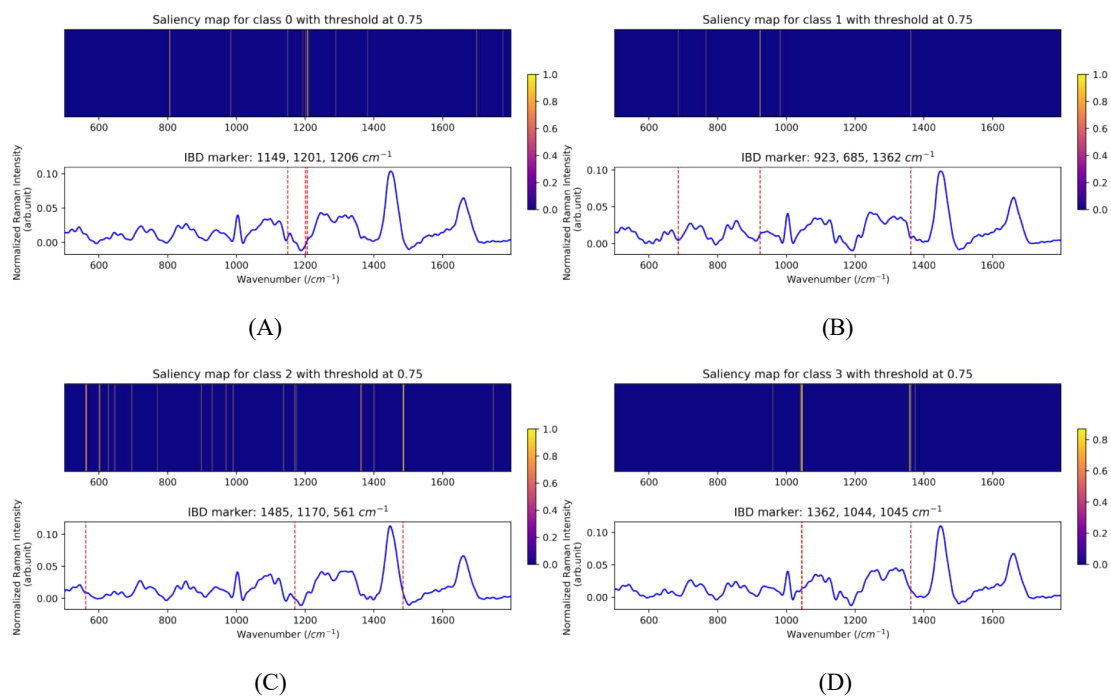


Figure S3. The interpretation of important wavenumber regions shown by saliency map visualizations with a threshold set at 0.75. Each subplot shows at least 10 important Raman bands. The colour in the saliency map signifies the importance of a particular Raman band. A high saliency score (yellow) indicates important contribution of wavenumber regions to each Mayo score classification. The upper and lower panel in each subfigure shows saliency maps and mean spectra for each Mayo score 0, 1, 2 and 3, respectively. The ten most important wavenumber positions for all the Mayo score are given in Table 4.

DIFFERENCE SPECTRUM	MODE	WAVENUMBERS (CM ⁻¹)				
		0 AND 1	min	1207	1206	1208
	<i>Amide III & CH₂ wagging vibrations (proteins) and DNA</i>					
max	1449		1450	1448	1648	1647
		<i>C-H vibration (proteins and lipids) and protein bands</i>				
1 AND 2	min	1446	1447	1445	1448	1444
		<i>CH₂ bending mode (proteins & lipids)</i>				
	max	1133	1132	1134	1131	861
		<i>Fatty acids (lipids)</i>				
2 AND 3	min	1361	1360	1362	1359	1363
		<i>Guanine (DNA)</i>		<i>Tryptophan (proteins)</i>		<i>Guanine (DNA)</i>
	max	1070	1069	1071	1068	1072
		<i>Triglycerides (lipids) or DNA</i>				
0 AND 3	min	1448	1447	1449	1446	1450
		<i>CH₂ bending mode (proteins & lipids)</i>				
	max	1138	1139	1137	1140	1136
		-				

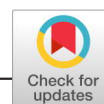
Table S1 A list of five Raman bands with minimum and maximum difference between the Mayo endoscopic scores 0-1, 1-2, 2-3, 0-3. The difference spectrum between 0 and 1 shows variation in Amide III, CH₂ wagging vibrations, DNA, C-H vibration (proteins and lipids) and protein bands. The difference spectrum between 1 and 2 shows variation in CH₂ bending mode and fatty acids. The difference spectrum between 2 and 3 shows variation in Guanine, Tryptophan and Triglycerides. The difference spectrum between 0 and 3 shows variation in CH₂ bending mode.

P2 DEEP LEARNING A BOON FOR BIOPHOTONICS?

Reprinted with permission from [P. Pradhan, S. Guo, O. Ryabchykov, J. Popp, T. W. Bocklitz, Deep learning a boon for Biophotonics?, 2020, *Journal of Biophotonics*, Vol. 13(6): e201960186, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim]. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

The declared individual contributions of the doctoral candidate and the other doctoral candidates participate as co-authors in the publications are listed below.

P. Pradhan ¹ , S. Guo ² , O. Ryabchykov ³ , J. Popp ⁴ , T. W. Bocklitz ⁵ , Deep learning a boon for Biophotonics?, 2020, <i>Journal of Biophotonics</i> , Vol. 13(6): e201960186, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim					
Involved in (Please tick the boxes that apply.)					
	1	2	3	4	5
Conceptual research design				X	X
Planning of research activities	X				X
Data collection	–	–	–	–	–
Data analysis and interpretation	–	–	–	–	–
Manuscript writing	X	X	X	X	X
Suggested publication equivalence value	1.0				


REVIEW ARTICLE

Deep learning a boon for biophotonics?

Pranita Pradhan^{1,2} | Shuxia Guo^{1,2} | Oleg Ryabchykov^{1,2} | Juergen Popp^{1,2} | Thomas W. Bocklitz^{1,2}

¹Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller-University, Jena, Germany

²Leibniz Institute of Photonic Technology (Leibniz-IPHT), Member of Leibniz Research Alliance 'Health Technologies', Jena, Germany

Correspondence

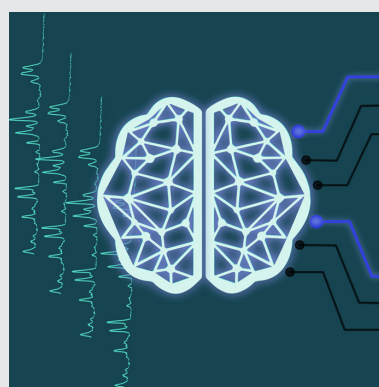
Thomas W. Bocklitz, Institute of Physical Chemistry and Abbe Center of Photonics (IPC), Friedrich-Schiller-University, Helmholtzweg 4, D-07743 Jena, Germany. Email: thomas.bocklitz@uni-jena.de

Funding information

Bundesministerium für Bildung und Forschung, Grant/Award Number: Uro-MDD (FKZ 03ZZ0444J); Deutsche Forschungsgemeinschaft, Grant/Award Numbers: BO 4700/1-1, BO 4700/4-1, PO 563/30-1, STA 295/11-1; Thüringer Ministerium für Wirtschaft, Wissenschaft und Digitale Gesellschaft, Grant/Award Number: DigLeben (5575/10-9)

Abstract

This review covers original articles using deep learning in the biophotonic field published in the last years. In these years deep learning, which is a subset of machine learning mostly based on artificial neural network geometries, was applied to a number of biophotonic tasks and has achieved state-of-the-art performances. Therefore, deep learning in the biophotonic field is rapidly growing and it will be utilized in the next years to obtain real-time biophotonic decision-making systems and to analyze biophotonic data in general. In this contribution, we discuss the possibilities of deep learning in the biophotonic field including image classification, segmentation, registration, pseudostaining and resolution enhancement. Additionally, we discuss the potential use of deep learning for spectroscopic data including spectral data preprocessing and spectral classification. We conclude this review by addressing the potential applications and challenges of using deep learning for biophotonic data.


KEYWORDS

artificial neural networks, biophotonics, deep learning, spectroscopy

1 | INTRODUCTION

Biophotonics is a rapidly growing multidisciplinary field that utilizes the interaction of light with biological systems and investigates these biological systems at the cellular, molecular and tissue level. Since the past decade, these biophotonic technologies are globally established in biotechnology companies, healthcare organizations,

medical instrument suppliers and pharmaceutical manufacturers. For instance, laser-based therapy is an important part of medical sciences today, and is used for light-guided therapies in various organs. Other light-based technologies like multiphoton microscopy (MPM), optical coherence tomography (OCT), Raman spectroscopy, infrared spectroscopy (IR), photoacoustic imaging (PAI) and fluorescence life-time imaging microscopy (FLIM)

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Journal of Biophotonics* published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

are further useful tools in biomedical and biophotonic research [1, 2]. For example, nonlinear multimodal imaging which includes two-photon excited fluorescence microscopy (TPEF), second-harmonic generation (SHG) and coherent anti-stokes Raman scattering (CARS), is widely used in dermatology, physiology, neurobiology and embryology. Similarly, technologies like OCT are mainly used in ophthalmology and cardiology, while spectroscopic techniques have various clinical and pharmaceutical applications.

Nowadays, biophotonic technologies are witnessing a rapid development in the instrumentation of the optical devices which is fastening the imaging speed, increasing the penetration depth and enhancing the resolution of the optical images. These developments make it possible to measure label-free molecular information of samples like cells or tissue. As all of the biophotonic technologies are label-free, the spectral and image data is untargeted. That means it is difficult to interpret a specific contrast associated to a chemical structure or a biomolecule in biophotonic data. Therefore, the interpretation of biophotonic data has to be generated using appropriate analysis techniques like statistics, chemometrics or machine learning. Additionally, the technical improvement of these biophotonic technologies has given rise to large datasets, which require big data analysis methods to be applied to biophotonic data [3]. Overall, interpreting and handling biophotonic data are two obvious challenges for the biophotonic community (Table 1).

TABLE 1 List of mathematical symbols

Symbol	Explanation
x	An input as a scalar (integer or real)
\mathbf{x}	An input vector or 1D data
\mathbf{X}	An input matrix or 2D data
Θ	A set of trainable parameters
\mathcal{W}	A set of weights
\mathcal{B}	A set of biases
L	Number of layers in a model
l	Index of L layers
η	Learning rate
\mathbb{R}^D	A D dimensional set of real numbers
\mathbb{R}^N	A N dimensional set of real numbers
τ	An index for iteration
P	A probability distribution over discrete variable
χ	An input space representation
\mathcal{F}	A latent space representation
E	Loss function
$\mathbb{E}_{x,y}[f(x)]$	Expectation of $f(x)$ with respect to $P(y x)$

In this context, the well-established statistical pattern-recognition methods are employed which extract “features” or “patterns” from the biophotonic data. These techniques are called “feature extraction” methods. Feature extraction is a process of dimension reduction used to transform high dimensional data to low dimensional data. Subsequently, the low dimensional data commonly called “features” can be used to construct learning algorithms. This procedure is shared by most of the machine learning algorithms where feature extraction is followed by prediction of the outcome or probabilities [4]. Classification or regression models are common examples of machine learning algorithms where features from images (like shape, texture, color features) or features of spectra (like intensity values at specific wavenumbers in Raman spectroscopy) are extracted to construct a predictive model. These machine-learning algorithms in combination with a high computational power can be utilized to interpret the biophotonic data. A subset of machine learning algorithms is called “deep learning,” which requires least manual intervention for feature extraction and can be employed as a decision-making algorithm with high accuracy. Since a decade, deep learning algorithms have achieved promising results in clinical radiology covering a wide range of applications from cancer diagnosis to personalized therapies [5]. Similar to clinical radiology, introduction of deep learning algorithms in biophotonics has also revolutionized the data analysis in this field. The respective research will be further discussed in this article.

This review article aims to give an overview of deep learning techniques for spectroscopic data intended for the multidisciplinary readership of *J. Biophotonics*. We aim to stimulate the interest of researchers and data scientists to foster applications of deep learning in biophotonics by discussing the ongoing evolution in the field of biophotonics and deep learning. Additionally, we emphasize potential applications and challenges encountered while applying deep learning for biophotonic data. We structure our article in the following manner: section 2 discusses the commonly used deep learning architectures to analyze biophotonic data. Section 3 presents the applications of deep learning for preprocessing, classifying and segmenting microscopic imaging data. Section 4 presents the preprocessing and analysis for spectroscopic data using deep learning. Further, section 5 addresses the challenges faced by a researcher while analyzing biophotonic data using deep learning and we introduce the approaches for overcoming these challenges. Lastly, we conclude our review in section 6 by answering the question “Is deep learning a boon for biophotonics?”

2 | DEEP LEARNING—AN OVERVIEW

With rising complexity of spectroscopic datasets and the need to achieve good decision-making systems, more advanced machine learning algorithms are required. Briefly, a machine-learning algorithm is an algorithm that is able to learn from data. A special kind of machine learning algorithm is deep learning algorithm. A deep learning algorithm is based on four major components which are an optimization algorithm, a cost function, a dataset and a deep learning model. Shortly, an optimization algorithm is an iterative method to compare various solutions for a problem until an optimal solution is obtained. A cost function is a mathematical formula used to evaluate the performance of a deep learning model. A dataset is one of the major components for training the deep learning models and can be split into three parts: training, validation and testing dataset. The training dataset is used for training the deep learning model, the validation dataset is used to tune the hyperparameters of the deep learning model and the independent test dataset or holdout set is used to evaluate the performance of the model in an unbiased manner [4, 6–8]. The last necessary component is a deep learning model which is made of a series of layers and hyperparameters depending on various architectures, which are discussed in the further course of the section.

Deep learning algorithms have widespread applications in speech recognition, natural language processing, healthcare and so on. Particularly in healthcare, deep learning is often applied to radiology data. Similar to clinical radiology, traditional artificial neural networks [9] were applied since the 1990s to biophotonic data [10, 11] the recently developed deep learning models especially convolutional neural networks have achieved state-of-the-art performance in the biophotonic field. This section summarizes a few deep learning models that are commonly used to analyze spectroscopic data. Each subsection gives a brief overview of a specific deep learning architecture combined with an illustration of how to apply these deep learning architectures for image and spectral data.

2.1 | Feed-forward neural network

Feed-forward neural network commonly called artificial neural network (ANN) or multilayer perceptron (MLP) [9, 12, 13] are the basis of most of the deep learning models utilized today. MLPs are loosely inspired by the human neural system. These models are called feed-forward neural network as the input flows only in the

forward direction without a feedback from the output into the model. Specifically, a feed-forward neural network passes the input $\mathbf{x} = \{x_i\} \in \mathbb{R}^D$, through a series of neurons with an activation a and a set of trainable parameters $\Theta = \{\mathcal{W}, \mathcal{B}\}$ to obtain an output y . An activation function $a = \sigma(\mathbf{w}^T \mathbf{x} + b)$ introduces an elementwise nonlinearity $\sigma(\cdot)$ to the output of a neuron, which is a linear combination of the neuron's input and the parameters Θ (see Figure 1). A composition of many such transformations forms the basis of a feed-forward neural network where the input is passed through a series of “hidden layers” to obtain the output. A neuron output y_k of an MLP with M and D neurons in two hidden layers l and $l-1$ respectively, can be represented as

$$y_k(\mathbf{x}; \Theta) = \sigma^l \left(\sum_{j=1}^M W_{kj}^l \sigma^{l-1} \left(\sum_{i=1}^D W_{ji}^{l-1} x_i + b_j^{l-1} \right) + b_k^l \right), \quad (1)$$

where Θ is the set of trainable parameters, W_{ji} is a weight matrix of size $j \times i$, with i inputs and j activations of $(l-1)$ th layer. During the training of a feed-forward neural network, the model parameters Θ are iteratively updated using an optimizer until convergence is achieved. A stochastic gradient descent (SGD) optimizer is commonly used in the literature [15, 16], which performs typically the minimization of a loss or a cost function E by the

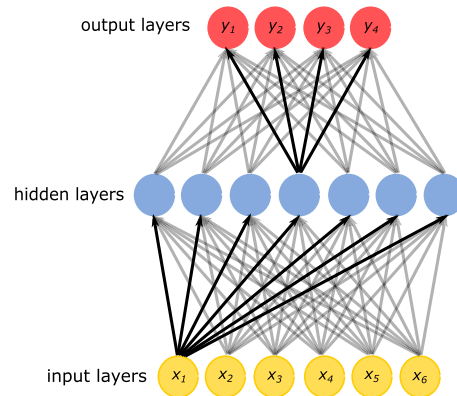


FIGURE 1 A feed-forward neural network or a multilayer perceptron with an input $\mathbf{x} \in \mathbb{R}^D$, $D = 6$ and output $y \in \mathbb{R}^N$, $N = 4$ is shown. The input to the network (depicted in yellow) can be features (like histogram features, local binary patterns [14]) obtained from an image or features (like intensity values of different wavenumbers) obtained from a spectrum which passes through the neurons of the hidden layer depicted in blue. The connections between the neurons are weighted by \mathcal{W} and the data is further passed through the layers with activation function a to obtain an output shown in red. The weights are updated using back-propagation as explained in Section 2.1

back-propagation method [17, 18]. Back-propagation minimizes the loss function in the parameter space Θ by computing a gradient of the loss function $E(\Theta)$ [17]. Based on the gradient of the loss function $\nabla E(\Theta)$ computed for all the layers, the model parameters $\Theta = \{\mathcal{W}, \mathcal{B}\}$ can be updated in each iteration τ using the formula given below:

$$\Theta^{(\tau+1)} = \Theta^{(\tau)} - \eta \nabla E(\Theta)^{(\tau)}. \quad (2)$$

Here, τ represents an iteration index and η is the learning rate. In addition to the SGD optimizer, other optimizers like Adam [19], Adadelta [20] and Adagrad [21] have also been reported in the literature.

MLPs have widespread applications in image and spectral classification as illustrated in Figure 2. The figure shows an MLP that utilizes image features or spectral features as the input. These features are further propagated through the network to emerge at the output neuron as class outputs (see Figure 2). The class outputs can be tumor/normal for a diagnostic task, disease stages for a disease assessment task or the type of pollen grains for a classification task of pollen grains. Mostly MLPs require the extraction of features from image or spectral data, which is one of the limitations of these basic neural networks. Therefore, more advanced deep learning architectures like convolutional neural networks are required.

2.2 | Convolutional neural network

A convolutional neural network (CNN) [22] is a variant of a MLP, which can work on grid data, for instance spectra or images. Unlike MLPs, CNNs consider the spatial

information of an image or temporal/spectral information of a signal directly. This is achieved by convolving the input, like an image \mathbf{X} , with trainable kernels or weights \mathbf{W}_k to generate a feature map \mathbf{X}_k . Mathematically, a feature map \mathbf{X}_k for the l th layer of a CNN is given by

$$\mathbf{X}_k^l = \sigma(\mathbf{W}_k^{l-1} * \mathbf{X}_k^{l-1} + b_k^{l-1}), \quad (3)$$

where $\mathcal{W} = \{\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_K\}$ are K trainable kernels and $\mathcal{B} = \{b_1, b_2, \dots, b_K\}$ are the biases. The illustration of a CNN architecture in Figure 3 shows a kernel \mathbf{W}_1 of size 3×3 , which is convolved with an image \mathbf{X} in a raster pattern with a stride value of 1 pixel (first layer). This forms a feature map or a linearly convolved image \mathbf{X}_1 . The linearly convolved image is further subjected to an elementwise nonlinear transformation σ which is typically a rectified linear unit (ReLU) [23], tanh [24] and sigmoid [25] function. The activation function σ is important in CNNs to introduce a nonlinearity to the model. Generally, a softmax activation function [6] in the last layer of a model utilized for classification tasks is used. The softmax activation layer maps the activations of the final layer to a probability distribution of classes $P(y|\mathbf{X}; \Theta)$ given as

$$P(y|\mathbf{X}; \Theta) = \text{softmax}(\mathbf{X}; \Theta) = \frac{e^{(\mathbf{w}_i^t)^T \mathbf{x} + b_i^t}}{\sum_{c=1}^C e^{(\mathbf{w}_c^t)^T \mathbf{x} + b_c^t}}, \quad (4)$$

where \mathbf{W}_i^t and b_i^t are the kernel and bias of the l th layer leading to a normalized probability distribution of class i . In contrast to other traditional activation functions, the output of a softmax activation function is normalized

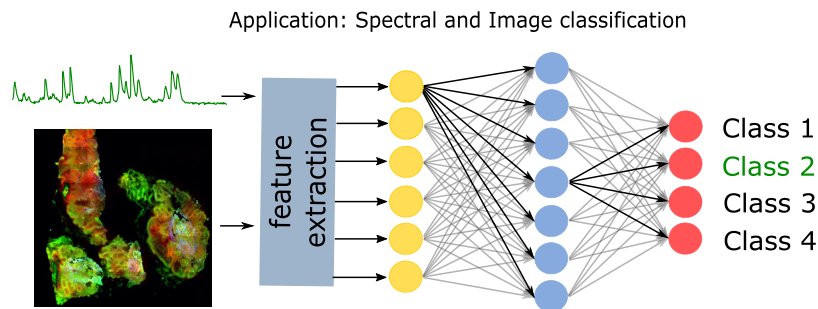
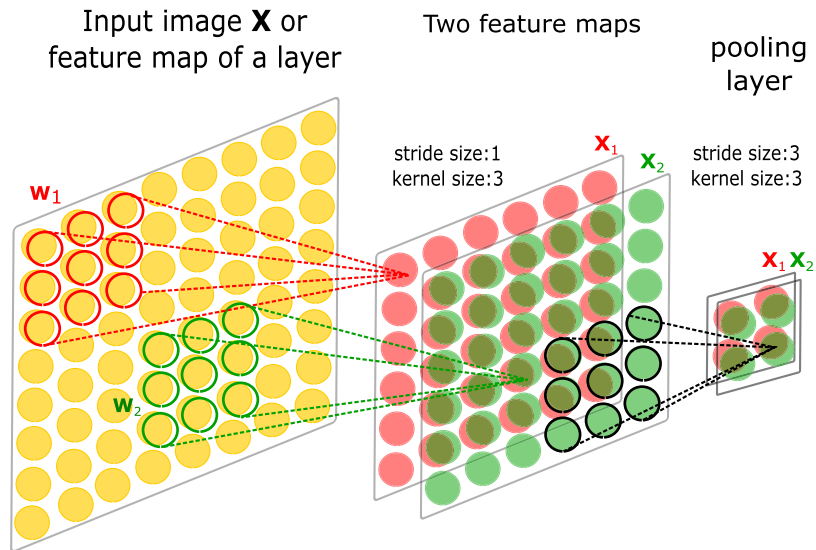


FIGURE 2 Applications of MLPs are shown, where each input neuron utilizes the features obtained from a Raman spectrum (top) and nonlinear multimodal image (bottom). The nonlinear multimodal image is composed of CARS signal as red channel, TPEF signal as green channel and SHG signal as blue channel. The input vector at the first layer is a vector of image features or spectral features. The output neuron of the MLP is a label or a class probability of the input spectrum or of the input image. CARS, coherent anti-stokes Raman scattering; MLP, multilayer perceptron; SHG, second-harmonic generation; TPEF, two-photon excited fluorescence microscopy

FIGURE 3 A general structure of a convolution neural network (CNN) is shown. The input image \mathbf{X} or a feature map of a layer is convolved by two kernels \mathbf{W}_1 and \mathbf{W}_2 . Each kernel of size 3×3 is convolved with a small section of the input image and is shifted with a stride of 1 pixel (first layer) in a raster pattern to obtain a whole feature map \mathbf{X}_1 and \mathbf{X}_2 . The figure also shows a pooling layer of a CNN, which condenses the spatial information of the feature maps making CNNs computationally efficient



between 0 and 1, and the sum of all outputs is equal to 1. A softmax activation function can be used as a last layer for both CNN and MLP in classification tasks. Similar to MLPs, back-propagation in CNNs is performed to update the weights in each kernel, which are computed using the gradients of the loss function determined in forward pass.

Unlike MLPs, CNNs utilize three other important concepts including weight sharing, pooling layers and receptive field (see Figure 3). A weight sharing reduces the number of parameters by sharing weights for all neurons in a feature map. Pooling layers aggregate the neighboring pixel values to reduce the spatial dimension of the input images or the feature maps. A receptive field is a region in the input space that is affected by a kernel. The pixels of an image closer to the center of the receptive field contribute more to the output feature [6].

CNNs are immensely used in biophotonics for image and spectrum classification (see Figure 4), disease characterization and microorganism identification. These applications are further explained in section 3 and section 4. CNNs are also used in other deep learning architectures like auto-encoders and generative adversarial networks explained in section 2.4 and section 2.5, respectively.

2.3 | Recurrent neural network

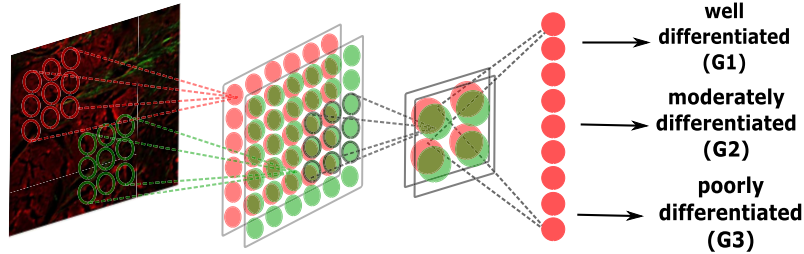
Standard neural networks like MLPs have certain limitations while working with sequence data like spectroscopic data or time series. One of the limitations is that

MLPs fail to consider the entire history of a sequenced input vector for obtaining an output [28] whereas, recurrent neural networks (RNNs) [17] incorporate neurons that span the input over time. Moreover, RNNs have hidden layers that add memory to the network over time.

RNNs can have three types of architectures to solve the sequence data problem: (a) the one-to-many RNN architecture has one input neuron and a sequenced or many output neurons, which are used for image captioning [29], (b) the many-to-one RNN architecture comprises a sequenced or many input neurons and one output neuron, which is used for text classification [30] and lastly (c) the many-to-many RNN architecture has a sequenced or many input neurons and a sequenced or many output neurons, which is mostly used for machine translation [31]. In addition to the earlier mentioned applications, RNNs have obtained promising results in natural language processing, speech recognition and machine translation tasks [32]. Moreover, a recent study reported the use of RNNs for the analysis of genetic data [33]. Despite the enormous development of RNNs, they are underexplored in the field of biophotonics as compared to MLPs and CNNs. Nevertheless, RNNs can build intelligent systems and its use in spectrum preprocessing, wavenumber calibration or intensity calibration, spectrum classification, decoding biomolecular markers from bio-spectroscopic data, learning spatial-spectral-temporal features for spectral data and phase retrieval of nonlinear optical spectroscopic data can be investigated in the future.

A typical many-to-many RNN structure is shown in Figure 5. The figure shows three unit types, an input vector, a hidden state vector and an output vector. For

Application: Image classification



Application: Cell localization and segmentation

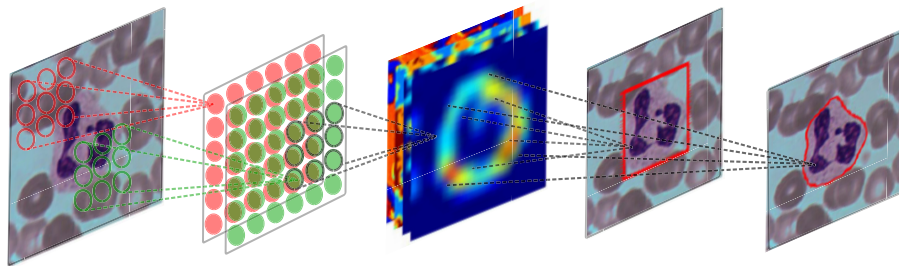


FIGURE 4 This figure shows application of CNNs like image classification (upper panel), localization and segmentation (lower panel). For image classification, a multiphoton image is used for classifying three grades of hepatocellular carcinoma (upper panel). For cell localization task, a leukocyte mask was generated using a CNN to localize and segment leukocytes in blood smear images (lower panel). These images are reproduced and modified from references [26, 27]. CNN, convolution neural network

sequenced input data ($\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_T$), an RNN can have many outputs ($\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \dots, \mathbf{y}_{T+N}$) or the same number of outputs like the input data ($\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \dots, \mathbf{y}_T$) or just one output unit y . The intermediate layer represents the hidden state of the RNN. The hidden state \mathbf{h}_t is the memory of the network and is calculated using the hidden state of the previous step $\mathbf{h}_{(t-1)}$ and the input vector at the current step \mathbf{x}_t . The hidden state at the first time step is initialized with zeros

$$\mathbf{h}_t = 0 \text{ for } t = 0. \quad (5)$$

The hidden state for the intermediate time steps is calculated by

$$\mathbf{h}_t = \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{t-1} + b) \text{ for } t \neq 0. \quad (6)$$

Here, \mathbf{U} is the weight vector of the hidden layer, \mathbf{W} is the weight vector of the input layer and \mathbf{V} is the weight vector of the output layer shared over time (see Figure 5). Applications of many-to-many RNNs for spectra preprocessing, where the input vector for the many-to-many RNN is a raw spectrum and the output vector is a preprocessed spectrum, still requires investigation.

However, many-to-one and many-to-many RNNs can also be used for classification purposes. In such cases, a softmax activation layer is added to the output sequence of the RNN model in order to achieve posterior probabilities for the classes.

Nevertheless, standard RNNs report some shortcomings. Firstly, RNNs require higher computational power and larger training data than usual CNNs. A standard RNN calculates an output at each time step utilizing just the past and the present element of the input vector. For spectroscopic data, the past, present and future states (or wavenumbers) of the spectra influence the output at a particular time step, and the application of bidirectional RNNs can be investigated. A bidirectional RNN utilizes hidden states from opposite directions to update the output sequence at a particular time step. Another shortcoming of RNNs is the problem of vanishing gradients, which occurs due to the deep structure of RNNs. To circumvent this problem, other variations of RNN including long short-term memory (LSTM) and gated recurrent unit (GRU) networks are used and have achieved better performances [34]. A comprehensive discussion of the variations of RNNs is out of scope of this review.

2.4 | Auto-encoder

Auto-encoders (AE) [35, 36] are ANNs consisting of two parts: an encoder and a decoder. The encoder transforms a D dimensional input $\mathbf{x} \in \mathbb{R}^D = \chi$ to a N dimensional hidden states $\mathbf{h} \in \mathbb{R}^N = \mathcal{F}$, $N < D$, where χ is the input space and \mathcal{F} is the latent space representation. The latent space \mathcal{F} is represented by the bottleneck of the model (see Figure 6). The bottleneck layer compresses the input space representation χ to capture the most salient features of the input data. The representation of the hidden states \mathbf{h} in the bottleneck layer can be written as

$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}). \quad (7)$$

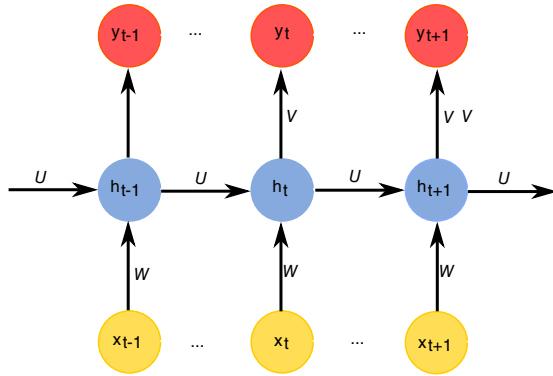
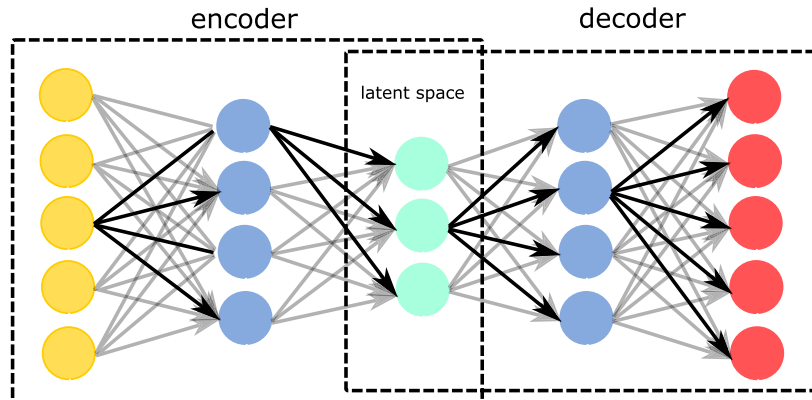


FIGURE 5 A structure of recurrent neural network is shown. A set of sequenced data \mathbf{x} with T time steps is given as an input (yellow) to reconstruct a sequenced output vector \mathbf{y} (red) with equal number of time steps. The hidden states (blue) store the features or act as memory unit of the RNN network. The weight matrices W, U, V are updated during the training of RNNs

FIGURE 6 An auto-encoder (AE) structure with two parts, an encoder and a decoder, is shown. An encoder transforms the input information (shown in yellow) into a latent space representation (shown in cyan) \mathcal{F} , which is transferred by the decoder to reconstruct an output (shown in red) in the same space representation as the input. Both the parts can be constructed using a CNN or an MLP. CNN, convolution neural network; MLP, multilayer perceptron



The dimension of the bottleneck layer is smaller as compared to the dimensions of the input layer to avoid the encoder from learning an identity function.

A decoder transforms the bottleneck features of the hidden states \mathbf{h} back to a reconstructed input \mathbf{x}' of the same dimension as \mathbf{x} . The reconstructed input \mathbf{x}' can be given as

$$\mathbf{x}' = \sigma(\mathbf{W}'\mathbf{h} + \mathbf{b}'). \quad (8)$$

Here, \mathbf{W}' and \mathbf{b}' are the weight matrix and bias of the decoder respectively. The training of an auto-encoder is performed through back-propagation of reconstruction error calculated between the original and the reconstructed input.

Traditionally, auto-encoders were used for dimensionality reduction [6]. Simple auto-encoders find its application for denoising, image deblurring and semantic segmentation (see Figure 7), which will be discussed in section 3 [39]. Additionally, variations of auto-encoders like stacking auto-encoder, sparse auto-encoder, denoising auto-encoder, convolutional auto-encoder, variational auto-encoder and contractive auto-encoder are used to prevent the learning of an identity function by the encoder, as stated earlier [40]. Moreover, auto-encoders can be a part of adversarial networks discussed in section 2.5.

2.5 | Generative adversarial network

A generative adversarial network (GAN) [41] is a special type of ANN that consists of two networks: a generator and a discriminator, which are trained simultaneously. The input to the generator is either a random noise vector \mathbf{z} or a real data, like an image \mathbf{X} , sampled from a prior distribution p_{data} . The generator is

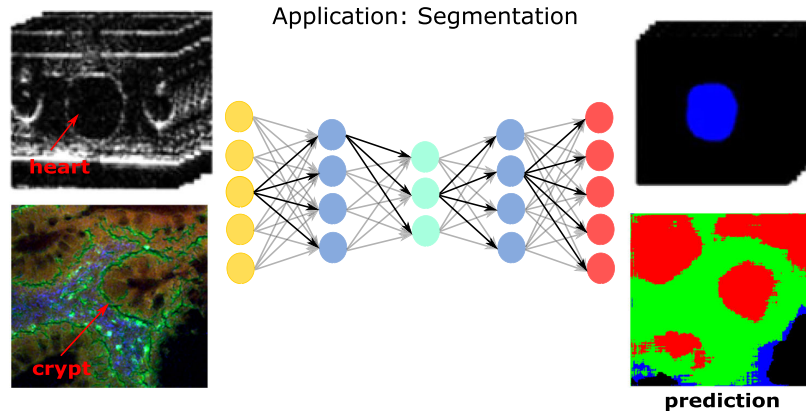


FIGURE 7 Applications of auto-encoders are shown in this figure. The upper panel shows a segmentation of a *Drosophila* heart, where optical coherence tomography images are used as an input to the AE model and the output is a segmented *drosophila* heart. Similarly, the lower panel shows a segmented crypt in a nonlinear multimodal image. The nonlinear multimodal image is used as an input to the AE model to obtain a false-color image with four distinct regions (crypt region in red) at the output. Similar to Figure 2, the nonlinear multimodal image is composed of CARS, TPEF and SHG signal. The images are reprinted from references [37, 38] with permissions. AE, auto-encoder; CARS, coherent anti-stokes Raman scattering; SHG, second-harmonic generation; TPEF, two-photon excited fluorescence microscopy

a differentiable function represented by an MLP (or an AE) that maps this input to an output y_G , such that $\mathcal{G}(\mathbf{x}; \Theta_G) : \{\mathbf{X}, \mathbf{z}\} \rightarrow y_G$. The generator $\mathcal{G}(\mathbf{x}; \Theta_G)$ aims to learn the distribution p_G to approximate the prior distribution of the real data p_{data} from where the input \mathbf{X} was drawn. The output y_G of the generator has visual similarity with the real data, e.g. images. In addition to the output from the generator, a real input image is also fed to the discriminator \mathcal{D} . The output of the discriminator $\mathcal{D}(y_G; \Theta_D) : y_D \rightarrow [0, 1]$ represents a probability that y_G is retrieved from p_{data} rather than p_G (see Figure 8). Both the networks \mathcal{G} and \mathcal{D} follow a min-max game where \mathcal{D} minimizes the probability of y_G belonging to p_{data} , and simultaneously \mathcal{G} maximizes this probability by generating more realistic images that cannot be distinguished by \mathcal{D} . This adversarial training is achieved by optimizing the loss function

$$E(\mathcal{G}, \mathcal{D}) = \mathbb{E}_{\mathbf{x}, y_G} [\log \mathcal{D}(\mathbf{x}, y_G)] + \mathbb{E}_{\mathbf{x}, \mathbf{z}} [\log (1 - \mathcal{D}(\mathbf{x}, \mathcal{G}(\mathbf{x}, \mathbf{z})))] \quad (9)$$

with back-propagation technique. During back-propagation, the gradient calculated over the loss function is back-propagated from the discriminator to the generator, in order to update the parameters of the generator. While training a GAN network certain challenges are encountered. Foremost, it is difficult to obtain convergence of both the networks due to

simultaneous training of the networks. Additionally, an early convergence of the discriminator network can cause the generated images to be easily distinguished from the true images. This is a consequence of the gradient of the discriminator reaching zero and thus providing no guidance to the generator for further training. After a few iterations, when convergence between the two networks is achieved, ($p_G = p_{\text{data}}$ and $\mathcal{D}(\mathbf{x}) = \frac{1}{2}$) the generator can produce realistic images, which are difficult to identify as “fake” images [41] by the discriminator.

Such an adversarial training of GANs have gained popularity in industrial and academic research due to their capability of domain adaptation and generating new images. Generative adversarial networks (GANs) are potentially used for biophotonic applications including denoising of images, correcting stitching artifacts in microscopic images, increasing spatial resolution [42, 43], virtual H&E staining of fluorescence images [44] and biological image synthesis of fluorescence images [45, 46] (see Figure 9). The applications of GANs are elaborated in chapter 3.

All the above mentioned deep learning architectures are huge and have many layers. With increasing architecture and dataset size, the memory requirements increases as well. Therefore, high computational power and efficient software are needed. A detailed explanation of the hardware requirements and commonly used software is given in the section 2.6.

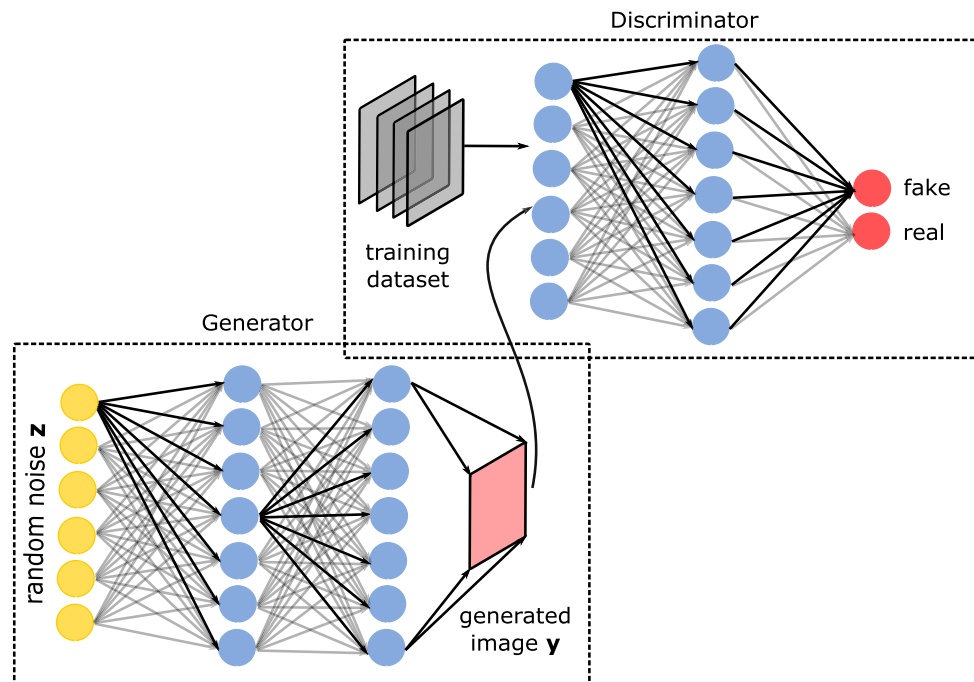


FIGURE 8 Generative adversarial network shows two adversaries, a generator and a discriminator. A generator's input is either random noise z or an image X . The output from a generator y_G , is fed to the discriminator D which distinguishes the generated output as real or fake. Both the networks are adversaries of each other as both the networks optimize different objective functions

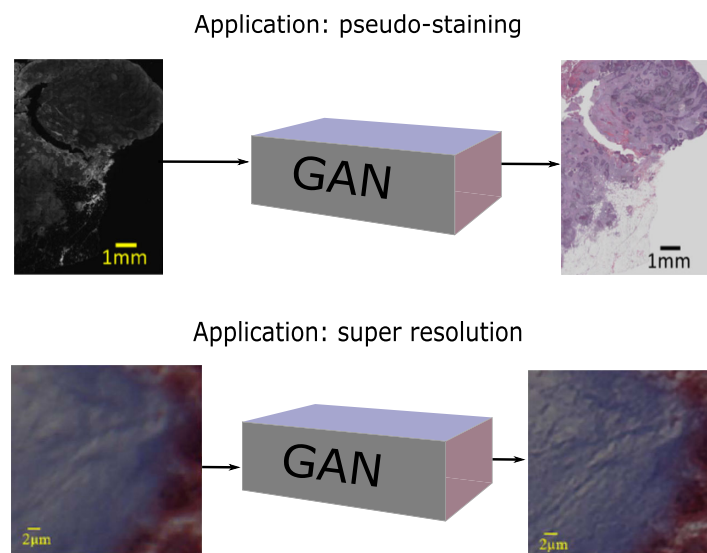


FIGURE 9 The two common applications of GANs including pseudostaining (upper panel) and resolution enhancing (lower panel) are shown. The image in the upper panel utilizes an autofluorescence image as an input and the GAN network produces H&E stained image at the output. Similarly, the image in the lower panel shows that a GAN model was used to enhance the resolution of a Masson's trichrome stained lung tissue section. The images are reprinted from earlier researches [42, 44] with permission. GAN, generative adversarial network

2.6 | Hardware throughput and software libraries

Deep learning algorithms perform complex matrix multiplications of millions of parameters in the hidden layers. This limits the performance of deep learning models due to the need for higher computational power and memory size. The recently introduced GPUs provide higher computational power as compared to conventional CPUs, thereby, accelerating the training of deep learning models to a greater extent.

In addition to the hardware, the availability of various software packages can facilitate the use of deep learning models in biophotonics. A range of open-source deep learning libraries like Caffe [47], Torch [48], Theano [49], Tensorflow [50], Keras [51] and Lasagne [52] are developed along with their interfaces in C++, Python and Lua programming languages. These packages can be efficiently implemented with GPUs, thus accelerating the training of deep learning models. Various researches using these libraries have been conducted for spectroscopic data which is discussed in section 3 and section 4.

2.7 | Educational resources

The above sections provide brief information about deep learning and various architectures. However, to make deep learning algorithms profitable for the biophotonic community, various educational resources are mentioned in this section.

In this context, books by Nielsen [53], Ripley [7], Russel and Norvig [8], Bishop [54], Goodfellow [6] and many more [55–58] are recommended sources for deep learning. Additionally, there are several online courses for deep learning which can give an overall hands-on experience of using various deep learning models with Python and R programming languages. Some of them are listed here: Coursera (<https://www.coursera.org>), deeplearning.ai (<https://www.deeplearning.ai>) and others (<https://www.datacamp.com>, <https://machinelearningmastery.com>, <https://www.pyimagesearch.com>).

Furthermore, applications of deep learning are showcased at a number of international conferences dedicated to biophotonics. A few of them include, but are not limited to, SPIE conferences, OSA conferences, IEEE conferences and FACCS conferences. Likewise, many peer-reviewed journals fully dedicated to the field of biophotonics have embraced the applications of deep learning and attracted interdisciplinary readership.

In the next two sections, applications of deep learning are elaborated.

3 | DEEP LEARNING FOR BIOPHOTONIC IMAGING

In the past decade, biomedical optical imaging has witnessed a vast development ranging from fast scanning systems to automated image analysis algorithms. In addition, developments like increased penetration depth, molecular specificity, faster image acquisition and high spatial resolution are advantageous for bed-side patient monitoring and diagnostics for personalized treatments. However, due to practical limitations of optical systems, certain challenges are encountered with the fast acquisition of highly resolved and noise-free data. Recently, deep learning algorithms have been used to address these unmet needs in biophotonic imaging and has shown overwhelming results for a broad range of applications. These applications will be further discussed in this section.

3.1 | Image denoising/deblurring

Deep neural networks can be designed for virtually any kind of input–output combination. One way to employ deep neural networks is to feed noisy or low-resolution images to the input of a generative network and use the images with desired resolution or noise level as an output. The generative network, which learns features from the high-resolution images, can be subsequently used for image enhancement. Generative networks using mean square error or similar type of loss function often lead to overly smoothed images at the output. A common way to preserve high-frequency features is to build a generative adversarial network (GAN), which was described in more details in section 2.5. Shortly, the GAN network contains a generative network to produce an image and a discriminator network to estimate the quality of the image produced by the generator. A variation of this architecture, called Wasserstein generative adversarial network (WGAN), which uses Wasserstein distance as a loss function, was recently utilized for resolution enhancement of OCT images [59]. An alternative, edge-sensitive conditional generative adversarial network (cGAN) was reported efficient against speckle noise. This speckle noise reduction was demonstrated for OCT images which utilized an edge-sensitive cGAN [60]. Another implementation of the GAN approach with additional content loss metrics was proposed for simultaneous denoising and super-resolution generation of optical coherence tomography [61]. This content loss was calculated from the difference between features extracted from the true image and the generated image. Besides OCT images, the GAN approach was successfully applied for fluorescence

microscopic images, making a cross-modality super-resolution possible without employing overly sophisticated setups [43]. The approach of achieving super-resolution by deep learning is additionally discussed in section 3.7.

In all above examples, deep neural networks learned patterns from the data, which makes it possible to increase the resolution and the signal-to-noise ratio simultaneously. This makes these methods advantageous in comparison with classical image enhancement methods, which usually improve one of the two quality parameters at the expense of the other parameter.

3.2 | Semantic segmentation

Semantic segmentation is a pixel classification task, where every pixel of an image represents a class. Semantic segmentation is widely used in digital pathology for applications, like tissue segmentation, nuclei segmentation and lesion detection [62]. Similarly, semantic segmentation of microscopic images, like nonlinear multimodal images [37, 63], OCT images [64] and fluorescence images, using auto-encoders (see section 2.4) is gathering researcher's interest. The above-mentioned works utilize U-net [65] type networks, which is an auto-encoder architecture with special connections between the encoder and the decoder network. Another striking feature of U-net is the weighted loss function, which heavily penalizes the misclassification of boundary pixels of an object, thus allowing to segment closely located objects efficiently. Previous research showed the semantic segmentation of nonlinear multimodal images (CARS, TPEF, SHG) of lung tissue using the U-net [63] architecture and of gastrointestinal tissue regions using the SegNet architecture (see Figure 7 bottom) [37, 66], respectively. Similarly, the authors of a recent research article [38] segmented a *Drosophila* heart in optical computed tomography images based on an U-net architecture (see Figure 7 top). Furthermore, it is shown in a recent work [37] that CNN based semantic segmentation achieved better performances as compared to traditional machine learning methods.

In addition to CNNs, recurrent neural networks have also shown promising results for semantic segmentation of the CamVid dataset [67, 68]. A recent work [69] used RNNs for perimysium segmentation in H&E stained skeletal microscopic images and achieved better performance as compared to the U-net architecture. RNNs can retrieve global spatial information of an image, which improves the semantic segmentation performance [69]. However, training a RNN can be computationally expensive and therefore it is underexplored in biophotonics.

3.3 | Disease recognition

Disease recognition using MLPs and CNNs is a very common application in the field of biophotonics. Out of all deep learning architectures discussed in section 2, MLPs are widely used for disease recognition and assessment. For example, MLPs were used to classify FLIM data of cervical neoplastic tissue sections, which achieved a significant discrimination between the normal and the pre-cancerous group as well as between the low-risk group and the high-risk group [70]. Another application of MLPs was reported using Raman spectroscopic data for classification of patients with Alzheimer's disease, other types of dementia and healthy individuals. Comparison of MLP results with conventional classifiers, like the radial basis function (RBF) classifier, showed that MLPs outperformed the conventional classifiers for the tested classification tasks [71]. In addition to MLPs, CNNs are the second most widely used deep learning architectures for disease classification. A recently proposed CNN application [72], classified malaria infected blood smear from healthy controls using Leishman stained images. The malaria-infected images were further used to segment the infected RBCs. Similarly, a very recent research reported the use of CNNs to assign cervical cancer into three stages using CARS, SHG, TPEF microscopic data [73].

Mostly, the data acquired by spectroscopic techniques is small, due to larger acquisition times. Therefore, the training of MLPs or CNNs is always challenging due to the small datasets available. In such cases, deep learning networks using transfer learning strategies can be applied [26, 74–76]. Transfer learning utilizes CNN models pretrained on a (large) source dataset and transfer the learned features to classify a (small) target dataset. For example, pretrained CNN models including GoogleNet [77], Inceptionv3 [78] and VGG16 [79] were used to classify breast cancer in OCT images [75], head and neck cancer in 3-D OCT images [80], lung cancer in CARS images [74] and hepatocellular carcinoma in multiphoton microscopic images [26] (see Figure 4 top), respectively. Here, the CNN models are first trained on large nonbiological datasets like ImageNet [81] and the parameters of these pretrained models are fine-tuned on the new biophotonic dataset.

To summarize, MLPs and CNNs are majorly used for disease classification. Recently reported transfer learning strategies [26, 74, 75] using CNNs have worked best for small datasets, which are accounted in biophotonic studies.

3.4 | Cell or organ localization

In addition to the segmentation tasks described in section 3.2, there is a specific segmentation application

known as the “localization” task. In biomedical imaging, localization can be used for counting cells of a specific type within the sample or its image. Subsequently, the localized cells can be segmented and analyzed through descriptive statistics over cell sizes, shapes and the cell morphology. Alternatively, segmented cells can be automatically classified or investigated manually by pathologists. It was shown that leukocytes can be localized within blood smear images and segmented using deep neural networks efficiently [27]. For the leukocyte localization, a multistep workflow that included a feature extraction by a feature pyramid network inspired by the ResNet architecture [82] was utilized. This was followed by the determination of a region of interest. Thereafter, a localization box was predicted and the leukocytes were segmented (see Figure 4 bottom). On every step of this workflow, convolutional or fully connected layers were used instead of user-defined features.

Another biomedical application of deep learning is organ localization within 3D computed tomography (CT) scans, which is an essential preprocessing step for the analysis of the scans. Recently, the organ localization and segmentation within 3D scans was demonstrated using a 3D U-net approach [83] and a 2D multichannel SegNet model [84].

3.5 | Pseudostaining

In imaging of biological tissue and cell samples often histological staining needs to be applied in order to enhance the contrast and highlight tissue features. This staining is usually performed during the sample preparation prior to the microscopic investigation of the sample. Both manual and automated microscopic image analysis often require such stained images. Some stains, like the hematoxylin and eosin (H&E) stain are used over many decades as “gold-standard” techniques in pathology. The main drawback of the conventional staining techniques is that they require additional time and effort. Recent studies showed that in certain cases deep learning can be employed instead of the actual sample staining. It was shown that cGAN architecture can be used to generate H&E stained images from hyperspectral microscopic images of unstained samples [85]. Another study employed a CNN-GAN approach in order to obtain H&E stained images from unlabeled tissue autofluorescence images (see Figure 9 top) [44]. Both studies performed virtual H&E staining by using different imaging techniques in a combination with deep learning instead of actual staining the sample. On the other side, it was shown that deep learning makes it possible to restain H&E stained microscopic

images into immunohistochemical (IHC) stained images [86]. The advantage of such approach is that H&E is a conventional and simple staining but the IHC staining is more costly and labor intensive. For such restaining, a conditional CycleGAN (cCGAN) architecture was used. Oversimplified, this CycleGAN approach is a combination of two generators (encoder and decoder) and discriminators. The first generator produces an IHC stained image from a H&E stained image, subsequently, the second generator transforms the generated IHC image into a virtually stained H&E image. This cycle makes it possible to introduce cycle identity loss and classification cycle loss in the network architecture.

In prospect, deep learning in combination with various imaging techniques may provide a fast and flexible alternative to the histological staining, making it possible to switch between virtual stains without additional sample preparation and measurements.

3.6 | Image registration

Nowadays, it is a common practice to measure one sample with multiple modalities in order to achieve a comprehensive characterization of the biological tissue specimen. For the joint analysis of the images obtained from two or more modalities, a perfect overlay of the two images is required. This is termed as image registration. The basic idea of the image registration methodology is to minimize or maximize an objective or a cost function computed on the overlapping region of the two (moving and fixed) images. The optimization of the objective function is achieved by iteratively searching a geometric transformation for the moving image. Various semiautomatic approaches have been proposed by researchers to register secondary ion mass spectrometry images with optical image [87], Raman microscopic images with mass spectrometric MALDI-TOF images [88] and FTIR images of tissue microarray (TMA) cores against H&E images [89]. However, these methods are not fully automated and require manual intervention. A recently developed automatic approach based on a sparse search strategy deals with sub region registration of FTIR microscopic images in whole-slide histopathological staining images. Additionally, the FTIR imaged cores of tissue microarrays were registered with their histopathologically stained counterparts. This work also presented the registration of CARS images within histopathological staining images [90]. Although this approach is robust and reliable for diverse microscopic technologies, it requires preprocessing of the samples acquired from various modalities. In such cases,

CNN based registration can potentially register images obtained from different modalities without the need of image preprocessing.

Recently, CNN based registration methods have been reported in radiology, which learn geometric transformation parameters for registering MRI and CT images [91, 92]. The results from the CNN based registration have shown surprisingly good results and are efficiently applied in a multiresolution scenario. However, a CNN based image registration of spectroscopic images is still under-explored and requires further investigation.

3.7 | Image super-resolution

An earlier section (see section 3.1) discussed that GAN architectures can be employed for both image denoising and resolution enhancement. Although the improvement of the signal-to-noise ratio is very important for the interpretation of the images, it can also be improved by increasing the number of collected images. On the other side, the resolution of the obtained image is often limited due to the technical properties, like the diffraction limit. There are various sophisticated technical solutions, which allow an imaging below the diffraction limit. A class of such techniques is called super-resolution imaging.

Besides technical solutions, overcoming the diffraction limit is also possible by employing image processing techniques, and in particular, deep learning. Studies showed that CNN can be applied to effectively improve the resolution of the stained tissue section (see Figure 9 bottom) [42]. A fully convolutional encoder-decoder network was successfully constructed for imaging of quantum dots and microtubes using single-molecule localization microscopy [93]. Another imaging limitation was pushed by deep learning in the area of lens-free holographic microscopy (LFHM). Due to the absence of the lens, the resolution is limited by the pixel size of the detector. To overcome this issue a CNN network, inspired by an U-net architecture was employed for LFHM, which made it possible to perform pixel super-resolution imaging [94]. Another example of generating super-resolution images was implemented for OCT images using a GAN-based approach [61]. Besides achieving super-resolution, this GAN-based approach decreased the image noise simultaneously.

In addition to the above-mentioned applications, deep learning is vastly applied for vibrational spectroscopic data including applications like preprocessing and classification of spectra. These applications are discussed in the following section.

4 | DEEP LEARNING FOR VIBRATIONAL SPECTROSCOPY

Until recently, data analysis in vibrational spectroscopy employed well-established classical machine learning techniques adapted to the structures of specific spectroscopic data. The general workflow in these scenarios is composed of preprocessing, feature extraction or feature selection and statistical modeling [95]. In contrast to the widespread use of artificial neural networks in spectral analysis [96–98], the application of deep learning in this field is growing but still in the early stage. This is because, on the one hand, classical machine learning does a great job in most cases, and on the other hand, deep learning in spectral analysis encounters many difficulties. Most of the existent deep neural networks were developed for image analysis or speech recognition and cannot be directly transferred to spectral analysis. Building a deep neural network for spectral analysis from scratch requires a lot of hyperparameter tuning and is tedious. Unlike in image analysis, there is rarely a pre-trained deep learning model for spectral data. The lack of large spectral datasets forms another difficulty to apply deep learning in spectral analysis. Nevertheless, the spectral analysis does see benefits from deep learning, which will be discussed in the following section from the perspectives of spectral preprocessing and statistical analysis.

4.1 | Preprocessing

Spectral preprocessing aims to remove corrupting contributions from the measured spectra, which is often done by smoothing, baseline correction, standardization, and so on. Preprocessing is a burden, not only because of the computation time, but also because it is not straightforward to select the preprocessing techniques that perform best on each specific dataset [99]. Deep learning can be a time saver assuming that the deep neural network is powerful enough to tolerate the corrupting effects and can be trained on raw data without any preprocessing to reach a satisfying performance. This has been shown in references utilizing convolutional neural networks or stacked contractive auto-encoders [100–104]. The kernels of the trained network were shown to work as smoothing, derivative/slope recognizers, thresholding and spectral region selection, which are basically preprocessing steps [101]. Unlike conventional preprocessing approaches, however, the outputs of the kernels are not necessarily physically meaningful, but rather a mathematical representation of preprocessing for the given data. This representation is best suited for the following regression or

classification models. Nevertheless, a close inspection of the outputs of the kernels does give a hint about the features that are the most significant for the regression or classification [101, 103].

While most investigations are engaged to construct deep learning methods utilizing the raw data and skip preprocessing, there are indeed efforts to apply deep learning as a preprocessing approach, especially for issues that cannot be solved easily with conventional preprocessing methods. As it is widely known, a sufficiently long integration time is normally needed for a usable spectrum, especially for Raman spectroscopy considering the small Raman cross-section. The slow measurement, especially in the case of Raman imaging, has hindered Raman spectroscopy to be applied for the investigations of dynamic processes. In such cases, fast measurements are needed but they suffer from bad data quality, such as extremely high noise or low spectral/spatial resolution. Deep learning has shown its capability of handling this issue in recent publications [105, 106]. For example, an U-net was applied to stimulated Raman spectra to reduce noise in the data and hence improve the sensitivity, which helps shorten the spectral acquisition time down to 20 μ s without losing sensitivity [105]. In another investigation [106] the authors applied a deep convolutional neural network to improve the spatial resolution of the Raman hyperspectral data. In this way, the line-scan Raman measurement was largely accelerated.

Following the spectral preprocessing, investigating the spectral data by using multivariate statistics and classification models is commonly performed. The next section discusses the statistical modeling of spectral data using deep neural networks.

4.2 | Statistical modeling

It is commonly hypothesized that deep neural networks are capable of feature learning [107], that is, they do not require hand-engineered features, which are needed to apply conventional classifiers. With multiple layers of linear and/or nonlinear units, deep neural networks show huge potential to learn hierarchical representations of features from complex data. It is thus advantageous to apply deep neural networks for the analysis of vibrational spectra, which are a complex superposition of all vibrational information within the sample. Applications of deep learning were reported for both infrared and Raman spectroscopy in order to achieve tasks like brain function investigations [108, 109], biological diagnostics [102, 110, 111], cytopathology [112], microbial identification [113], pathogenic bacteria identification [113], food science investigations [114, 115], tobacco leaves characterization

[116] and mineral analysis [117]. Furthermore, it was reported in references that deep learning can perform better than classical machine learning methods [100, 103]. A deep convolutional neural network was also used for an un-mixing tasks, i.e., to resolve pure components and their abundances from mixture spectra. Thereby, N one-component identification models were trained with data composed of spectra of a pure component, negative and positive samples in terms of this pure component. The N models could successfully solve the un-mixing task at the end [118].

In addition to the different applications discussed above, strategies were reported to improve the performance of deep learning. In particular, a hierarchical deep convolutional neural network was employed on Raman microscopic data, in which neighboring spectral pixels were merged hierarchically in order to combine the spatial information with spectral information. This combination finally led to a better classification between healthy and cancer cells [112]. In addition, different searching algorithms such as grid search [103], particle swarm optimization (PSO) [114] and artificial bee colony algorithm (ABC) [117] have been utilized to automatically find the optimal hyperparameters of a deep neural network. A combination of a CNN and an extreme learning machine (ELM) was reported to speed up the training and improve the generalization performance of the trained network. The optimal values of ELM were sought by an artificial bee colony algorithm (ABC) [117].

Despite the investigations included in previous paragraphs, deep learning is far less developed in vibrational spectral analysis in comparison with image analysis and speech recognition. One of the reasons is that the deep neural networks are extremely data starving, but measuring spectral data from a large number of samples is limited by practical reasons, especially for biological samples. Data augmentation can be utilized to solve this issue, which is normally done by randomly shifting the wavenumber axis, adding random noise and/or (linearly) combining multiple spectra [100, 101]. However, these data augmentation techniques can introduce unknown (spectral) features into the data, especially if the variations of interest are very subtle. This is perhaps the reason, why the best model achieved in reference [101] was trained by utilizing an additional EMSC after data augmentation. A generative adversarial network may play a role for better data augmentation, but there is yet no application reported to the authors' best knowledge.

Besides the intrinsic complexity of the spectra and limited sample size, vibrational spectroscopy is remarkably sensitive to measurement conditions and there exist significant variations among multiple measurements. Hence, it is important during spectral analysis to learn

features of interest but not those related to the measurement in order to achieve an optimal prediction on new measurements. Deep learning can play a role in this context as it was reported in previous research [101]. Therein a CNN was used to predict a test dataset comprising of drug concentrations higher than the concentrations of the training dataset. In this case, the test performance of the CNN can be improved, only if the hyperparameters of the network was tuned based on the validation set. Tuning with a randomly selected validation set did not provide significantly better predictions. In fact, it is more than difficult to build a deep neural network, which tolerates unwanted variations and generalizing well between measurements. Data augmentation can help in this situation, as it was discussed in reference [101], but the improvement was limited. Another strategy is transfer learning, which has been discussed in the previous section. Its capability for dealing with unwanted spectral variations was shown in reference [102] where the deep network was pretrained on embedded tissues and finetuned to classify fresh frozen tissues.

Another important issue of applying deep learning for vibrational spectral analysis is a proper validation. As it was mentioned in the last paragraph, vibrational spectra often vary from measurement to measurement and device to device. It is thus important and necessary to validate a deep neural network using measurements independent to the training data. A random separation between training and testing data should be avoided. In addition, the testing data cannot be included in any procedure that affects the final modeling, including model-based preprocessing such as EMSC [119]. Otherwise, an overestimation of the network is highly possible. Similar challenges and issues related to deep learning methods are discussed in the next section.

5 | DISCUSSIONS AND CRITICAL ISSUES

Deep learning was already applied several times in biophotonic data analysis, but its potential is much larger. To use this potential an immense amount of data for training is needed. If such large datasets are not available, then increasing the dataset size by data augmentation or using transfer learning methods to achieve good model performances are commonly used approaches. Furthermore, class imbalances are predominantly seen in clinical studies, which affect the training of deep neural networks. Another issue about using deep learning methods is the lack of interpretability of model predictions, which restricts the use of deep learning methods for newly developed measurement modalities in the biophotonic

field. Additionally, proper model validation techniques are needed, which will be elaborated in this section.

5.1 | Current challenges

This subsection elaborates the challenges which are related to the dataset, training and understanding of the deep neural network encountered by data scientists in biophotonics.

5.1.1 | Lack of data

Biophotonic technologies are emerging techniques with restricted use in clinical practice as compared to other radiological and conventional histopathological techniques. Therefore, the dataset size is often limited. Moreover, the systematic accessibility of data and open repositories is limited in the biophotonics field. This leads to one of the major challenges to use deep learning for biophotonic data, which is the shortage of data. Deep learning models are data-driven and require a large amount of data depending on the task and the number of parameters in the model [120, 121] (Table 2).

Small datasets can easily lead to over-fitting causing poor generalizability on a new dataset. The problem of small datasets can be overcome by increasing datasets using data augmentation techniques. The basic idea of data augmentation is to artificially expand the training dataset by creating modified versions of the original dataset. For example, commonly used data augmentation techniques for image data are translation, rotation, shifting, increasing or decreasing brightness and magnification of the images. Other commonly used data augmentation techniques for images are adding Gaussian noise and transforming the color space of the images [122]. Likewise, data augmentation of spectral data can also be performed by adding noise to the spectral data or shifting the wavenumber axis for spectroscopic data [100, 101]. However, it is worth noting that slight perturbations in the images or the spectra can also degrade the model performance [123]. To prevent the degradation of the model performance and also to avoid too large dataset sizes, we discuss some practical considerations with the perspective of data augmentation in section 5.2.1.

In addition to data augmentation, transfer learning is another alternative technique to train deep learning models on small datasets. This technique focuses on transferring features of a deep neural network learned on a larger dataset to a small dataset. Research has shown that transfer-learning strategies lead to promising results

TABLE 2 Abbreviation in alphabetical order

Acronym	Explanation
ABC	Artificial bee colony algorithm
AE	Auto-encoders
ANN	Artificial neural network
BRNN	Bidirectional recurrent neural network
CARS	Coherent anti-stokes Raman scattering
cGAN	Conditional generative adversarial network
cCGAN	Conditional cycle GAN
CGAN	Cycle GAN
CNN	Convolutional neural network
CPU	Central processing unit
DBN	Deep belief network
ELM	Extreme learning machine
EMSC	Extended multiplicative signal correction
GAN	Generative adversarial network
GPU	Graphical processing unit
GRU	Gated recurrent unit
H&E	Hematoxylin and eosin
IHC	Immunohistochemical
FLIM	Fluorescence life-time imaging
FTIR	Fourier-transform infrared spectroscopy
LFHM	Lens-free holographic microscopy
LSTM	Long short-term memory
MALDI-TOF	Matrix assisted laser desorption-ionization (time of flight)
MLP	Multilayer perceptron
MPM	Multiphoton microscopy
OCT	Optical coherence tomography
PAI	Photoacoustic imaging
PSO	Particle swarm optimization
RBF	Radial basis function
RBM	Restricted Boltzmann machine
ReLU	Rectified linear unit
RNN	Recurrent neural network
SAE	Stacked auto-encoder
SERS	Surface enhanced Raman spectroscopy
SGD	Stochastic gradient descent
SHG	Second harmonic generation
SMOTE	Synthetic minority over-sampling technique
TPEF	Two-photon excitation fluorescence
WGAN	Wasserstein generative adversarial network

when applied for small spectroscopic dataset [26, 74, 75]. However, transferring features of a deep neural network which is pretrained on a dataset like ImageNet, to

perform classification or regression tasks on spectroscopic data, is debatable. Prior research has shown that with increase in the distance of the tasks (like classification, regression) and domains (like biological, non-biological), the transfer of the specific features learned in the last layers of a deep neural network can negatively affect the model performance. Thus, leading to “negative transfer learning” [124]. A practical advice on applying transfer learning approaches on small dataset is given in section 5.2.2.

5.1.2 | Imbalanced dataset

A second challenge in training deep learning models is an imbalanced data distribution, which is a key issue in all biological datasets. Training a deep neural network with unbalanced datasets affects the sensitivity of the loss function towards the majority class. To circumvent such biases, data-level and method-level approaches are used. Data-level methods address the class imbalance problem by random over-sampling the minority class or under-sampling the majority class. Although data-level methods are simple, over-sampling can introduce over-fitting of the model and under-sampling can cause loss of important information. Another complex sampling method is synthetic minority over-sampling technique (SMOTE), which creates synthetic data for the minority class. However, this method is limited due to the issue of generalizability and variance [125]. Also, creating synthetic spectral data is not straight forward due to the complexity of the spectral features.

An alternative to this imbalance issue are model-level methods, which have significantly improved the training results of deep learning models. In these cases, the loss function is penalized by the weight of the classes, which is defined by the number of samples in each class. However, sometimes it is difficult to define a customized loss function for a multiclass classification task. Many researchers have reported the use of a hybrid approach, where data-level and model-level methods are combined. Furthermore, other methods dealing with the loss function to overcome class imbalances have also been reported in the literature [126, 127].

5.1.3 | Bias-variance trade off

The third challenge encountered while constructing any machine learning method is the bias-variance trade off. There is always a competition to find a perfect balance between high bias (under-fitting) and high variance (over-fitting) for complex models. Model complexity can

be defined as the number of trainable parameters in a model and an increase in number of trainable parameters also increases the model complexity. With increasing model complexity, like encountered in deep neural networks, the increase of variance is more likely. A high variance in deep learning models can be due to three major reasons: the first reason is the sampling variance, the second reason is the model complexity variance and lastly is the model initialization variance. Sampling variance is a consequence of a high biological variance between the samples and within a sample (eg, the variance between biological replicates and within the replicates). Therefore, acquiring more balanced data and maintaining a consistent data acquisition protocol is essential. Additionally, comparison of the data acquired in different laboratories and devices should be encouraged in order to avoid such biases.

Model complexity and model initialization variance is controlled by the depth and width of the deep neural networks. Research has shown that an increasing the depth of a deep neural network by adding layers to a neural network can be a source of over-fitting, whereas increasing the width of the deep neural network decreases the model-related variance [128]. Therefore, designing a deep neural network should be done with focus on the generalization capabilities of the model. Even though, bias-variance trade-off is also observed in classical machine learning models, research shows that deep learning methods can efficiently find a balance between the bias and variance [129, 130].

5.1.4 | Interpretability of the “black-box”

Deep learning models have achieved breakthrough performance in various domains of medical imaging including biophotonics (see section 3 and section 4). As these models are intended to be utilized in modern healthcare systems, the interpretation of their decision-making is a key issue. It is important to know if deep neural networks make their predictions based on the biomolecular information instead of some background effect or noise in the spectroscopic data. An example of missing interpretability of the “black-box” models can be seen in a recent research [37] where an auto-encoder like model was used to segment nonlinear multimodal images of CARS, TPEF and SHG into four tissue regions. The segmentation results from the auto-encoder were satisfactory compared to the classical machine learning approach using hand-engineered texture features. However, the contribution of the three modalities CARS, TPEF and SHG for the segmentation of crypts was unknown. Similarly, by using deep learning models the contributions of spectral

features to a prediction, like the presence or absence of a disease, is difficult to interpret. This drawback hinders the usage of deep learning models especially in newly developed biophotonic technologies. Nevertheless, researchers are now developing various decomposition techniques for understanding complex deep learning models [131–135].

A recent research [136] utilizes Taylor series expansion for interpreting the output function of nonlinear models like ANNs on Raman spectroscopic data. Within this approach, the degree of nonlinearity of ANN model was realized using a second-order Taylor expansion. This allowed an interpretation of the patterns learned by ANN models based on wavenumber combinations to predict a particular class. Another approach [131] uses the layer-wise decomposition of features from hidden layers to understand the contribution of all pixels in an image to detect a particular class. While all these techniques are mostly developed for computer vision tasks, its utility can be expanded for spectroscopic data and this needs further investigations.

5.1.5 | Standardization for biophotonics

Biophotonics has an outstanding potential for clinical healthcare. However, in contrast to the well-established radiological or histopathological techniques, biophotonic technologies lack the adoption of standard procedures. There are no international consensus of assessing the performance of biophotonic devices which largely affects the reproducibility of data. Subsequently, the machine learning models trained on such data are less reliable. In this regard, several publications [137–139] have presented standardization procedures for various biophotonic technologies.

Improving the quality of clinical studies, comparing data from different laboratories and systems, facilitating the use of open databases, allowing quantitative comparisons between different models are critical factors for developing the best computational models. Validating the strength of these machine-learning models is also important and is further discussed in section 5.2.4.

5.2 | Practical considerations: *do's* and *dont's*

Researchers often encounter challenges as it was discussed in section 5.1 while training a deep learning model. To overcome these challenges, various approaches including data augmentation, transfer learning and model validation are established. However, these

approaches have pitfalls that can generate poor deep learning models, increase the training time and cause memory issues. Thus, it is important that developers circumvent common pitfalls while constructing deep learning models. In the following section a practical advice for constructing these models and avoiding common mistakes are given.

5.2.1 | Data augmentation

The choice of data augmentation should be made depending on the dataset. Data augmentation strategies like horizontal flips, random rotations, scaling and shearing are simple to implement, however these strategies fail to add new information or patterns into the training dataset [140, 141]. Moreover, random rotations and translation can introduce zero values in the corners of the image, which causes a bias in training the deep neural network. Therefore, the image regions with zero values are removed or filled with a reflection of the original image. In addition to geometric transformations, adding noise like jitter or Gaussian noise has improved regularization properties of the deep neural network for medical image classification [140, 142]. For fluorescence images, Gaussian and Poisson noise are commonly observed. These can be simulated to generate synthetic fluorescence images. Another data augmentation technique is the style transformation using GANs, commonly known as style transfer. In style transfer methods, the color and texture information from one image is transferred to another image to generate a completely new image [141, 143]. However, style transfer in biophotonics requires systematic investigation, as it may cause subtle alterations in the color and texture of the newly generated image which are associated to the biomolecular information under investigation. Thus, data augmentation techniques like style transfer should be performed cautiously for medical imaging, because it may also require changing the labels respectively. Another method to create large datasets from small dataset is the extraction of patches of the images. This method was implemented in a recent research [37] for semantic segmentation of nonlinear multimodal images. Utilizing patches for data augmentation not only increases the dataset size but also retains the biomolecular information of the images without the need to change labels. However, extracting patches of large spectroscopic images fails to generate new independent data and contrarily increases the dataset size. This can cause memory requirement issues. In such cases, large images should be down sampled and non-informative patches should be removed.

As mentioned above, data augmentation can increase the dataset size and memory requirement depending on the data augmentation scheme applied. To tackle this issue, online and offline data augmentation strategies can be chosen. If the dataset is relatively small, offline data augmentation can be performed. Offline data augmentation increases the dataset size by a factor equal to the number of transformations performed. If the whole augmented dataset is used for model construction, it can increase the memory requirements. The second option is online data augmentation which performs transformations of the mini-batches used while training the deep neural network model. This approach reduces the memory requirements but increases the training time.

In addition to the above-mentioned points, there are further important considerations for data augmentation. First, data augmentation should be performed for the training dataset only. Moreover, all the images should be rescaled to the same size before adding any kind of noise and various levels of noise can be tested to achieve the best validation accuracy. Overall, the benefit of data augmentation in biophotonics is an open issue that should be investigated systematically.

5.2.2 | Transfer learning

The previous section explains that data augmentation is an effective method to work with small datasets and this section introduces transfer learning as a strategy for small datasets. There are two transfer learning strategies which are commonly followed: first, a pretrained deep neural network are used as feature extractor and those features are utilized to build an easy model for classification or regression. The second strategy is to fine-tune the weights of a pretrained deep neural network using the new dataset. Fine-tuning of the weights can be conducted for all the layers of the network or restricted only to the last layers where most specific features are learned. Based on the two transfer learning strategies, the size of the dataset, the similarity between the datasets and the similarity between the tasks (classification or regression) involved, four major approaches can be utilized [144]:

- If the new dataset is small and similar to the original dataset, then the generic features from the top layers of a pretrained deep neural network will be relevant for the new dataset and thus these generic features can be used to train an easy classifier.
- If the new dataset is large and similar to the original dataset, then fine-tuning of the whole pretrained deep neural network can be performed.

- If the new dataset is small and different from the original dataset, then it is best to train a linear classifier (linear discriminant analysis or support vector machine) by using activations from the top and intermediate layers of a pretrained deep neural network. Previous research reported that this method works best for small spectroscopic datasets [26, 74, 75]. However, for biophotonics this needs proper investigation depending on the dataset.
- If the new dataset is large and different from the original dataset, then it is beneficial to train a deep neural network from scratch and initialize the weights using a similar pretrained deep neural network model.

5.2.3 | Splitting the dataset

Splitting of the dataset depends on the dataset size. In many machine-learning applications, large datasets are divided into two parts: 80% training dataset and 20% test dataset. A classifier or a regressor will be fitted using the 80% training dataset and the performance of the model will be evaluated on the remaining test dataset. For small datasets, k -fold cross validation techniques are generally used, where the whole dataset is resampled k times to train the model k times and evaluate its performance on the unused fold. Although the cross validation techniques allow a proper estimation of the generalization performance of the constructed model, its use in deep learning is limited due to the large training time and memory requirement. Thus, in deep learning applications the dataset is mostly divided into three parts: training, test and validation dataset. The training dataset is used to fit the deep learning model. The validation dataset provides an unbiased evaluation of the fitted deep learning model and simultaneously optimizes the hyperparameters of the model. And finally, the test dataset is used for evaluating the performance of the final model fitted on the training dataset. The division of the dataset into parts should be made at the highest hierarchical level. For instance, in a clinical setting, the highest hierarchical level is at the patient-level or device-level. Images or spectra obtained from the same patient should be a part of either the training, validation or the test dataset, to avoid any training bias [145]. A training bias is introduced when both the training and validation dataset originate from the same source (patient or device), thus reaching a high training and validation accuracy but a poor test accuracy. In prospect, splitting the dataset plays a major role in training deep learning models. Thus, it is beneficial for the biophotonic community to encourage proper model validation.

5.2.4 | Model validation and assessment of model performance

Establishing common procedures for model validation is important for biophotonics as explained in section 5.1.5. This facilitates a fair comparison between different models and systems. It is a common practice to test a final model on a third “independent test set” (also referred to as “holdout set”) beside the “training set” and the “validation set.” The latter mainly serves the purpose of model selection and hyperparameter optimization [4, 7, 8]. However, this requires a lot of data which represents the whole underlying population. To deal with small datasets cross-validation using the k -fold strategy is a commonly used approach. [145]

While training a deep neural network, the accuracy on training and validation dataset rises gradually with the number of iterations. If not, then several possibilities are responsible to lower the performance including overfitting of the model on the training dataset, a small dataset size, a noisy dataset, the choice of hyperparameters and the depth of the model. In such cases, increasing the dataset by data-augmentation techniques, removing redundant data by filtering noisy images or spectra, optimizing the hyperparameters and performing cross validation can be considered. Nevertheless, reducing overfitting requires systematic studies depending on the dataset.

In addition to the above-mentioned techniques, early-stopping of the model training can also be utilized to improve the generalization performance [146, 147]. Early stopping is a regularization technique that stops the training of the deep learning models before the performance on the validation dataset begins to decline. In cross validation of deep learning the model with the best validation accuracy can be used to predict the test data. In the case of comparison of two or more models, the performances on test dataset should be reported.

5.2.5 | Reduce over-fitting

As explained earlier (see section 5.1.3), a deep learning model trained with high variance can predict well on the training data but shows a poor generalizability to the test data. Adjusting the generalizability and constructing robust models is done by reducing overfitting. This is often termed as “regularization” [6, 142] and can be achieved by several methods. Augmentation of training data explained in section 5.2.1 is often considered as one of the regularization methods [148]. Another method is to add dropout layers to the model. Adding dropout layer is based on the principle: “learn less to learn better.” In

this regularization technique, the outputs of some neurons in the hidden layers are ignored, thereby, forcing the remaining neurons to learn a sparse representation of the data [149, 150]. Several variations of the dropout method reported in the literature have shown to improve model performances [151–154]. In addition to the dropout methods, early stopping (explained in section 5.2.4) and weight regularization techniques are other regularization methods for reducing over-fitting.

Weight regularization like L1 and L2 regularization penalizes the model during training based on the magnitude of the learned weights [155, 156], because large weights of a deep neural network can be a sign of an unstable network [157]. Regularization techniques encourage the sum of absolute values of the weights (L1) or sum of squared values of the weights (L2) to be minimum and thereby generating sparse weights that reduce over-fitting. Another method to check over-fitting is to reduce the capacity of deep learning models by decreasing the number of layers in the model or number of parameters in each layer [4].

Besides these regularization techniques, batch normalization technique is a well-known method to overcome over-fitting of deep neural networks [158]. This technique standardizes the inputs to a layer of deep neural network for each mini-batch. In this way, training of the deep neural network is stabilized and the training process is accelerated [159].

In summary, all the earlier explained topics are complementary to each other with a common goal of reducing over-fitting and constructing robust deep learning models. However, the effects of each of these regularization methods on biophotonic data need systematic investigation.

6 | CONCLUSION AND FUTURE OUTLOOK

Biophotonics is a rapidly growing field with a great potential to be a part of clinical practice. Current technological advancements in biophotonics are pushing the limits by increasing the resolution of optical systems, achieving larger penetration depths and faster scanning speeds. Additionally, current optical systems are capable of probing from micro to macroscopic scales, detectors are becoming more specific and efforts for miniaturizing devices using fibers are observed [3, 160]. All these technological advancements are enriching the information content of the biophotonic data and advanced data analysis methods, like deep learning techniques, are needed. In this regard, researchers are developing deep learning methods for

various biophotonic applications, which were elaborated in this review article.

Out of all the contributions discussed in this review article, a majority of work includes deep learning methods for biophotonic image data, whereas deep learning for spectral data is still underexplored. Almost 60% of the research used image data for early detection of diseases and assessment of disease stages. The remaining work majorly focused on virtual staining, increasing the resolution of fluorescence images and segmentation of cells, tissues and organs in spectroscopic images. In addition, a small part of the reviewed papers focused on preprocessing and classification of vibrational spectroscopic data. Although deep learning methods are underexplored for spectral data, we foresee that its development for vibrational spectroscopic data can transform the biophotonics field. Therefore, we discuss some potential applications of deep learning to analyze image and spectral data in this review.

Deep learning architectures can be used for spectral classification without the need of complex preprocessing steps [100]. On the other side, architectures like RNNs can be used for spectral preprocessing including denoising or despiking. Due to the basic similarities in the shape of the spectra, classification models can be trained with spectral data obtained from different domains using transfer-learning methods [100]. We speculate that transfer learning can complement the model-transfer methods [161] built for spectroscopic data by transferring high-level features of training data obtained in one domain to new data acquired in another domain. Until now, transfer learning methods have proven beneficial for fluorescence imaging data especially for cases where large datasets were not available [26, 74, 75].

Deep learning for vibrational spectroscopy has some challenges like the lack of data, the complexity of spectra, inter and intra-class-variances within the spectra and interpretability of the deep learning models. The issue of lack of data can be addressed by creating and facilitating access to large databases of spectroscopic data and efforts have already been initialized in this direction. Recent studies have reported large databases comprising images of three modalities including confocal, two-photon and wide-field fluorescence microscopy depicting biological samples [162–164]. Along with creating large databases, it is equally important to adopt standardized data acquisition protocols, acquire balanced datasets and reliable annotations to increase the current state-of-the-art performances of the models. In order to achieve robust and reliable deep learning models and to use them in clinical setting, it is required to apply online training, updating the model parameters with the arrival of new data and check the data and model reproducibility. At the same

time, the biophotonic community should adopt validating standards in order to avoid publishing over-fitted deep learning models. Despite of the outstanding progress of deep learning methods in biophotonics field, their reliability as decision-making systems is always questionable due to their “black-box” behavior. Thus, researchers are developing methods to understand the deep learning predictions [136, 165]. Nevertheless, this topic needs more investigations.

Finally, we have to answer our initial question “Is deep learning a boon for biophotonics?” We think that deep learning is eventually going to be a boon to biophotonics, which will revolutionize the decision-making approaches for pathologist, clinicians and doctors. A motivating example of deep learning used in optical systems is the IDx-DR device, a clinically accepted deep learning model to detect diabetes retinopathy in optical coherence tomography images [166]. Another potential example is GAN-based modeling for virtual staining of autofluorescence images which can bypass the long staining protocols and help the pathologist to compare new biophotonic technologies with the “gold-standard” staining methods. However, deep learning for biophotonics is still in an infant stage and requires overcoming various hurdles before coming into clinical usage. A large amount of data, quality check for the data, providing reliable annotations, appropriate model validation, interpreting model predictions and improving hardware capacities are vital for overcoming these hurdles. Overcoming these challenges and achieving optimal decision-making algorithms based on deep learning for modern healthcare systems is potentially the future of biophotonics.

ACKNOWLEDGMENTS

Funding of the German Research Foundation (DFG) for the projects BO 4700/1-1, BO 4700/4-1, PO 563/30-1 and STA 295/11-1 is highly acknowledged. This work received financial support by the Ministry for Economics, Sciences and Digital Society of Thuringia (TMWWDG), under the framework of the Landesprogramm ProDigital (DigLeben-5575/10-9).


CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

ORCID

Pranita Pradhan  <https://orcid.org/0000-0002-0558-2914>

Thomas W. Bocklitz  <https://orcid.org/0000-0003-2778-6624>

REFERENCES

- [1] C. Krafft, *J. Biophotonics* **2016**, 9(11-12), 1362.
- [2] N. Vogler, S. Heuke, T. W. Bocklitz, M. Schmitt, J. Popp, *Annu. Rev. Anal. Chem.* **2015**, 8, 359.
- [3] L. Marcu, S. A. Boppart, M. R. Hutchinson, J. Popp, B. C. Wilson, *J. Biomed. Opt.* **2017**, 23(2), 021103.
- [4] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford **1995**.
- [5] M. P. McBee, O. A. Awan, A. T. Colucci, C. W. Ghobadi, N. Kadom, A. P. Kansagra, S. Tridandapani, W. F. Auffermann, *Acad. Radiol.* **2018**, 25(11), 1472.
- [6] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, Cambridge, MA **2016**.
- [7] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge **1996**.
- [8] Stuart Russel, Peter Norvig, *EUA: Prentice Hall* **2003**, 178.
- [9] F. Rosenblatt, *Psychol. Rev.* **1958**, 65(6), 386.
- [10] B. R. Upadhyaya, G. Mathai, E. Eryurek, in *29th IEEE Conference on Decision and Control*, IEEE, **1990**, pp. 3277-3282.
- [11] T. Bocklitz, M. Putsche, C. Stüber, J. Käs, A. Niendorf, P. Rösch, J. Popp, *J. Raman Spectrosc.* **2009**, 40(12), 1759.
- [12] W. S. McCulloch, W. Pitts, *Bull. Math. Biophys.* **1943**, 5(4), 115.
- [13] D. H. Hubel, T. N. Wiesel, *J. Physiol.* **1963**, 165(3), 559.
- [14] S. Brahmam, L. C. Jain, L. Nanni, A. Lumini, *Local Binary Patterns: New Variants and Applications*, Springer, Berlin, Heidelberg **2013**.
- [15] Léon Bottou, *On-Line Learning in Neural Networks*, Cambridge University Press, New York **1998**, 17 (9), 142.
- [16] L. Bottou, in *Proceedings of COMPSTAT2010*, Springer, **2010**, pp. 177-186.
- [17] D. E. Rumelhart, G. E. Hinton, R. J. Williams, *Nature* **1986**, 323(6088), 533.
- [18] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, L. D. Jackel, in *Advances in Neural Information Processing Systems*, **1990**, pp. 396-404.
- [19] D. P. Kingma, J. Ba, *arXiv preprint arXiv:1412.6980* **2014**.
- [20] M. D. Zeiler, *arXiv preprint arXiv:1212.5701* **2012**.
- [21] J. Duchi, E. Hazan, Y. Singer, *J. Mach. Learn. Res.* **2011**, 12, 2121.
- [22] Y. Le Cun, L. D. Jackel, B. Boser, J. S. Denker, H. P. Graf, I. Guyon, D. Henderson, R. E. Howard, W. Hubbard, *IEEE Commun. Mag.* **1989**, 27(11), 41.
- [23] A. F. Agarap, *arXiv preprint arXiv:1803.08375* **2018**.
- [24] B. Karlik, A. V. Olgac, *Int. J. Artif. Intell. Expert Syst.* **2011**, 1 (4), 111.
- [25] C. Nwankpa, W. Ijomah, A. Gachagan, S. Marshall, *arXiv preprint arXiv:1811.03378* **2018**.
- [26] H. Lin, C. Wei, G. Wang, C. Hu, L. Lin, M. Ni, J. Chen, S. Zhuo, *J. Biophotonics* **2019**, 12(7), e201800435.
- [27] H. Fan, F. Zhang, X. Liang, Z. Li, G. Liu, Y. Xu, *J. Biophotonics* **2019**, 12(7), e201800488.
- [28] Z. C. Lipton, J. Berkowitz, C. Elkan, *arXiv preprint arXiv:1506.00019* **2015**.
- [29] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille, *arXiv preprint arXiv:1412.6632* **2014**.

- [30] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts, *EMNLP* **1631**, 2013, 1631.
- [31] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, **2014**, pp. 1724-1734. <https://www.aclweb.org/anthology/D14-1179>.
- [32] W. Zaremba, I. Sutskever, O. Vinyals, *arXiv preprint arXiv:1409.2329* **2014**.
- [33] P. Baldi, G. Pollastri, *J. Mach. Learn. Res.* **2003**, 4, 575.
- [34] X. Zhu, P. Sobihani, H. Guo, in *International Conference on Machine Learning*, **2015**, pp. 1604-1612.
- [35] D. H. Ballard, in *Proceedings of the Sixth National Conference on Artificial Intelligence – Volume 1*, AAAI Press, of AAAI'87, **1987**, pp. 279-284.
- [36] H. Bouldard, Y. Kamp, *Biol. Cybern.* **1988**, 59, 291.
- [37] P. Pradhan, T. Meyer, M. Vieth, A. Stallmach, M. Waldner, M. Schmitt, J. Popp, T. Bocklitz, in *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods – Volume 1: ICPRAM, INSTICC, SciTePress*, **2019**, pp. 396-405.
- [38] L. Duan, X. Qin, Y. He, X. Sang, J. Pan, T. Xu, J. Men, R. E. Tanzi, A. Li, Y. Ma, C. Zhou, *CoRR* **2018**, *abs/1803.01947*.
- [39] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, *Med. Image Anal.* **2017**, 42, 60.
- [40] K. Raza, N. K. Singh, *arXiv preprint arXiv:1812.07715* **2018**.
- [41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets. in *Advances in Neural Information Processing Systems 27* (Eds: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger), Curran Associates, Inc., New York **2014**, p. 2672.
- [42] Y. Rivenson, Z. Göröcs, H. Günaydin, Y. Zhang, H. Wang, A. Ozcan, *Optica* **2017**, 4(11), 1437.
- [43] H. Wang, Y. Rivenson, Y. Jin, Z. Wei, R. Gao, H. Günaydin, L. A. Bentolila, C. Kural, A. Ozcan, *Nat. Methods* **2019**, 16(1), 103.
- [44] Y. Rivenson, H. Wang, Z. Wei, K. Haan, Y. Zhang, Y. Wu, H. Gunaydin, J. Zuckerman, T. Chong, A. Sisk, L. Westbrook, W. Wallace, A. Ozcan, *Nat. Biomed. Eng.* **2019**, 3, 466.
- [45] C. Fu, S. Lee, D. J. Ho, S. Han, P. Salama, K. W. Dunn, E. J. Delp, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, **2018**, pp. 2221-2229.
- [46] A. Osokin, A. Chessel, R. E. Carazo Salas, F. Vaggi, in *Proceedings of the IEEE International Conference on Computer Vision*, **2017**, pp. 2233-2242.
- [47] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, in *Proceedings of the 22nd ACM International Conference on Multimedia*, **2014**, pp. 675-678.
- [48] R. Collobert, S. Bengio, J. Mariétoz, *Torch: A Modular Machine Learning Software Library*, Idiap, Martigny, Switzerland **2002**.
- [49] The Team, R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, A. Belopolsky, Y. Bengio, A. Bergeron, J. Bergstra, V. Bisson, J. B. Snyder, N. Bouchard, N. Boulanger-Lewandowski, X. Bouthillier, Y. Zhang, *arXiv preprint arXiv:1605.02688* **2016**.
- [50] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Retrieved from <https://www.tensorflow.org/>, **2015**.
- [51] F. Chollet, and others *Keras*, <https://keras.io>, **2015**.
- [52] S. Dieleman, J. Schlüter, C. Raffel, E. Olson, S. K. Sønderby, D. Nouri, D. Maturana, M. Thoma, E. Battenberg, J. Kelly, J. De Fauw, M. Heilman, D. M. de Almeida, B. McFee, H. Weideman, G. Takács, P. de Rivaz, J. Crall, G. Sanders, K. Rasul, C. Liu, G. French, J. Degraeve, *Lasagne: First Release*, **2015**. <https://doi.org/10.5281/zenodo.27878>.
- [53] M. Nielsen, Y. Bengio, A. Couville, Retrieved from <http://neuralnetworksanddeeplearning2017>.
- [54] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York **2006**.
- [55] L. Deng, D. Yu, *Foundations and Trends in Signal Process.* **2014**, 7(3-4), 197.
- [56] K. Gurney, *An Introduction to Neural Networks*, CRC Press, London, UK **2014**.
- [57] H. D. Beale, H. B. Demuth, M. T. Hagan, *PWS Publish Company* **1996**, 11, 1-47.
- [58] S. Haykin, *Neural Networks and Learning Machines*, 3/E, Pearson Education India, Hoboken, NJ **2010**.
- [59] K. J. Halupka, B. J. Antony, M. H. Lee, K. A. Lucy, R. S. Rai, H. Ishikawa, G. Wollstein, J. S. Schuman, R. Garnavi, *Biomed. Opt. Express* **2018**, 9(12), 6205.
- [60] Y. Ma, X. Chen, W. Zhu, X. Cheng, D. Xiang, F. Shi, *Biomed. Opt. Express* **2018**, 9(11), 5129.
- [61] Y. Huang, Z. Lu, Z. Shao, M. Ran, J. Zhou, L. Fang, Y. Zhang, *Opt. Express* **2019**, 27(9), 12289.
- [62] A. Janowczyk, A. Madabhushi, *J. Pathol. Inf.* **2016**, 7(1), 29.
- [63] X. Xu, J. Cheng, M. Thrall, Z. Liu, X. Wang, S. Wong, *Biomed. Opt. Express* **2013**, 4, 2855.
- [64] Y. Li, R. Zheng, Y. Wu, K. Chu, Q. Xu, M. Sun, Z. J. Smith, *J. Biophotonics* **2019**, 12(9), e201800410.
- [65] O. Ronneberger, P. Fischer, T. Brox, in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, Vol. 9351 of LNCS, **2015**, pp. 234-241 (available on [arXiv:1505.04597](https://arxiv.org/abs/1505.04597) [cs.CV]). <http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a>.
- [66] V. Badrinarayanan, A. Kendall, R. Cipolla, *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, 39, 2481.
- [67] F. Visin, K. Kastner, A. C. Courville, Y. Bengio, M. Matteucci, K. H. Cho, *CoRR* **2015**, *abs/1511.07053*.
- [68] J. Gabriel, J. S. Brostow, J. Fauqueur, R. Cipolla, in *European Conference on Computer Vision*, Springer, **2008**, pp. 44-57.
- [69] Y. Xie, Z. Zhang, M. Sapkota, Y. Lin, *Medical Image Computing and Computer-assisted Intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* **2016**, 9901, 185-193.
- [70] G. Jun, F. C. Yaw, B. K. Ng, S. G. Razul, S. K. Lim, *J. Biophotonics* **2014**, 7(7), 483.

- [71] E. Ryzhikova, O. Kazakov, L. Halamkova, D. Celmins, P. Malone, E. Molho, E. A. Zimmerman, I. K. Lednev, *J. Biophotonics* **2015**, *8*(7), 584.
- [72] G. I. Gopakumar, M. Swetha, S. S. Gorthi, G. S. Subrahmanyam, *J. Biophotonics* **2017**, *11*, e201700003.
- [73] K. Aljakouch, Z. Hilal, I. Daho, M. Schuler, S. D. Krauß, H. K. Yosef, J. Dierks, A. Mosig, K. Gerwert, S. F. El-Mashtoly, *Anal. Chem.* **2019**, *91*(21), 13900.
- [74] S. Weng, X. Xu, J. Li, S. T. C. Wong, *J. Biomed. Opt.* **2017**, *22* (10), 106017.
- [75] N. Singla, K. Dubey, V. Srivastava, *J. Biophotonics* **2019**, *12* (3), e201800255.
- [76] N. Ali, E. Quansah, K. Köhler, T. Meyer, M. Schmitt, J. Popp, A. Niendorf, T. Bocklitz, *Transl. Biophotonics* **2019**, *1*(1-2), e201900003.
- [77] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, **2015**, pp. 1-9.
- [78] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, **2016**, pp. 2818-2826.
- [79] K. Simonyan, A. Zisserman, *arXiv 1409.1556* **2014**.
- [80] A. E. Heidari, T. T. Pham, I. Ifegwu, R. Burwell, W. B. Armstrong, T. Tjosen, S. Whyte, C. Giorgioni, B. Wang, B. J. F. Wong, Chen, Z., *J. Biophotonics* **2019**, *13* (3), e201900221.
- [81] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255.
- [82] K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, **2016**, pp. 770-778.
- [83] X. Feng, K. Qing, N. J. Tustison, C. H. Meyer, Q. Chen, *Med. Phys.* **2019**, *46*(5), 2169.
- [84] Y. Liu, W. Fu, V. Selvakumaran, M. Phelan, W. Paul Segars, E. Samei, M. Mazurowski, J. Y. Lo, G. D. Rubin, R. Henao, in *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications* (Eds: P.-H. Chen, P. R. Bak), International Society for Optics and Photonics, SPIE **2019**, p. 319. <https://doi.org/10.1117/12.2512887>.
- [85] N. Bayramoglu, M. Kaakinen, L. Eklund, J. Heikkila, in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, **2017**, pp. 64-71.
- [86] Z. Xu, C. F. Moro, B. Bozóky, Q. Zhang, *arXiv preprint arXiv:1901.04059* **2019**.
- [87] T. Gregory Schaaff, J. M. McMahon, P. J. Todd, *Anal. Chem.* **2002**, *74*(17), 4361.
- [88] T. Bocklitz, A. Crecelius-Vitz, C. Matthäus, N. Tarcea, F. von Eggeling, M. Schmitt, U. Schubert, J. Popp, *Anal. Chem.* **2013**, *85*(22), 10829.
- [89] J. T. Kwak, S. Hewitt, S. Sinha, R. Bhargava, *BMC Cancer* **2011**, *11*, 62.
- [90] C. Yang, D. Niedieker, F. Grosserueschkamp, M. Horn, A. Tannapfel, A. Kallenbach-Thieltges, K. Gerwert, A. Mosig, *BMC Bioinformatics* **2015**, *16*(1), 396.
- [91] H. Uzunova, M. Wilms, H. Handels, J. Ehrhardt, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, **2017**, pp. 223-231.
- [92] H. Li, Y. Fan, *arXiv preprint arXiv:1709.00799* **2017**.
- [93] E. Nehme, L. E. Weiss, T. Michaeli, Y. Shechtman, *Optica* **2018**, *5*(4), 458.
- [94] Z. Luo, A. Yurt, R. Stahl, A. Lambrechts, V. Reumers, D. Braeken, L. Lagae, *Opt. Express* **2019**, *27*(10), 13581.
- [95] O. Ryabchykov, S. Guo, T. Bocklitz, *Phys. Sci. Rev.* **2018**, *4*(2), 20170043.
- [96] S. Seifert, V. Merk, J. Kneipp, *J. Biophotonics* **2016**, *9* (1-2), 181.
- [97] V. Tafintseva, E. Vigneau, V. Shapaval, V. Cariou, E. M. Qannari, A. Kohler, *J. Biophotonics* **2018**, *11*(3), e201700047.
- [98] H. Chen, X. Li, N. Broderick, Y. Liu, Y. Zhou, J. Han, W. Xu, *J. Biophotonics* **2018**, *11*(9), e201800016.
- [99] J. Gerretzen, E. Szymańska, J. Jansen, J. Bart, H.-J. Manen, E. Van den Heuvel, L. Buydens, *Anal. Chem.* **2015**, *87*(24), 12096.
- [100] J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon, S. J. Gibson, *Analyst* **2017**, *142*(21), 4067.
- [101] E. J. Bjerrum, M. Glahder, T. Skov, *arXiv preprint arXiv:1710.01927* **2017**.
- [102] A. Raulf, J. Butke, C. Küpper, F. Großertüschkamp, K. Gerwert, A. Mosig, *Bioinformatics (Oxford, England)* **2020**, *36*(1), 287.
- [103] J. Acquarelli, T. van Laarhoven, J. Gerretzen, T. Tran, L. Buydens, E. Marchiori, *Anal. Chim. Acta* **2016**, *954*, 22.
- [104] J. Dong, M. Hong, Y. Xu, X. Zheng, *J. Chemometr.* **2019**, *33* (11), e3184.
- [105] H. Lin, F. Deng, K.-C. Huang, H. J. Lee, J.-X. Cheng, *CLEO: Applications and Technology*, Optical Society of America, San Jose, CA **2019**, ATu3K.
- [106] H. He, M. Xu, Z. Cheng, P. Zheng, L. Luo, L. Wang, B. Ren, *Anal. Chem.* **2019**, *91*, 7070.
- [107] Y. Bengio, A. Courville, P. Vincent, *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*(8), 1798.
- [108] T. Trakoolwilaiwan, B. Behboodi, J. Lee, K. Kim, J.-W. Choi, *Neurophotonics* **2017**, *5*, 1.
- [109] S. Hiwa, K. Hanawa, R. Tamura, K. Hachisuka, T. Hiroyasu, *Comput. Intell. Neurosci.* **2016**, *2016*, 1.
- [110] M. Yu, H. Yan, J. Xia, L. Zhu, T. Zhang, Z. Zhu, X. Lou, G. Sun, M. Dong, *Photodiagn. Photodyn. Ther.* **2019**, *26*, 430.
- [111] W. B. Sohn, S. Y. Lee, S. Kim, *J. Raman Spectrosc.* **2020**, *51* (3), 414.
- [112] S. Krauß, R. Roy, H. Yosef, T. Lechtonen, S. El-Mashtoly, K. Gerwert, A. Mosig, *J. Biophotonics* **2018**, *11*(10), e201800022.
- [113] C.-S. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. A. E. Saleh, S. Ermon, J. Dionne, *Nat. Commun.* **2019**, *10*(1), 1.
- [114] Salim Malek, Farid Melgani, Yakoub Bazi, *J. Chemometr.* **2017**.
- [115] X. Yu, L. Tang, X. Wu, H. Lu, *Food Anal. Methods* **2017**, *11*, 1.
- [116] Z. Jianqiang, L. Weijuan, H. Ying, Q. Changgui, Y. Shuangyan, L. Changyu, N. Linru, *Anal. Lett.* **2018**, *51*(7), 1029.
- [117] B. Le, X. Dong, Y. Mao, D. He, *Infrared Phys. Technol.* **2018**, *93*, 34.
- [118] X. Fan, W. Ming, H. Zeng, Z. Zhang, H. Lu, *Analyst* **2019**, *144*, 1789.

- [119] K. Liland, A. Kohler, N. Afseth, *J. Raman Spectrosc.* **2016**, 47, 643.
- [120] D. Soekhoe, P. Van Der Putten, A. Plaat, in *International Symposium on Intelligent Data Analysis*, Springer, **2016**, pp. 50-60.
- [121] J. Cho, K. Lee, E. Shin, G. Choy, S. Do, *arXiv preprint arXiv:1511.06348* **2015**.
- [122] C. Shorten, T. M. Khoshgoftaar, *J. Big Data* **2019**, 6(1), 60.
- [123] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, A. Madry, *arXiv preprint arXiv:1712.02779* **2017**.
- [124] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, in *Advances in Neural Information Processing Systems*, **2014**, pp. 3320-3328.
- [125] M. Buda, A. Maki, M. A. Mazurowski, *Neural Netw.* **2018**, 106, 249.
- [126] J. M. Johnson, T. M. Khoshgoftaar, *J. Big Data* **2019**, 6(1), 27.
- [127] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, P. J. Kennedy, in *2016 international Joint Conference on Neural Networks (IJCNN)*, IEEE, 2016, pp. 4368-4374.
- [128] B. Neal, S. Mittal, A. Baratin, V. Tantia, M. Scicluna, S. Lacoste-Julien, I. Mitliagkas, *arXiv preprint arXiv:1810.08591* **2018**.
- [129] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, J. Xu, E. M. Cofer, C. A. Lavender, S. C. Turaga, A. M. Alexandari, Z. Lu, D. J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. S. Segler, S. M. Boca, S. Joshua Swamidass, A. Huang, A. Gitter, C. S. Greene, *J. R. Soc. Interface* **2018**, 15(141), 20170387.
- [130] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, *arXiv preprint arXiv:1611.03530* **2016**.
- [131] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, *PLoS One* **2015**, 10(7), e0130140.
- [132] G. Montavon, W. Samek, K.-R. Müller, *Dig. Signal Process.* **2018**, 73, 1.
- [133] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, *Pattern Recogn.* **2017**, 65, 211.
- [134] W. Samek, T. Wiegand, K.-R. Müller, *CoRR* **2017**, abs/1708.08296.
- [135] K. Simonyan, A. Vedaldi, A. Zisserman, *preprint* **2013**.
- [136] T. Bocklitz, in *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods – Volume 1 ICPRAM, INSTICC, SciTePress*, 2019, pp. 874-880.
- [137] R. Horstmeyer, R. Heintzmann, G. Popescu, L. Waller, C. Yang, *Nat. Photonics* **2016**, 10(2), 68.
- [138] M. Koch, P. Symvoulidis, V. Ntziachristos, *Nat. Photonics* **2018**, 12(9), 505.
- [139] *Nat. Photonics* **2018**, 12, 117.
- [140] Z. Hussain, F. Gimenez, D. Yi, D. Rubin, *AMIA Annual Symposium Proceedings* **2017**, 979-984.
- [141] A. Mikołajczyk, M. Grochowski, *Int. Interdisc. PhD Workshop (IIPhDW)* **2018**, 2018, 117.
- [142] C. M. Bishop, *Neural Comput.* **1995**, 7, 108.
- [143] L. Perez, J. Wang, *CoRR* **2017**, abs/1712.04621.
- [144] S. J. Pan, Q. Yang, *IEEE Trans. Knowl. Data Eng.* **2009**, 22(10), 1345.
- [145] S. Guo, T. Bocklitz, U. Neugebauer, J. Popp, *Anal. Methods* **2017**, 9, 4410.
- [146] L. Prechelt, *Neural Netw.* **1998**, 11(4), 761.
- [147] L. Prechelt, in *Early Stopping—But When?* (Eds: G. Montavon, G. B. Orr, K.-R. Müller), Springer, Berlin Heidelberg, Berlin, Heidelberg **2012**, p. 53.
- [148] J. Kukačka, V. Golkov, D. Cremers, *arXiv preprint arXiv:1710.10686* **2017**.
- [149] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, *J. Mach. Learn. Res.* **2014**, 15(1), 1929.
- [150] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, *arXiv preprint arXiv:1207.0580* **2012**.
- [151] J. Ba, B. Frey, in *Advances in Neural Information Processing Systems*, **2013**, pp. 3084-3092.
- [152] X. Bouthillier, K. Konda, P. Vincent, R. Memisevic, *arXiv preprint arXiv:1506.08700* **2015**.
- [153] Shin-ichi Maeda, *arXiv preprint arXiv:1412.7003* **2014**.
- [154] Y. Gal, Z. Ghahramani, in *International Conference on Machine Learning*, **2016**, pp. 1050-1059.
- [155] D. C. Plaut, S. J. Nowlan, G. E. Hinton, *Experiments on Learning by Back Propagation*. ERIC, Washington, DC **1986**.
- [156] A. Krogh, J. A. Hertz, in *Advances in Neural Information Processing Systems 4* (Eds: J. E. Moody, S. J. Hanson, R. P. Lippmann), Morgan-Kaufmann, Santa Cruz, CA **1992**, p. 950.
- [157] R. Reed, R. J. MarksII, *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*, MIT Press, Cambridge, MA **1999**.
- [158] Y. LeCun, L. Bottou, G. B. Orr, K. R. Müller, in *Efficient BackProp* (Eds: G. B. Orr, K.-R. Müller), Springer, Berlin Heidelberg, Berlin, Heidelberg **1998**, p. 9.
- [159] S. Ioffe, C. Szegedy, *arXiv preprint arXiv:1502.03167* **2015**.
- [160] C. Brian, M. J. Wilson, F. Leblond, *J. Biomed. Opt.* **2018**, 23(3), 030901.
- [161] S. Guo, R. Heinke, S. Stöckel, P. Rösch, J. Popp, T. Bocklitz, *J. Raman Spectrosc.* **2018**, 49(4), 627.
- [162] M. Riffle, T. N. Davis, *BMC Bioinformatics* **2010**, 11(1), 263.
- [163] Y. Zhang, Y. Zhu, E. Nichols, Q. Wang, S. Zhang, C. Smith, S. Howard, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, **2019**, pp. 11710.
- [164] E. Czech, B. A. Aksoy, P. Aksoy, J. Hammerbacher, *BMC Bioinformatics* **2019**, 20(1), 1.
- [165] J. de La Torre, A. Valls, D. Puig, *Neurocomputing* **2019**.
- [166] M. D. Abramoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, M. Niemeijer, *Investig. Ophthalmol. Vis. Sci.* **2016**, 57(13), 5200.

How to cite this article: Pradhan P, Guo S, Ryabchykov O, Popp J, Bocklitz TW. Deep learning a boon for biophotonics? *J. Biophotonics*. 2020; e201960186. <https://doi.org/10.1002/jbio.201960186>

Nonlinear Multimodal Imaging Characteristics of Early Septic Liver Injury in a Mouse Model of Peritonitis

Melina Yarbakht,^{†,‡,||,#} Pranita Pradhan,^{‡,||,#} Nilay Köse-Vogel,[†] Hyeonsoo Bae,^{||} Sven Stengel,[†] Tobias Meyer,^{||,‡} Michael Schmitt,^{‡,||} Andreas Stallmach,[†] Jürgen Popp,^{‡,||} Thomas Wilhelm Bocklitz,^{*,‡,||} and Tony Bruns^{*,†,‡,‡}

[†]Department of Internal Medicine IV (Gastroenterology, Hepatology, Infectious Disease) and [‡]Center for Sepsis Control and Care (CSCC), Jena University Hospital, 07747 Jena, Germany

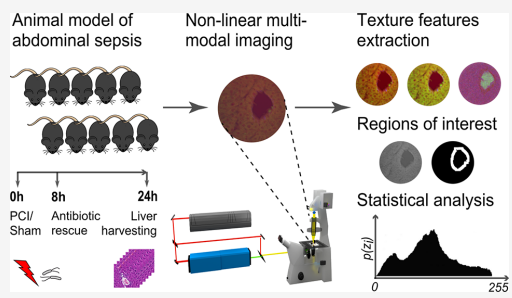
[‡]Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller University, 07743 Jena, Germany

^{||}Leibniz Institute of Photonic Technology, Member of Leibniz Health Technology, 07745 Jena, Germany

^{*}Department of Medicine III, University Hospital RWTH Aachen, 52074 Aachen, Germany

Supporting Information

ABSTRACT: Sepsis constitutes a life-threatening organ failure caused by a deregulated host response to infection. Identifying early biomolecular indicators of organ dysfunction may improve clinical decision-making and outcome of patients. Herein we utilized label-free nonlinear multimodal imaging, combining coherent anti-Stokes Raman scattering (CARS), two-photon excited autofluorescence (TPEF), and second-harmonic generation (SHG) to investigate the consequences of early septic liver injury in a murine model of polymicrobial abdominal infection. Liver tissue sections from mice with and without abdominal sepsis were analyzed using multimodal nonlinear microscopy, immunofluorescence, immunohistochemistry, and quantitative reverse transcription polymerase chain reaction (qRT-PCR). Twenty-four hours after the induction of sepsis, hepatic mRNA of inflammatory cytokines and acute phase proteins was upregulated, and liver-infiltrating myeloid cells could be visualized alongside hepatocellular cytoplasmic translocation of high mobility group box 1. According to the statistical analysis based on texture feature extraction followed by the combination of dimension reduction and linear discriminant analysis, CARS (AUC = 0.93) and TPEF (AUC = 0.83) showed an excellent discrimination between liver sections from septic mice and sham-treated mice in contrast to SHG (AUC = 0.49). Spatial analysis revealed no major differences in the distribution of sepsis-associated changes between periportal and pericentral zones. These data suggest early alterations in hepatic lipid distribution and metabolism during liver injury and confirm nonlinear multimodal imaging as a promising complementary method for the real-time, label-free study of septic liver damage.



Sepsis is the primary cause of death from infection and the leading cause of death in intensive care units worldwide. On the basis of the third international definition of sepsis, it is currently defined as a life-threatening organ dysfunction which results from dysregulated host responses to infection.¹ In addition to being a critical regulator of the inflammatory host response during systemic infections,² the liver constitutes an important target organ during sepsis, manifesting as hepatocellular injury and cholestasis.^{3,4} In addition, patients with pre-existing liver disease are predisposed to developing systemic inflammation and organ failure in response to bacterial infections.⁵ As early recognition and timely treatment determine the outcome of sepsis,⁶ patients at risk should be identified before organ dysfunction has been established.⁷ Therefore, novel markers of imminent organ damage need to

be identified in order to improve the outcome of septic liver injury.

The nondestructive characterization of intact biological samples in their native condition using multiphoton optical microscopy methods introduces optical imaging techniques as a promising diagnostic method for future medical applications.⁸ Classical imaging techniques, such as positron emission tomography (PET), computed tomography (CT), and magnetic resonance imaging (MRI), provide large penetration depth but only limited spatial resolution.⁹ Linear optical microscopy methods including linear Raman, fluorescence microscopy, or optical coherence tomography often lack

Received: April 10, 2019

Accepted: July 30, 2019

Published: July 30, 2019

sufficient imaging speed, tissue penetration, or contrast.¹⁰ These limitations can be overcome by employing nonlinear optical imaging techniques, which allow high-speed imaging with dwell times of microseconds. For example, three-photon fluorescence using 1700 nm excitation penetrates up to 1.2 mm into tissue.¹¹ Multimodal imaging provides information about different constituents of biological samples, such as the distribution of lipids with coherent anti-Stokes Raman scattering (CARS) at the symmetric aliphatic CH₂-stretching vibration of methylene groups CH₂ at 2850 cm⁻¹, the analysis of autofluorophores, such as fiber elastin, nicotinamide adenine dinucleotide (phosphate), and flavin adenine dinucleotide, with two-photon excited fluorescence (TPEF), and collagen distribution with second harmonic generation (SHG). The diagnostic potential of CARS/TPEF/SHG imaging has been studied using various chemometric techniques in different diseases, such as brain tumors,¹² inflammatory bowel disease,^{13,14} nonmelanoma skin cancer,¹⁵ and head and neck carcinoma.⁹

This research undertook to evaluate the efficiency and utility of multimodal imaging as a label-free complementary method in combination with image analysis in early septic liver injury in a mouse model of polymicrobial abdominal sepsis.^{16,17} To that aim, unstained liver tissue sections were analyzed using nonlinear multimodal imaging alongside classical techniques in order to prove the concept that this method can visualize early biochemical changes in septic liver injury.

MATERIALS AND METHODS

Animals and Experimental Procedure. All animal experiments were approved by the local government authority of Thuringia, Thüringer Landesverwaltungsamt, and performed based on the approved guidelines (reg. no. 02-010/15). Inbred male C57BL/6J mice at 8–12 weeks of age were provided by the Institute of Laboratory Animal Science and Welfare of Jena University Hospital and were housed under controlled light/dark cycles plus ad libitum access to food and water. We used the polymicrobial contamination and infection (PCI) model of sepsis based on an intraperitoneal (ip) injection of a defined volume of human stool.^{16,17} Fecal suspension (3.5 μL/g body weight) or 0.9% NaCl as a control was injected intraperitoneally in five animals per group. After 8 h, animals received 25 μg/g meropenem as antibiotic rescue. Twenty-four hours after the septic injury, animals were sacrificed by an isoflurane overdose. Spleen and livers were harvested, fixed in 10% natural buffered formalin (Sigma-Aldrich, Taufkirchen, Germany) or shock-frozen in liquid nitrogen for further studies. To obtain plasma, cardiac EDTA blood samples were taken and centrifuged at 2000g for 15 min at 4 °C. Organs and plasma were stored at -80 °C until further analysis (Figure 1A).

Hepatic Gene Expression Analysis. Two-step qRT-PCR¹⁸ was used to evaluate the expression level of *Il6*, *Il1b*, and *Saa3* in liver tissue. Total RNA was isolated from liver samples with the use of TriFast solution (VWR/PEQLAB, Darmstadt, Germany) according to the manufacturer's instructions and purified with the use of NucleoSpin RNA kit columns (Macherey-Nagel, Düren, Germany). Conversion of purified RNA to cDNA was performed with the use of High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Darmstadt, Germany) based on the manufacturer's standard protocol. qRT-PCR was performed with the gene specific primers *Il6* (forward: 5'-GACAAAGCCAGA-

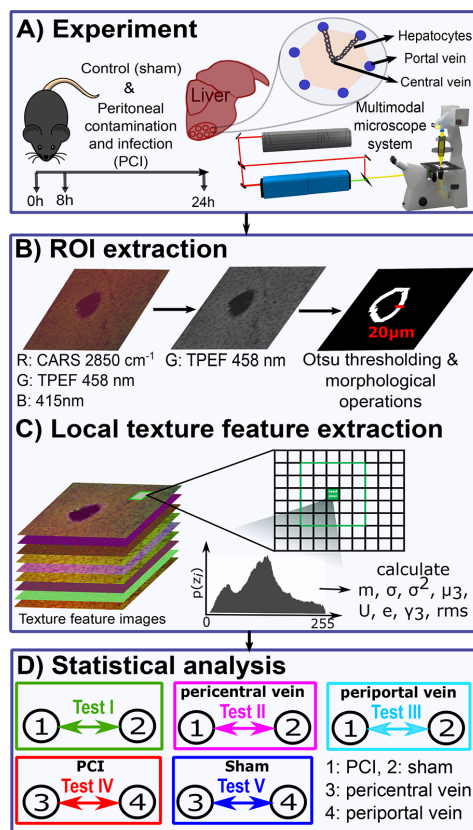


Figure 1. Schematic representation of the analysis workflow. (A) Mice underwent peritoneal contamination and infection (PCI) or sham experiments, and livers were explanted 24 h after the onset of sepsis for nonlinear multimodal imaging. (B) 20 μm regions around the pericentral and periportal veins were identified, and (C) local texture features were extracted. (D) Five statistical tests were conducted to identify spatial differences in liver texture transformation between animals with and without early septic liver damage.

GTCCCTCAGAGAG- 3' reverse: 5'-CTAGGTTTGCCG-AGTAGATCTC-3'), *Il1b* (forward: 5'-AAGGAGAACCAGCAACGACAAAA-3', reverse: 5'-TGGGGAAGCTCTGCAGACTCAAAC- 3'), and *Saa3* (forward: 5'-TGCCATCATTCTTGCATCTTGA- 3', reverse: 5'-CCGTGAAGCTTCTGAACAGCCT- 3') using Thermo Scientific Maxima SYBR Green qPCR Master Mix (Thermo Fisher Scientific, Dreieich, Germany) on a Rotor-Gene Q thermal cycler (QIAGEN, Hamburg, Germany). Normalization of threshold Cycles for the selected genes was performed against *Actb* (forward: 5'-ATGGAGGGGAATACAGCCC- 3', reverse: 5'-TTCTTTGCAGCTCCTTCGTT- 3') with the use of Rotor Gene Q-series software. The 2^{-ΔΔCt} method was used to calculate the relative fold change.¹⁹

Plasma Markers of Hepatic Damage. Hepatic cell death was investigated using plasma concentrations of alanine aminotransferase (ALT) on an Architect plus 16200 (Abbott, Wiesbaden, Germany).

Immunofluorescence and Immunohistochemistry.

Representative tissue liver was fixed in 10% natural buffered formalin (Sigma-Aldrich, Taufkirchen, Germany) for 24 h at room temperature, embedded in paraffin, cut into 4 μm thick sections, and mounted on Superfrost/Plus-coated glass slides (Fisher Scientific, Pittsburgh, PA) for hematoxylin and eosin (H&E).

For immunohistochemistry analysis of high mobility group box 1 protein (HMGB1), 4 μm sections were incubated at 59 $^{\circ}\text{C}$ for 45 min. Following the antigen retrieval step (Dako, Fisher Scientific), blocking was performed in peroxidase block buffer in the humid chamber for 5 min. Slides were stained using rabbit polyclonal antibody (Abcam ab18256) at 1:1000 dilutions and HRP-conjugated anti-rabbit IgG (Dako Envision +System-HRP kit).

For immunofluorescence analysis, liver cryosections were fixed in ice-cold acetone for 10 min. After washing with Tris buffer 0.1% Tween-20 (TBS-T), incubation with primary antibodies against F4/80 (clone Cl:A3-1, BioRad, Munich, Germany) or Gr-1 (clone RB6-8C5, Biolegend, San Diego, CA) was performed overnight at 1:1000 dilutions and followed by incubation with Cy3-labeled donkey anti-rat IgG (1:400 Jackson Immuno Research Europe) for 1 h. DAPI Fluoromount-G (SouthernBiotech) served as a fluorescent stain for cell nuclei.

Multimodal Imaging. A continuous wave frequency-doubled Neodymium-Vanadate laser at 532 nm (Verdi V18, Coherent, Santa Clara, CA) with an average power of 16.7 W was used to pump a titanium-sapphire (Ti:Sa) laser (Mira HP, Coherent) to generate pulses at 832 nm with 2–3 ps pulse duration, 76 MHz repetition rate, and an average output power of 3.5 W. The Ti:Sa output was split into two parts: the first used as Stokes beam and the second coupled into an Optical Parametric Oscillator (OPO, APE, Germany) for frequency conversion. The OPO was tuned to 672.5 nm for generating the pump beam to excite the symmetric aliphatic C–H stretching vibration of methylene groups CH_2 at 2850 cm^{-1} . Both pump and Stokes laser beams were guided into an LSM 510 Meta, Zeiss, Germany), optimizing their temporal and spatial overlap by a mechanical delay line and a dichroic beam combiner and focused on the sample with an average power of 43 mW and 31 mW for the pump and Stokes, respectively, by a 20 \times objective of 0.8 NA (Plan-Apochromat 20 \times /0.8, Zeiss, Germany). The nonlinear modalities CARS, TPEF, and SHG were simultaneously acquired by non-descanned photomultiplier tubes (PMT, Hamamatsu Photonics, Japan) in forward (CARS, SHG) and backward direction (TPEF) with a pixel dwell time of 1.6 μs and a resolution of 2048 \times 2048 pixel using specific band-pass filters at 564 nm (CARS 2850 cm^{-1}), 426–490 nm (TPEF), and 415 nm (SHG), respectively.

Image Preprocessing and Feature Extraction of Multimodal Images. The image preprocessing and statistical analysis were performed using Python with packages such as sklearn,²⁰ scipy,²¹ and numpy²² on a commercially available PC system Intel CoreTM i5-7500 CPU, 3.40 GHz, 16 GB RAM.

The data set composed of six multimodal images obtained from each of the five PCI (sepsis) and five sham (control) mice. Here six multimodal images represent analytical replicates of each mouse, resulting in 60 multimodal images in total. Each multimodal image was a combination of three modalities including CARS as the red channel, TPEF as the

green channel, and SHG as the blue channel (Supporting Information Figures 1 and 2). The spatial resolution of each multimodal image was 2048 \times 2048 pixels for a 450 \times 450 μm^2 tile scan. Multimodal images containing the identified periportal or pericentral vein were downsampled to 1024 \times 1024 pixels. To assess a biomolecular difference in periportal and pericentral zones, a region of 20 μm around these veins was used for statistical analysis. The number of pixels in this 20 μm region for every multimodal image is different due to the varying sizes of vein cross-sections (Figure 1B). The perivenous regions were obtained using a binary mask based on the green channel (TPEF) of the multimodal image as TPEF channel showed the best contrast between the vein lumen (background) and hepatic tissue (foreground). SHG and CARS signals were less contrasting due to sparse perivenous signals and nonresonant background caused by underlying CaF_2 slides, respectively (Supporting Information Figure S1). Thus, the green channel was binarized using global image thresholding²³ based on Otsu's method,²⁴ followed by morphological operations including region opening, region closing, erosion, and dilation (Figure 1B).

Finally, for this 20 μm region a set of nine first-order histogram features including mean, median, standard deviation, variance, skewness, energy, entropy, kurtosis, and RMS were calculated locally for all three modalities: CARS/TPEF/SHG (Supporting Information Table S1). First-order histogram features are the statistical moments of the histogram which give an intuitive understanding of the texture of the underlying tissue region (Figure 1C). Detailed information on the local extraction of the first-order histogram features and the nine feature images can be found in Supporting Information (Table S1). We refer to the first-order histogram features as “texture features”. To summarize the texture feature information on the selected region, a median of the texture features images in the selected region was calculated. The median values of the texture feature images are stable to the varying number of pixels of the analyzed region caused due to different vein sizes. Thus, 27 median values of the texture features for the region in each multimodal image were obtained (9 median values acquired from 9 texture feature images of each channel). Further, these median values of the texture feature images were normalized to the range [0, 1] and were used for the statistical analysis.

Statistical Analysis and Modeling. Our primary aim was to investigate differences in multimodal imaging between liver sections from mice with abdominal sepsis (PCI) and controls without sepsis (sham). For this purpose, the median of the median values of each texture feature image obtained from the analytical replicates (six multimodal images) of each mouse was calculated. These continuous variables were compared with the nonparametric Mann–Whitney U test (test I in Figure 1D). In additional analyses, we investigated differences between the median values of texture feature images of mouse livers from the PCI and sham group for pericentral (test II) and periportal areas (test III), separately, as well as differences in periportal versus pericentral areas in animal with (test IV) and without (test V) sepsis. *P* values <0.05 in two-sided testing were considered significant. Owing to the exploratory nature of this study, we did not adjust for multiple testing.

To prove the efficiency of nonlinear multimodal imaging for early sepsis diagnosis, we trained a linear classifier to distinguish the PCI and sham group based on the median values of the texture feature images. The dimension of the

texture feature was reduced by principal component analysis, and the reduced features were used to train the linear classifier based on linear discriminant analysis (PCA-LDA). To further investigate the contribution of individual modalities for the classification of the two groups, we trained a linear classifier using median values of texture feature images obtained from CARS, TPEF, and SHG modalities separately. The linear classifiers were evaluated using leave-one-mouse-out cross-validation strategy, and the number of principal components is optimized.²⁵ A grid search was performed to acquire the best number of principal components. The binary classification results were visualized by plotting sensitivity against 1-specificity in a receiver-operating characteristic (ROC) curve. The area under the ROC curve (AUC) shows the ability of the binary classifier to discriminate between the PCI and sham group based on the features from all modalities combined and individual modalities.

RESULTS AND DISCUSSION

Confirmation of Septic Liver Injury. Plasma markers of liver damage showed mild hepatocellular cell death 24 h after the induction of sepsis (Figure 2A). The hepatic inflammatory

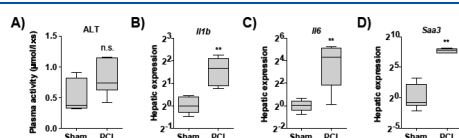


Figure 2. (A) Plasma activity of alanine aminotransferase (ALT) 24 h after the induction of sepsis in the peritoneal contamination and infection (PCI) model or after sham treatment. (B–D) Quantification of hepatic mRNA expression of *Il1b*, *Il6*, and *Saa3*. Expression data were normalized to *Actb* expression and the mean of sham. ** $P \leq 0.01$ in Mann–Whitney U test. n.s., not significant.

response could be confirmed by elevated hepatic mRNA levels of *Il6*, *Il1b*, and *Saa3* (Figure 2B–D). As suggested by plasma transaminase levels, H&E staining revealed only mild hepatocellular damage in conventional histology (Figure 3A). However, immunohistochemistry staining for HMGB1 revealed its cytoplasmic translocation in liver sections of septic mice but not in sham animals (Figure 3B) as observed in models of acute liver injury and in patients with acute liver failure.²⁶

To assess the hepatic immune cell infiltration, liver sections were stained for F4/80-expressing macrophages and Gr-1 (Ly-6G/Ly-6C)-expressing myeloid cells using immunofluorescence. There was no difference in F4/80-expressing cells in animals with sepsis (median 19.7 positive cells per 100 nuclei; range: 17.9–20.4) and sham animals without sepsis (20.4 positive cells per 100 nuclei, range: 14.2–25.2) (Figure 3C).

In contrast, septic liver damage was associated with an influx of Gr-1-expressing immune cells (median 11.5 positive cells per 100 nuclei, range: 7.9–22.5) as compared to sham animals (4.0 per 100 nuclei, range: 2.9 to 5.7) (Figure 3D) consistent with a hepatic accumulation of neutrophils during the early phase of abdominal sepsis.²⁷

Multimodal Imaging. Although the differences between multimodal images of liver sections from animals in the absence (Supporting Information Figure S1) or presence of septic liver injury (Supporting Information Figure S2) were not strikingly obvious at visual inspection, 7 out of 27 liver

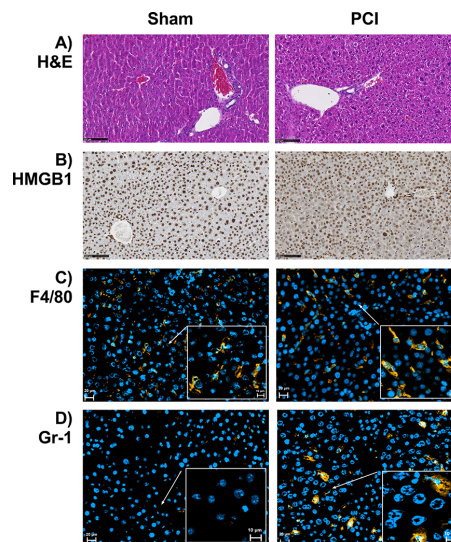


Figure 3. (A) Representative images of hematoxylin and eosin (H&E)-stained liver sections 24 h after peritoneal contamination and infection (PCI) or sham. Original magnification: 20 \times . Scale bar: 100 μ m. (B) Representative images of immunohistochemical staining of HMGB1 in livers from animals after PCI or sham. Immunohistochemistry was performed using an anti-HMGB1 antibody visualized with diaminobenzidine (brown). Original magnification: 10 \times . Scale bar: 10 μ m. Immunofluorescence staining of (C) F4/80-positive hepatic macrophages and (D) Gr-1-positive myeloid cells. Cells stained with primary antimouse monoclonal antibodies, a secondary Cy3-labeled antibody (yellow), and DAPI nuclear stain (blue). Original magnification: 20 \times (full) and 40 \times (insert). Scale bar: 20 μ m (full) and 10 μ m (insert).

texture features tested were significantly different between both groups of animals at the predefined significance level (test I). These texture features included data derived from CARS (mean, skewness, energy, entropy, and root-mean-square) and from SHG (standard deviation and variance) (Figure 4). When the analysis was restricted to the pericentral or periportal

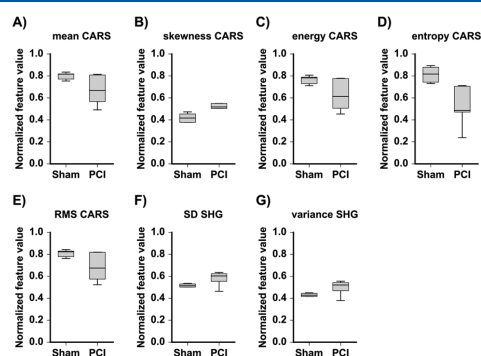


Figure 4. Results of test I: Boxplot of seven texture features indicating significant differences between livers from mice after peritoneal contamination and infection (PCI) and control animals (sham). RMS: root-mean-square. SD: standard deviation.

regions (tests II and III), the entropy of the CARS signal remained a robust discriminator between livers from septic mice and controls in both zones of the liver acinus (Supporting Information Figures S3 and S4). In contrast, we only observed a difference in the skewness of the TPEF signal when comparing the pericentral and the periportal regions of livers from animals with sepsis (test IV, Supporting Information Figure S5) and no significant differences between the two regions in nonseptic animals (test V, data not shown).

The result of multivariate analysis to distinguish between the two groups based on the texture features obtained from individual modalities was compared using receiver operating characteristics (Figure 5). The ROC of the classifier trained on

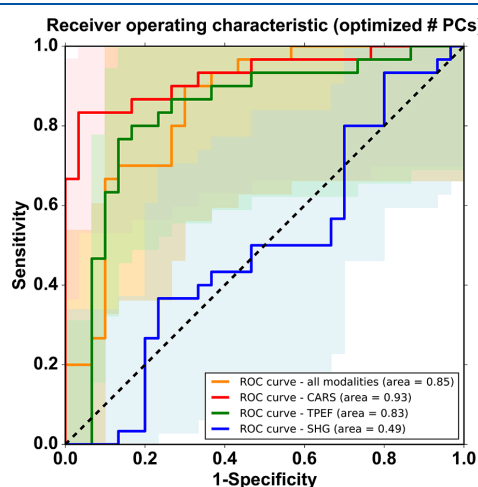


Figure 5. ROC curve analysis of the classifier trained on features extracted from all modalities (yellow), CARS-based features (red), TPEF-based features (green), and SHG-based features (blue). The classifier trained on CARS-based features achieved the highest AUC (AUC = 0.93). With TPEF-based features the classifier achieves an AUC of 0.83. Whereas, the classifier trained using SHG-based features shows poor diagnostic ability (AUC = 0.49) to discriminate between the PCI and sham group.

CARS and TPEF features show the highest AUC of 0.93 and 0.83, respectively. The CARS signal at 2850 cm^{-1} maps the CH_2 functional group which is abundant in lipids and studies have shown variations in the lipid profile during sepsis development.^{28,29} The TPEF signal excited at 672.5 nm and detected at $426\text{--}490\text{ nm}$ maps different autofluorophores, particularly the NAD(P)H level, which is possibly altered at the inflammation site due to higher cell activity or myeloid cell infiltration.^{13,30–32} Our results show that TPEF (AUC = 0.83) and CARS (AUC = 0.93) features can effectively characterize the texture of the tissue, thus achieving a perfect diagnostic ability for discriminating between the PCI and sham groups. While the classifier trained on SHG features shows an AUC of 0.49. The SHG signal at 415 nm maps fiber structures such as collagen, which can be used as a potential biomarker to monitor sepsis severity.^{33–35} However, SHG features illustrated a poor diagnostic ability (AUC = 0.49) to discriminate between the two groups, possibly due to poor collagen metabolism in the early sepsis stage which the SHG features fail to retain. Lastly, the classifier trained using features

obtained from all the three modalities achieved an AUC of 0.85. The low AUC of this classifier can be due to sample size ($n_{\text{PCI}} = n_{\text{sham}} = 5$) smaller than the number of features ($m = 27$). In conclusion, the classifier trained on the CARS- and TPEF-based features showed better performance than the classifier trained on the features obtained from SHG to classify the two groups.

CONCLUSION

We herein show that the application of multimodal imaging is able to identify biochemical characteristics of septic liver injury with high accuracy at an early time point, even in the absence of significant histological changes. The morphological and chemical composition changes extracted from multimodal imaging provide information on metabolic changes in the early phase of sepsis. As CARS proved most accurate to discriminate septic from control livers, our results suggest early changes in the hepatic intra- and extracellular lipid composition as possible underlying reasons for the observed liver texture transformation.

Although our analysis does not allow discrimination of a specific class of lipids to be differentially regulated, these findings are in agreement with reports of early changes of sphingomyelin and cholesterol composition and metabolism in animal models of septic liver damage occurring as early as 24 h after the induction of sepsis.^{17,36} According to analysis, these characteristics were found without anatomical preference in both the periportal and the pericentral regions of liver lobules. In summary, our data confirm nonlinear multimodal imaging as a promising complementary method for the real-time, label-free study of early septic liver damage.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: [10.1021/acs.analchem.9b01746](https://doi.org/10.1021/acs.analchem.9b01746).

Representative liver sections from sham-treated animals as derived from H&E staining and multimodal imaging. Boxplot of significant features comparing the pericentral regions of livers from animals with sepsis versus sham-treated animals. Definitions and representative images of local texture features (PDF)

AUTHOR INFORMATION

Corresponding Authors

*Phone: +49 (0) 241 80 80866, E-mail: tbruns@ukaachen.de.

*Phone: +49 (0) 3641 32 4401, E-mail: Thomas.bocklitz@uni-jena.de.

ORCID

Michael Schmitt: 0000-0002-3807-3630

Jürgen Popp: 0000-0003-4257-593X

Thomas Wilhelm Bocklitz: 0000-0003-2778-6624

Author Contributions

*M.Y. and P.P. contributed equally to this work. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was funded by the Federal Ministry of Education and Research (BMBF) Germany (FKZ: 01 E0 1002 and via the Integrated Research and Treatment Center “Center for Sepsis Control and Care” (CSCC), FKZ 01EO1502)). The German Research Foundation (DFG) provided research funding to T.B. (BR 4182/3-1), to T.W.B. (BO 4700/1-1) to J.P. (PO 563/30-1), and to A.S. (STA 295/11-1). Additionally, the funding of the CRC 1278 PolyTarget is highly acknowledged.

REFERENCES

- (1) Singer, M.; Deutschman, C. S.; Seymour, C. W.; Shankar-Hari, M.; Annane, D.; Bauer, M.; Bellomo, R.; Bernard, G. R.; Chiche, J. D.; Coopersmith, C. M.; Hotchkiss, R. S.; Levy, M. M.; Marshall, J. C.; Martin, G. S.; Opal, S. M.; Rubenfeld, G. D.; van der Poll, T.; Vincent, J. L.; Angus, D. C. *JAMA* **2016**, *315* (8), 801–10.
- (2) Bauer, M.; Press, A.; Trauner, M. *Curr. Opin Crit Care* **2013**, *19* (2), 123–127.
- (3) Strnad, P.; Tacke, F.; Koch, A.; Trautwein, C. *Nat. Rev. Gastroenterol. Hepatol.* **2017**, *14*, 55.
- (4) Geier, A.; Fickert, P.; Trauner, M. *Nat. Clin. Pract. Gastroenterol. Hepatol.* **2006**, *3*, 574.
- (5) Bruns, T.; Zimmermann, H. W.; Stallmach, A. *World journal of gastroenterology: WJG* **2014**, *20* (10), 2542.
- (6) Kumar, A.; Roberts, D.; Wood, K. E.; Light, B.; Parrillo, J. E.; Sharma, S.; Suppes, R.; Feinstein, D.; Zanotti, S.; Taiberg, L.; et al. *Crit. Care Med.* **2006**, *34* (6), 1589–1596.
- (7) Sartelli, M.; Kluger, Y.; Catena, F. *World Journal of Emergency Surgery* **2018**, *13* (1), 6.
- (8) Lee, J. H.; Kim, J. C.; Tae, G.; Oh, M.-k.; Ko, D.-K. *J. Biomed. Opt.* **2013**, *18* (7), 076009.
- (9) Heuke, S.; Chernavskaja, O.; Bocklitz, T.; Legesse, F. B.; Meyer, T.; Akimov, D.; Dirsch, O.; Ernst, G.; von Eggeling, F.; Petersen, I.; et al. *Head & Neck* **2016**, *38* (10), 1545–1552.
- (10) Smith, L.; MacNeil, S. *Skin Res. Technol.* **2011**, *17* (3), 257–269.
- (11) Horton, N. G.; Wang, K.; Kobat, D.; Clark, C. G.; Wise, F. W.; Schaffer, C. B.; Xu, C. *Nat. Photonics* **2013**, *7* (3), 205.
- (12) Meyer, T.; Bergner, N.; Krafft, C.; Akimov, D.; Dietzek, B.; Popp, J.; Bielecki, C.; Romeike, B. F.; Reichart, R.; Kalf, R. *J. Biomed. Opt.* **2011**, *16* (2), 021113.
- (13) Chernavskaja, O.; Heuke, S.; Vieth, M.; Friedrich, O.; Schürmann, S.; Atreya, R.; Stallmach, A.; Neurath, M. F.; Waldner, M.; Petersen, I.; Schmitt, M.; Bocklitz, T.; Popp, J. *Sci. Rep.* **2016**, *6*, 29239.
- (14) Pradhan, P.; M, T.; Vieth, M.; Stallmach, A.; Waldner, M.; Schmitt, M.; Popp, J.; Bocklitz, T. Semantic Segmentation of Non-linear Multimodal Images for Disease Grading of Inflammatory Bowel Disease: A SegNet-based Application. *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*; Feb 19–21, 2019, Prague, Czech Republic; Volume 1: ICPRAM, 10.
- (15) Heuke, S.; Vogler, N.; Meyer, T.; Akimov, D.; Kluschke, F.; Rowert-Huber, H. J.; Lademann, J.; Dietzek, B.; Popp, J. *Healthcare (Basel, Switzerland)* **2013**, *1* (1), 64–83.
- (16) Gonnert, F. A.; Recknagel, P.; Seidel, M.; Jbeily, N.; Dahlke, K.; Bockmeyer, C. L.; Winning, J.; Lösche, W.; Claus, R. A.; Bauer, M. *J. Surg. Res.* **2011**, *170* (1), e123–e134.
- (17) Chung, H.-Y.; Witt, C. J.; Jbeily, N.; Hurtado-Oliveros, J.; Giszas, B.; Lupp, A.; Gräler, M. H.; Bruns, T.; Stallmach, A.; Gonnert, F. A.; Claus, R. A. *Sci. Rep.* **2017**, *7* (1), 12348.
- (18) Bustin, S. A. *J. Mol. Endocrinol.* **2000**, *25* (2), 169–193.
- (19) Schmittgen, T. D.; Livak, K. J. *Nat. Protoc.* **2008**, *3* (6), 1101.
- (20) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. *Journal of machine learning research* **2011**, *12* (Oct), 2825–2830.
- (21) Jones, E.; Oliphant, T.; Peterson, P. *SciPy: Open source scientific tools for {Python}*, 2014.
- (22) Oliphant, T. E. *Comput. Sci. Eng.* **2007**, *9* (3), 10–20.
- (23) Vogler, N.; Bocklitz, T.; Mariani, M.; Deckert, V.; Markova, A.; Schelkens, P.; Rösch, P.; Akimov, D.; Dietzek, B.; Popp, J. *J. Opt. Soc. Am. A* **2010**, *27* (6), 1361–1371.
- (24) Otsu, N. *IEEE Transactions on Systems, Man and Cybernetics* **1979**, *9* (1), 62–66.
- (25) Guo, S.; Bocklitz, T.; Neugebauer, U.; Popp, J. *Anal. Methods* **2017**, *9* (30), 4410–4417.
- (26) Zhou, R.-R.; Zhao, S.-S.; Zou, M.-X.; Zhang, P.; Zhang, B.-X.; Dai, X.-H.; Li, N.; Liu, H.-B.; Wang, H.; Fan, X.-G. *BMC Gastroenterol.* **2011**, *11* (1), 21.
- (27) Zhang, P.; Xie, M.; Spitzer, J. A. *Shock* **1994**, *2* (2), 133–140.
- (28) Barati, M.; Nazari, M. R.; TalebiTaher, M.; Farhadi, N.; Comparison of lipid profile in septic and non-septic patients. *Iran J. Clin Infect Dis* **2011**, *6* (4).
- (29) Sunayana, P.; Renymol, B.; Ambili, N.; Fasting Lipid Profile and Disease Severity in Sepsis Patients. *Journal of Clinical & Diagnostic Research* **2017**, *11* (11). DOI: 10.7860/JCDR/2017/30268.10820
- (30) Hart, D. W.; Gore, D. C.; Rinehart, A. J.; Asimakis, G. K.; Chinkes, D. L. *J. Surg. Res.* **2003**, *115* (1), 139–47.
- (31) Protti, A.; Fortunato, F.; Artoni, A.; Lecchi, A.; Motta, G.; Mistraretti, G.; Novembrino, C.; Comi, G. P.; Gattinoni, L. *Critical care (London, England)* **2015**, *19*, 39.
- (32) Waldner, M. J.; Rath, T.; Schurmann, S.; Bojarski, C.; Atreya, R. *Front. Immunol.* **2017**, *8*, 1256.
- (33) Gaddnas, F.; Koskela, M.; Koivukangas, V.; Risteli, J.; Oikarinen, A.; Laurila, J.; Saarnio, J.; Ala-Kokko, T. *Critical care (London, England)* **2009**, *13* (2), R53.
- (34) Morrison, G.; Fraser, D. D. *Critical care (London, England)* **2009**, *13* (3), 154.
- (35) Cicchi, R.; Pavone, F. S. *J. Innovative Opt. Health Sci.* **2014**, *07* (05), 1330008.
- (36) Li, J.; Xia, K.; Xiong, M.; Wang, X.; Yan, N. *Exp. Ther. Med.* **2017**, *14* (6), 5635–5640.

Supporting Information

Non-linear multimodal imaging characteristics of early septic liver injury in a mouse model of peritonitis

Melina Yarbakht ^{†,‡,§}, Pranita Pradhan ^{⊥,§}, Nilay Köse-Vogel [†], Hyeonsoo Bae [⊥], Sven Stengel [†], Tobias Meyer ^{⊥,⊥}, Michael Schmitt ^{⊥,⊥}, Andreas Stallmach [†], Jürgen Popp ^{⊥,⊥}, Thomas Wilhelm Bocklitz ^{*,⊥,⊥} and Tony Bruns ^{*,†,‡,§}

[†]Department of Internal Medicine IV (Gastroenterology, Hepatology, Infectious Disease), Jena University Hospital, 07747 Jena, Germany, [‡] Center for Sepsis Control and Care (CSCC), Jena University Hospital, 07747 Jena, Germany, [⊥]Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller University, 07743 Jena, Germany, [⊥]Leibniz Institute of Photonic Technology, 07745 Jena, Germany, [§]Department of Medicine III, University Hospital RWTH Aachen, 52074 Aachen, Germany.

* Corresponding authors: tbruns@ukaachen.de and Thomas.bocklitz@uni-jena.de. §MY and PP contributed equally to this work.

Table of contents-

Figure S1 Representative liver sections from sham-treated animals as derived from H&E staining and multimodal imaging for sham treated animals showing pericentral and periportal veins

Figure S2 Representative liver sections from animals with septic liver injury as derived from H&E staining and multimodal imaging for sham treated animals showing pericentral and periportal veins

Figure S3 Boxplot of significant features comparing the pericentral regions of livers from animals with sepsis versus sham-treated animals.

Figure S4 Boxplot of significant features comparing the periportal regions of livers from animals with sepsis versus sham-treated animals.

Figure S5 Boxplot of significant features comparing the pericentral versus periportal regions of livers from animals with sepsis.

Table S1 Local texture feature extraction and local texture feature images

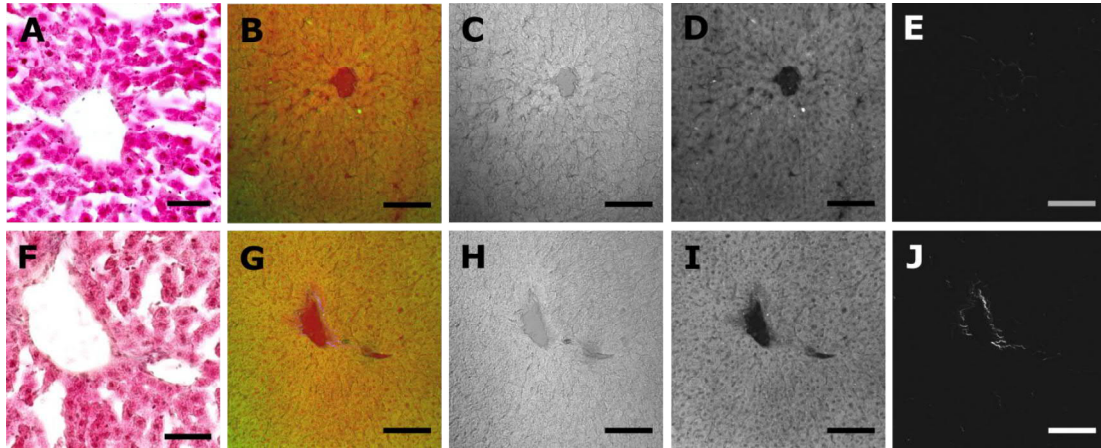


Figure S1. Representative liver sections from sham-treated animals as derived from H&E staining and multimodal imaging for sham treated animals showing pericentral (top panels) and periportal veins (bottom panels). (A, F) indicate H&E stained sections. (B, G) show multimodal images with (C, H) as the red channel indicating CARS signal. Similarly, (D, I) and (E, J) are the green and blue channel of the multimodal images indicating TPEF and SHG signal, respectively. The TPEF channel (D, I) was used for obtaining the binary mask due to its contrast between the tissue and vein region. Scale bar is 100 μ m for multimodal images and 500 μ m for H&E staining with original magnification 60 \times . Original magnification for multimodal images is 20 \times .

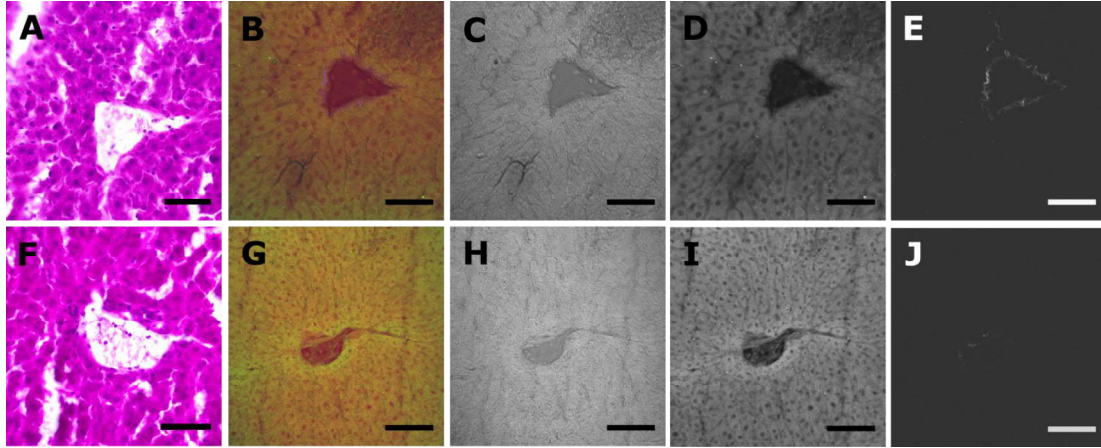


Figure S2. Representative liver sections from animals with septic liver injury as derived from H&E staining and multimodal imaging for sham treated animals showing pericentral (top panels) and periportal veins (bottom panels). (A, F) indicate H&E stained sections. (B, G) show multimodal images with (C, H) as red channel indicating CARS signal. Similarly, (D, I) and (E, J) are the green and blue channel of the multimodal images indicating TPEF and SHG signal, respectively. The TPEF channel (D, I) was used for obtaining the binary mask due to its contrast between the tissue and vein region. Scale bar is 100 μ m for multimodal images and 500 μ m for H&E staining with original magnification 60 \times . Original magnification for multimodal images is 20 \times .

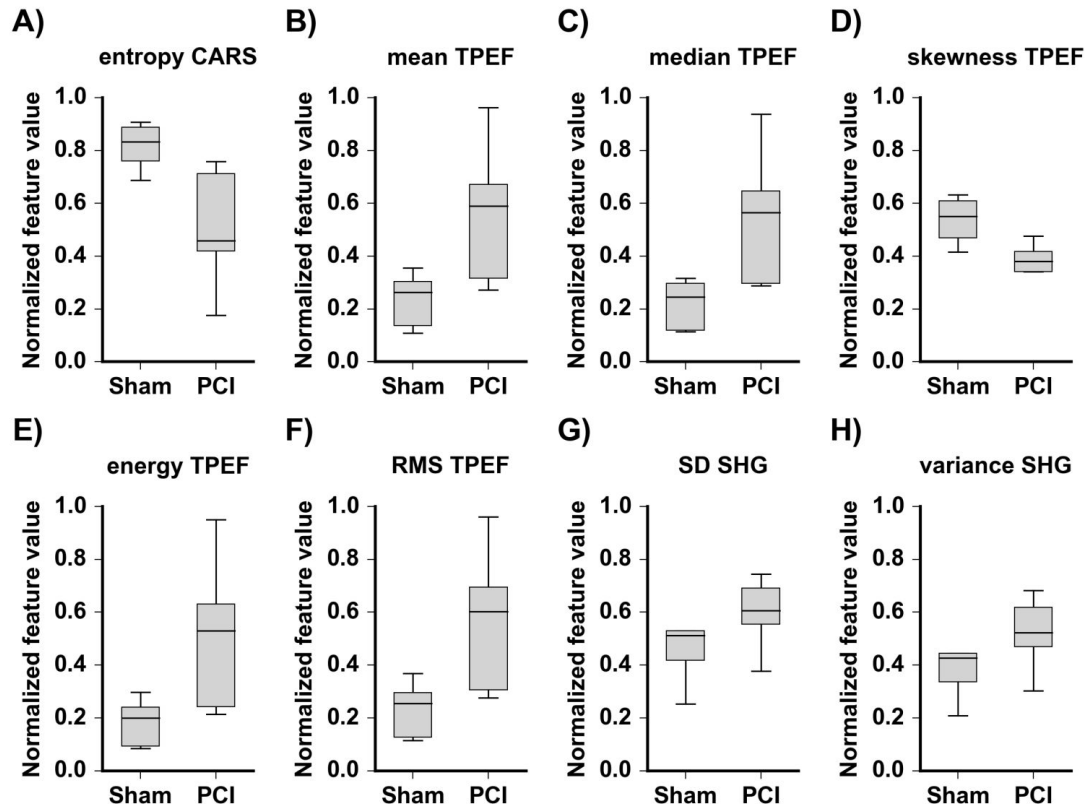


Figure S3. Results of test II- Boxplot of significant features comparing the pericentral regions of livers from animals with sepsis versus sham-treated animals. Eight statistically significant features ($p < 0.05$ in Mann-Whitney U test) that discriminate between sepsis and sham are shown. RMS: Root-Mean-Square. SD: Standard deviation.

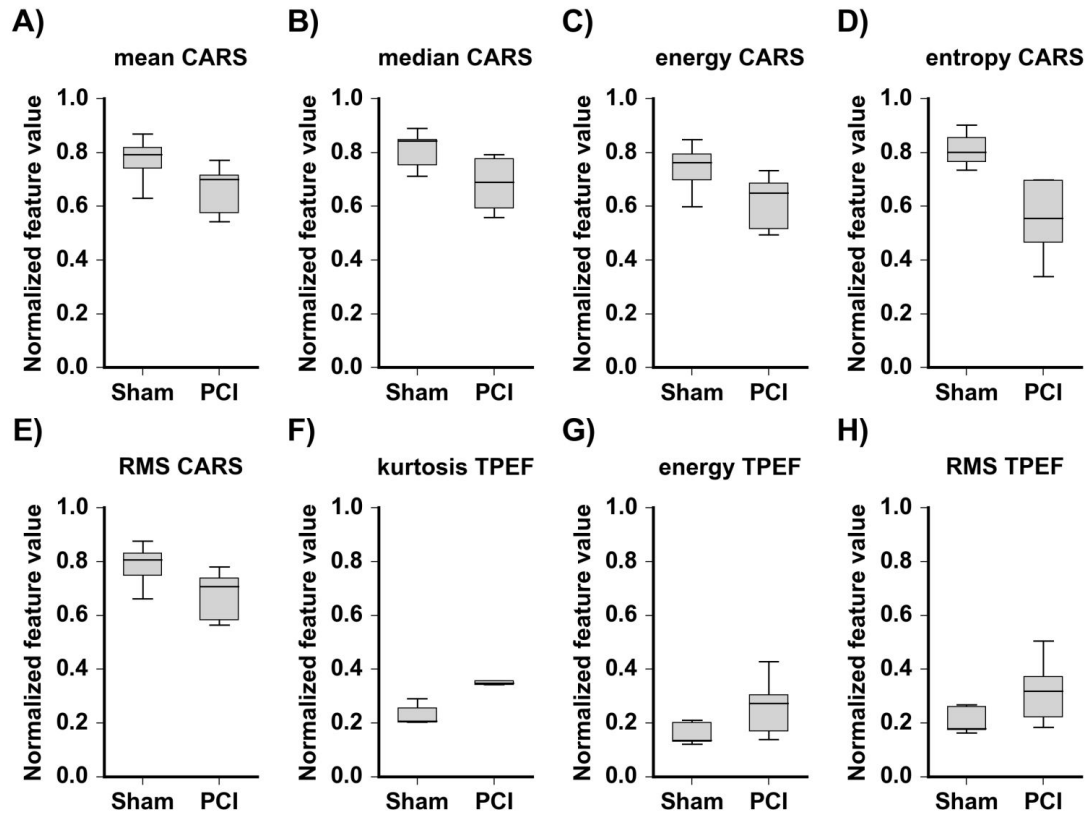


Figure S4. Results of test III- Boxplot of significant features comparing the periportal regions of livers from animals with sepsis versus sham-treated animals. Eight statistically significant features ($p < 0.05$ in Mann-Whitney U test) that discriminate between sepsis and sham are shown. RMS: Root-Mean-Square.

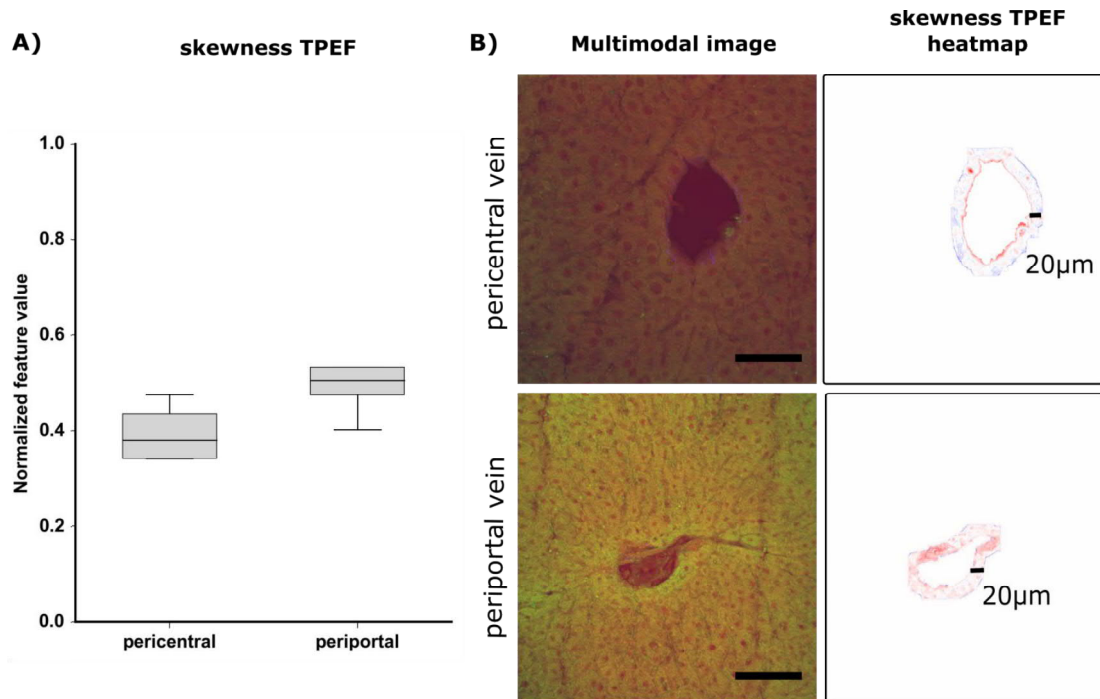
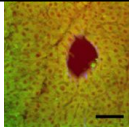
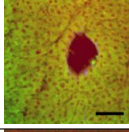
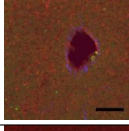
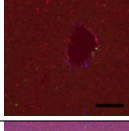
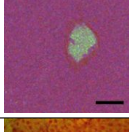
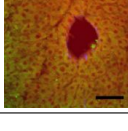


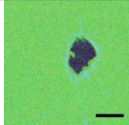
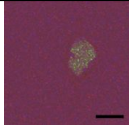
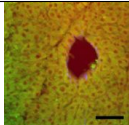
Figure S5. Results of test IV- (A) Boxplot of significant features comparing the pericentral versus periportal regions of livers from animals with sepsis. One statistically significant feature ($p < 0.05$ in Mann-Whitney U test) that discriminates between the periportal and the pericentral region is shown. (B) shows the multimodal image and its corresponding heatmap of skewness TPEF (normalized values) for the two veins. Red color indicates higher skewness values (normalized feature value = 1.00) whereas blue represents lower skewness values (normalized feature value = 0.0). Scale bar for multimodal image is 100µm.

Local first order histogram feature extraction.

Let $I(x,y,z)$ be a gray scale image of size $r \times c$ and J be the $20\mu\text{m}$ region selected around the veins in the image. Variables x and y denote the spatial coordinates of a seed pixel s.t. $(x,y) \in J$ and z denotes the intensity of the seed pixel. To calculate the first order histogram feature locally, a window w of size $m \times n$ (such that $m < r, n < c$) is used around each seed pixel (x,y) to obtain an intensity histogram $p(z_i)$. Based on the intensity histogram $p(z_i)$, nine statistical moments were calculated for a seed pixel. Here, $i = 0,1,2,\dots,T-1$, where T is the number of distinct intensity levels in window w . For our analysis the window size was 5 ($m = n = 5$) and nine statistical moments, $f_{single} \in \mathbb{R}^s$ ($s = 9$) for all pixels in J were calculated for CARS/TPEF/SHG channel. Nine feature images shown below indicate a liver tissue section of a septic mouse with the pericentral vein. Each feature image is a RGB image acquired by a combination of features obtained for all the three channels. For instance, an image of mean texture feature (first row, fourth column) is composed of mean texture feature of the CARS/TPEF/SHG signal as the red/green/blue channel respectively. The formula of the statistical moments, its description and feature images are given below. The feature images are shown on whole 1024×1024 pixel image for better visualization however, the features were extracted only for the selected region J , since it was computational efficient. The original multimodal image of these feature images can be seen in **Supporting Figure S5 (B)**.

Table S1. Nine statistical features calculated locally for every channel (CARS/TPEF/SHG) of the multimodal image. Scale bar is $100\mu\text{m}$.

Statistical feature	Description	Formula	Feature image
Mean	Describes the average tendency of the intensity in a region.	$m = \sum_{i=0}^{T-1} z_i p(z_i)$	
Median	Gives a rough idea about the shape of the histogram.	The median is the value that separates the lower and upper half of the sorted array of pixel values	
Standard deviation	Deviation of the intensity values of the histogram.	$\sigma(z) = \sqrt{\sum_{i=0}^{T-1} (z_i - m)^2 p(z_i)}$	
Variance	Describes how far a value lies from the mean.	$\sigma^2(z) = \sum_{i=0}^{T-1} (z_i - m)^2 p(z_i)$	
Skewness	Degree of asymmetry of the histogram.	$\mu_3(z) = \sum_{i=0}^{T-1} \left[\frac{z_i - m}{\sigma} \right]^3$	
Energy	The value is lowest of coarse texture	$U(z) = \sum_{i=0}^{T-1} p^2(z_i)$	

Entropy	Variability in intensity values.	$e(z) = - \sum_{i=0}^{T-1} p(z_i) \log_2 p(z_i)$	
Kurtosis	Characterizes the relative peakedness and flatness of the histogram.	$\gamma_2 = \sum_{i=0}^{T-1} \left\{ \left[\frac{z_i - m}{\sigma} \right]^4 \right\} - 3$	
RMS	It is a measure of differences of intensity values.	$rms = \sqrt{\sum_{i=0}^{T-1} z_i^2}$	

P4 SEMANTIC SEGMENTATION OF NON-LINEAR MULTIMODAL IMAGES FOR DISEASE GRADING OF INFLAMMATORY BOWEL DISEASE: A SEGNET-BASED APPLICATION

Reprinted with permission from [P. Pradhan, T. Meyer, M. Vieth, A. Stallmach, M. Waldner, M. Schmitt, J. Popp, T. Bocklitz, Semantic Segmentation of Non-linear Multimodal Images for Disease Grading of Inflammatory Bowel Disease: A SegNet-based Application, *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, 2019, Vol. 1: 396–405, Prague, SciTePress]. Copyright 2019 by SCITEPRESS – Science and Technology Publications, Lda.

The declared individual contributions of the doctoral candidate and the other doctoral candidates participate as co-authors in the publications are listed below.

P. Pradhan ¹ , T. Meyer ² , M. Vieth ³ , A. Stallmach ⁴ , M. Waldner ⁵ , M. Schmitt ⁶ , J. Popp ⁷ , T. Bocklitz ⁸ , Semantic Segmentation of Non-linear Multimodal Images for Disease Grading of Inflammatory Bowel Disease: A SegNet-based Application, <i>Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods</i> , 2019, Vol. 1: 396–405, Prague, SciTePress								
Involved in (Please tick the boxes that apply.)								
	1	2	3	4	5	6	7	8
Conceptual research design			X	X	X	X	X	X
Planning of research activities	X							X
Data collection		X						
Data analysis and interpretation	X							X
Manuscript writing	X	X	X	X	X	X	X	X
Suggested publication equivalence value	1.0							

Semantic Segmentation of Non-linear Multimodal Images for Disease Grading of Inflammatory Bowel Disease: A SegNet-based Application

Pranita Pradhan^{1,2}, Tobias Meyer², Michael Vieth³, Andreas Stallmach⁶, Maximilian Waldner^{4,5}, Michael Schmitt¹, Juergen Popp^{1,2} and Thomas Bocklitz^{1,2}

¹*Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller-University, Jena, Germany*

²*Leibniz Institute of Photonic Technology, Member of Leibniz Health Technologies Jena, Germany*

³*Institute of Pathology, Klinikum Bayreuth, Bayreuth, Germany*

⁴*Erlangen Graduate School in Advanced Optical Technologies (SAOT), Friedrich-Alexander University of Erlangen-Nuremberg, Germany*

⁵*Medical Department 1, Friedrich-Alexander University of Erlangen-Nuremberg, Erlangen, Germany*

⁶*Department of Internal Medicine IV (Gastroenterology, Hepatology, and Infectious Diseases), Jena University Hospital, Jena, Germany*

Keywords: Semantic Segmentation, Non-linear Multimodal Imaging, Inflammatory Bowel Disease.

Abstract: Non-linear multimodal imaging, the combination of coherent anti-stokes Raman scattering (CARS), two-photon excited fluorescence (TPEF) and second harmonic generation (SHG), has shown its potential to assist the diagnosis of different inflammatory bowel diseases (IBDs). This label-free imaging technique can support the 'gold-standard' techniques such as colonoscopy and histopathology to ensure an IBD diagnosis in clinical environment. Moreover, non-linear multimodal imaging can measure biomolecular changes in different tissue regions such as crypt and mucosa region, which serve as a predictive marker for IBD severity. To achieve a real-time assessment of IBD severity, an automatic segmentation of the crypt and mucosa regions is needed. In this paper, we semantically segment the crypt and mucosa region using a deep neural network. We utilized the SegNet architecture (Badrinarayanan et al., 2015) and compared its results with a classical machine learning approach. Our trained SegNet model achieved an overall F1 score of 0.75. This model outperformed the classical machine learning approach for the segmentation of the crypt and mucosa region in our study.

1 INTRODUCTION

Histopathological examination represents the 'gold-standard' for diagnosing inflammatory bowel disease (IBD), where the quantification of colonic inflammation is based on the visual appearance of the tissue. However, histopathology delays the diagnosis due to a long sample preparation protocol that includes taking biopsies, tissue embedding, tissue sectioning and staining. In this regard, label-free imaging techniques like multiphoton microscopy (MPM) has been recognized as an *in vivo* imaging technique for IBD diagnostics (Schürmann et al., 2013) (Chernavskaia et al., 2016) (Waldner et al., 2017). These label-free techniques allow a non-destructive investigation of biomolecules in tissue with high tissue penetration depth and spatial resolution (Cicchi and Pavone, 2014) (Vogler et al., 2015).

In the past, MPM techniques like two-photon ex-

cited fluorescence (TPEF) and second harmonic generation (SHG) along with coherent anti-stokes Raman scattering (CARS) were used to visualize biomolecular changes associated with IBDs. Biomolecular information like changed CARS, TPEF and SHG signal intensity along with the crypt morphometries served as predictive marker for an inflamed colon tissue. Likewise, Chernavskaia et al. presented a predictive modelling of histological indexes associated with IBD based on the biomolecular changes of the crypt and mucosa region. Such an automatic predictive modelling of histological indexes is beneficial to accelerate IBD diagnosis. In the work of Chernavskaia et al., the crypt and mucosa region were manually segmented, so a full automatization of the predictive modelling of histological indexes requires a semantic segmentation of crypt and mucosa region without manual effort.

Semantic segmentation of the crypt and mucosa

region is challenging due to several reasons. First, shape irregularities of the crypts add a large biological variance to the data. For example, an inflamed colon tissue reveals crypt deformations and a loss of crypt density, whereas regularly shaped crypts can be found in healthy colon tissue. Second, the crypts are located within the mucosa and therefore the two regions overlap, making the classification even more difficult. Third, the identification of the crypt boundaries is complicated as they are closely located to each other. Lastly, there is a limited amount of annotated medical data, which captures various tissue structures of an inflamed colon. The above mentioned reasons lead to a high morphological variance of the tissue structures thereby making the semantic segmentation of the crypt and mucosa challenging. For this segmentation task machine learning algorithms can be utilized, either classical machine learning or deep learning. Due to this challenging segmentation task mentioned above a high domain-specific representation is needed, which is difficult to obtain using hand-crafted features in classical machine learning.

On the other hand, deep convolutional neural networks (DCNNs) are capable of learning domain-specific representations of an image and have achieved successful results in image classification (Babaie et al., 2017) (Krizhevsky et al., 2012a), object recognition (Pathak et al., 2018) and semantic segmentation (Roth et al., 2015) (Long et al., 2014). Existing DCNNs like U-Net (Ronneberger et al., 2015) and SegNet (Badrinarayanan et al., 2015) have gained state-of-the-art results in biomedical image segmentation and in the field of digital pathology (Janowczyk and Madabhushi, 2016). In this study, we utilize DCNNs to semantically segment multimodal images into biologically significant regions for assisting the predictive modelling of histological indexes. Furthermore, we compare the segmentation results obtained by a DCNN with a classical machine learning approach.

The paper is organized as follows: In section 2 we introduce the previous work related to gland segmentation of histology images, in section 3 we outline our multimodal image dataset and our segmentation workflow. This is followed by a description of the evaluation metrics and a presentation of the results in section 4. We discuss and conclude our work in section 5 and 6, respectively.

2 RELATED WORK

Medical Image Segmentation (MIS) can be utilized for numerous applications like identifying tissue

structures, cell counting, lesion and tumour detection (Norouzi et al., 2014). The approaches for MIS can be categorized into three types. First, the segmentation using classical image processing techniques like thresholding, morphological operations and watershed transform (Wu et al., 2005). Second, training a classification model based on handcrafted image features (classical machine learning) like statistical features, grey level co-occurrence matrix features and local binary patterns (Farjam et al., 2007) (Doyle et al., 2007) (Naik et al., 2008) (Guo et al., 2018). And the third approach is the segmentation using high-level features obtained by a DCNN (Kainz et al., 2017) (Awan et al., 2017) (Chen et al., 2016).

Wu et al. utilized classical image processing algorithms including thresholding and seeded region growing for segmentation of the human intestinal glands. However, this method considered a prior knowledge of the morphological structures of the gland and was qualitatively evaluated (Wu et al., 2005). In another approach by Peng et al., k-means clustering and morphological operations were used to segment the prostate glandular structures. Based on these structures a linear classifier to distinguish normal and malignant glands was constructed (Peng et al., 2011). Peng et al. utilized a k-means clustering algorithm directly on the colour information. Therefore, the approach is not incorporating shape and texture features, which are important for crypt segmentation.

In the contribution by Farzam et al. and Doyle et al., texture, shape and graph-based features were extracted within a classical machine learning approach. Based on these features, a linear classifier to distinguish different pathological tissue sections of the prostate cancer patients was built (Farjam et al., 2007) (Doyle et al., 2007). In the work presented by Naik et al., a Bayesian classifier was used to identify true lumen areas and the false positive lumen areas were removed by applying size and structure constraints. Using the true lumen area, a level set curve (Li et al., 2005) was initialized and evolved until the interior boundary of the nuclei. Morphological features (like distance ratio, compactness, area overlap ratio) were calculated based on the boundaries of the detected lumen area and nuclei. This was followed by a manifold learning scheme called Graph Embedding algorithm (Shi and Malik, 2000) to reduce the dimension of the feature space. Based on the reduced feature space, a support vector machine (SVM) algorithm was used to classify the images into different Gleason grades of prostate cancer (Naik et al., 2008). The above-mentioned methods efficiently segmented regularly shaped gland structures but faced challenges in

segmenting irregularly shaped gland structure.

To tackle this problem, Gunduz-Demir et al. proposed an object-graph based approach that relies on decomposing an image into objects. Their approach used a three-step region growing algorithm, followed by boundary detection and false region elimination (Gunduz-Demir et al., 2010). In another work by Sirinukunwattana et al. (Sirinukunwattana et al., 2015), a Random Polygons Model (RPM) to segment glandular structure in human colon tissue was formulated. The glandular structures were modelled as polygons with random vertices that were located on the cell nuclei within the epithelium. Based on the spatial arrangement of the epithelial nuclei and neighbouring nuclei, an inference of the RPM was made via Reversible-Jump Markov Chain Monte Carlo simulation. False positive polygons were removed by post-processing procedures (Sirinukunwattana et al., 2015). While this technique is stochastic in nature, it can produce different results for the same image and thus a robust approach is needed.

Approaches using DCNNs like AlexNet (Krizhevsky et al., 2012b), VGGNet (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2014), U-Net (Ronneberger et al., 2015) and SegNet (Badrinarayanan et al., 2015) have achieved promising results in MIS. The recent *MICCAI 2015 Gland Segmentation Challenge* presented several innovative algorithms for segmentation of colon glands in histology images (Sirinukunwattana et al., 2015). Chen et al. achieved state-of-the-art performance on the Warwick-QU colon adenocarcinoma dataset by integrating multi-level feature representation with Fully Convolutional Network (FCN) (Chen et al., 2016). Likewise, Kainz et al. used two DCNN that were inspired by the LeNet-5 architecture (LeCun et al., 1998) (Kainz et al., 2017). The first DCNN was used to separate the closely located gland structures and the second DCNN was used to distinguish gland and non-gland regions (Kainz et al., 2017). In Awan et al., a DCNN was used to mark gland boundaries and based on the glandular shape, a two-class and three-class classification model for colorectal adenocarcinoma using histology image was designed (Awan et al., 2017).

In this contribution, we intend to use a SegNet model (Badrinarayanan et al., 2015) for the semantic segmentation of non-linear multimodal images into four distinct regions. Our method is different to the described previous works in the following ways:

- This work is the first to implement semantic segmentation of crypts and mucosa region in non-linear multimodal images. All the above methods have been implemented on H&E (Hematoxylin

and Eosin) stained image which needs a long sample preparation time and leads to sample destruction. In contrast, label-free non-linear multimodal imaging can be used as an *in vivo* technique and its automatic tissue classification can provide a real-time histological index prediction.

- Our method is adapted to multimodal images that show low SNR and are hard to analyze (Vogler et al., 2015).
- Unlike other machine learning methods, we perform a four-class semantic segmentation of multimodal images. In addition to the crypt region we also segment the mucosa region that can be used to assign a histological index.

3 MATERIAL AND METHODS

3.1 Dataset

For this study, we utilized an already published dataset composed of twenty multimodal images sampled from twenty IBD patients. Each multimodal image was converted to an RGB image, which was constructed based on the three modalities, CARS at 2850 cm^{-1} (red channel), TPEF at 458 nm (green channel) and SHG at 415 nm (blue channel). We followed the same image pre-processing steps as explained by Chernavskaja et al. which included downsampling of the multimodal image followed by median filtering, uneven illumination correction (Legesse et al., 2015), background estimation and contrast adjustment (Chernavskaja et al., 2016). The dataset was randomly divided into 11 training, 5 validation and 4 test images. The training dataset was augmented using a rotation angle 60° and 90° .

A histological index between 0 (healthy) and 2 (severe disease) based on crypt architecture, mucosal chronicity and activity was assigned to every image by a trained pathologist. In addition, manually annotated crypt and mucosa regions were obtained as a false-colour image (as shown in figure 1). The manually annotated image is partitioned into four subregions R_l , $l = \{0, 1, 2, 3\}$: mucosa without crypt (R_0) labelled as 0, crypt (R_1) labelled as 1, non-mucosa (R_2) labelled as 2 and background (R_3) labelled as 3.

Table 1: Overview of the dataset.

Dataset	# images	Rotation angle	Total # patches	# selected patches
Train	11	$0^\circ, 60^\circ, 90^\circ$	9.228	3.990
Validation	5	0°	1.168	1.168
Test	4	0°	880	880

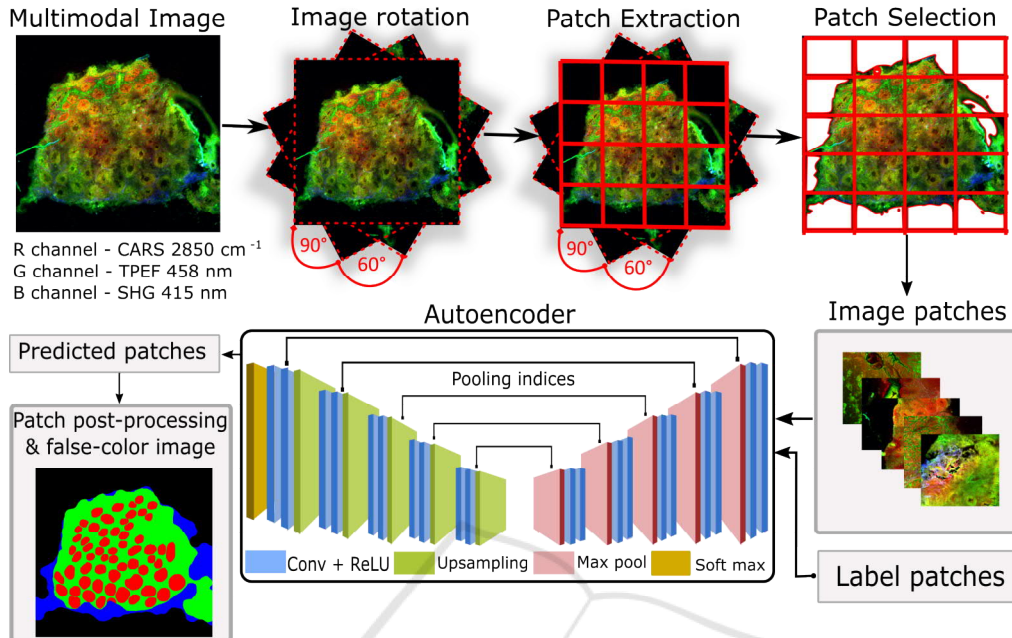


Figure 1: Overview of our proposed segmentation workflow. (1) Data augmentation using random rotations of the multimodal image. (2) Extracting patches of size 256×256 . (3) Selecting patches including only tissue regions. (4) Training a SegNet model using the patches and their label map. (5) Combining the predicted patches into a segmentation map. The segmentation map is depicted as false-colour image where green, red, blue and black represent R_0 (mucosa without crypt), R_1 (crypt), R_2 (non-mucosa), R_3 (background).

3.2 Classical Machine Learning Approach

We formulate the tissue segmentation task as a pixel classification problem. For the pixel classification we extracted texture features locally using first-order statistical moments of the histogram (Guo et al., 2018). These features give an intuitive understanding of the underlying texture in different tissue regions locally and are easy to calculate.

A set of 11 texture features (mean, standard deviation, skewness, kurtosis, median, energy, entropy, RMS, variance, minimum, maximum) was calculated using a window of (5×5) around each pixel for every channel of the multimodal images. This led to 33 texture features (11 features \times 3 channels) per pixel. To reduce the computational complexity, the multimodal images were resized by a factor of 3 and features were calculated for every fifth pixel. Background pixels were excluded with the help of a binary mask which was generated using k-means clustering ($k = 2$) and morphological operations. Morphological operations composed of dilation (kernel size: 9×9 , iterations: 2), closing (kernel size: 9×9 , iterations: 2) and open-

ing (kernel size: 3×3 , iterations: 2) were used. A linear classification model based on a principal component analysis in combination with a linear discriminant analysis (PCA-LDA) was built using the 33 texture features. The model was trained using texture features extracted from 11 training images and its performance was evaluated on the remaining images.

All the computations were performed using Python machine learning library Scikit-learn (Pedregosa et al., 2011), Numpy (Travis E, 2006) and Scipy (Jones et al., 2001). The total execution time was approximately 20 hours on a commercially available PC system Intel® Core™ i5-7500 CPU, 3.40 GHz, 16GB RAM.

3.3 Deep Learning Approach

3.3.1 Encoder-decoder Architecture

We used the SegNet architecture proposed by Badrinarayanan et al. (Badrinarayanan et al., 2015). This network proceeds with an encoder and a decoder with 13 convolutional layers in each. The input to the first layer of the encoder is an image of size $M \times N$.

An activation map of the $(m + 1)^{th}$ encoder layer is given as:

$$x_{m+1} = [\text{MAX}\{\text{ReLU}[\text{CONV}_m\{x_m\} + b_m]\}] \quad (1)$$

$\text{CONV}\{\cdot\}$ is the convolution operator, $\text{ReLU}[\cdot]$ is the rectified linear unit function: $f(x) = \max(0, x)$, $\text{MAX}\{\cdot\}$ is the max pooling layer with a receptive field of (2×2) and a stride of 2, $m \in \{1, 2, \dots, 13\}$ represent index of the convolution layer and b_m is the learned bias of m^{th} layer. The decoder consists of an upsampling, convolution and batch normalization layer. An activation map y_{m+1} of the $(m + 1)^{th}$ decoder layer is given as:

$$y_{m+1} = \text{NORM}[\text{CONV}_m\{\text{US}(y_m) + b_m\}]. \quad (2)$$

Here, y_m is the activation map of m^{th} layer, $\text{US}(\cdot)$ is the upsampling layer and $\text{NORM}[\cdot]$ is the batch normalization layer. The features from the last layer of the decoder are fed to a softmax activation layer. The output of the softmax layer is a c channel image, where c represents the number of classes. A segmented image is generated by assigning each pixel to the class, which had the maximum probability.

3.3.2 Segmentation Workflow

The semantic segmentation of the multimodal image into four regions was performed using a patch based convolutional neural network approach (Jaremenko et al., 2015). This workflow was implemented using Python with the Deep Learning Library Keras (Chollet et al., 2015) with Theano backend (Theano Development Team, 2016).

The model construction started with a patch extraction and a patch selection. Each multimodal image (denoted by \mathbf{I}) of size $M \times N$ was divided into patches (denoted by \mathbf{P}) of size 256×256 pixels without any overlap. The number of patches per image was different as our images were different in size. Each patch P_i can be partitioned into at most four sub-regions (denoted as R_0, R_1, R_2, R_3) such that $\bigcup_{l=0}^3 R_l = P$ where \bigcup represents union set. To remove the background patches from the training set, a homogeneity factor ($H = \sum_{s=1}^K \sum_{t=1}^K \frac{p_{st}}{1+|s-t|}$, where p_{st} is the probability of relative position of a pixel pair, K is the distinct intensity level) was calculated for each patch and a threshold of 60% was optimized such that all the patches belonged to the tissue section. This led to 9.228 training patches. The patches from validation and test set were used for model evaluation. Table 1 shows an overview of the dataset and the patches.

For patch training, the SegNet model (Badrinarayanan et al., 2015) was trained end-to-end to classify the pixels of the multimodal patch into the

four regions. The input of the SegNet model (Badrinarayanan et al., 2015) was a multimodal patch and the output of the model was a segmented patch. The weights of the encoder layers were initialized using VGG16 model pre-trained on ImageNet dataset (Simonyan and Zisserman, 2014) (Russakovsky et al., 2014). We trained the model using a mini-batch of five patches and the stochastic gradient descent optimizer to minimize the cross-entropy loss function. The learning rate was set to 10^{-4} and the training was terminated when the validation loss converged. The total training time was approximately 3 hours on a single NVIDIA GeForce GTX 1060 (6GB memory).

The model performance was evaluated on the test patches. The predicted patches were combined into a whole image, which was called 'segmented map'. This segmented map was post-processed using morphological operations like removing blobs and filling holes. The segmented map was visualized as a false-colour image, wherein the regions R_0 (mucosa without crypt), R_1 (crypt), R_2 (non-mucosa), R_3 (background) were indicated in green, red, blue and black, respectively. The segmented map was visually evaluated, and the quantitative evaluation of the segmented regions was performed by calculating the F1 score and recall as explained in section 4.

4 RESULTS

4.1 Qualitative Evaluation

We visually inspected the segmented map of the validation and the test images. The segmentation of regularly shaped crypts for images with architecture = 0, chronicity = 0, activity = 0, was good. On the other hand, the model performed poorly for segmenting irregularly shaped crypts observed in architecture > 0 and chronicity > 0. The segmentation of the mucosa region was good for all images. We believe that training the SegNet model (Badrinarayanan et al., 2015) with more images of histological index greater than 0 can improve the segmentation performance for images with higher histological indexes, e.g. with stronger altered crypt structures. Also a good quality image with high SNR is required for training the model.

4.2 Quantitative Evaluation

One of the evaluation metrics for classification problems is accuracy, which is misleading for unbalanced class sizes. In our case the number of background

Table 2: Comparison of machine learning and deep learning prediction for R_0 (mucosa without crypt), R_1 (crypt), R_2 (non-mucosa) and R_3 (background) based on F1 score and recall. The values correspond to mean (\pm standard deviation). The number in bold is the best score for classical machine learning and deep learning.

	F1 score				Recall			
	R_0 (mucosa without crypt)	R_1 (crypt)	R_2 (non-mucosa)	R_3 (background)	R_0 (mucosa without crypt)	R_1 (crypt)	R_2 (non-mucosa)	R_3 (background)
Deep learning	0.55 (± 0.17)	0.63 (± 0.13)	0.64 (± 0.14)	0.95 (± 0.02)	0.57 (± 0.22)	0.63 (± 0.17)	0.76 (± 0.15)	0.92 (± 0.03)
Machine learning	0.27 (± 0.11)	0.18 (± 0.18)	0.56 (± 0.23)	0.96 (± 0.02)	0.45 (± 0.12)	0.55 (± 0.24)	0.44 (± 0.25)	1 (± 0.00)

pixels is much higher than the number of pixels belonging to the crypt region, hence accuracy is an inappropriate choice for an evaluation metric.

We evaluated the model performance using F1 score and recall for each region. The number of pixels in the segmented map that intersect with its manually annotated image is considered as true positive. The higher the number of true positives, higher is the F1 score and recall. The two metrics are given by $F_1 = 2TP/(2TP+FN+FP)$ and $Recall = TP/(TP+FN)$, where TP is true positive, FP is false positive and FN is false negative.

In table 2, we report the mean and standard deviation of the F1 score and recall for each region of the segmented maps after post-processing. These values did not change significantly before and after post-processing. The overall segmentation accuracy for the region R_0 , R_1 , R_2 shows that SegNet model (Badrinarayanan et al., 2015) outperformed the classical machine learning approach. Specifically, an overlap of the predicted crypts with manually annotated crypts was 18% and 63% using classical machine learning and deep learning, respectively.

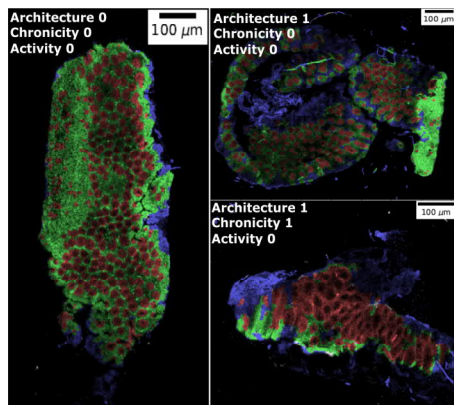


Figure 2: Segmented map superposed with the grey scale multimodal image along with the histological indexes. Regularly shaped crypts (left image) are well segmented whereas a poor segmentation of irregularly shaped crypts is observed (right bottom).

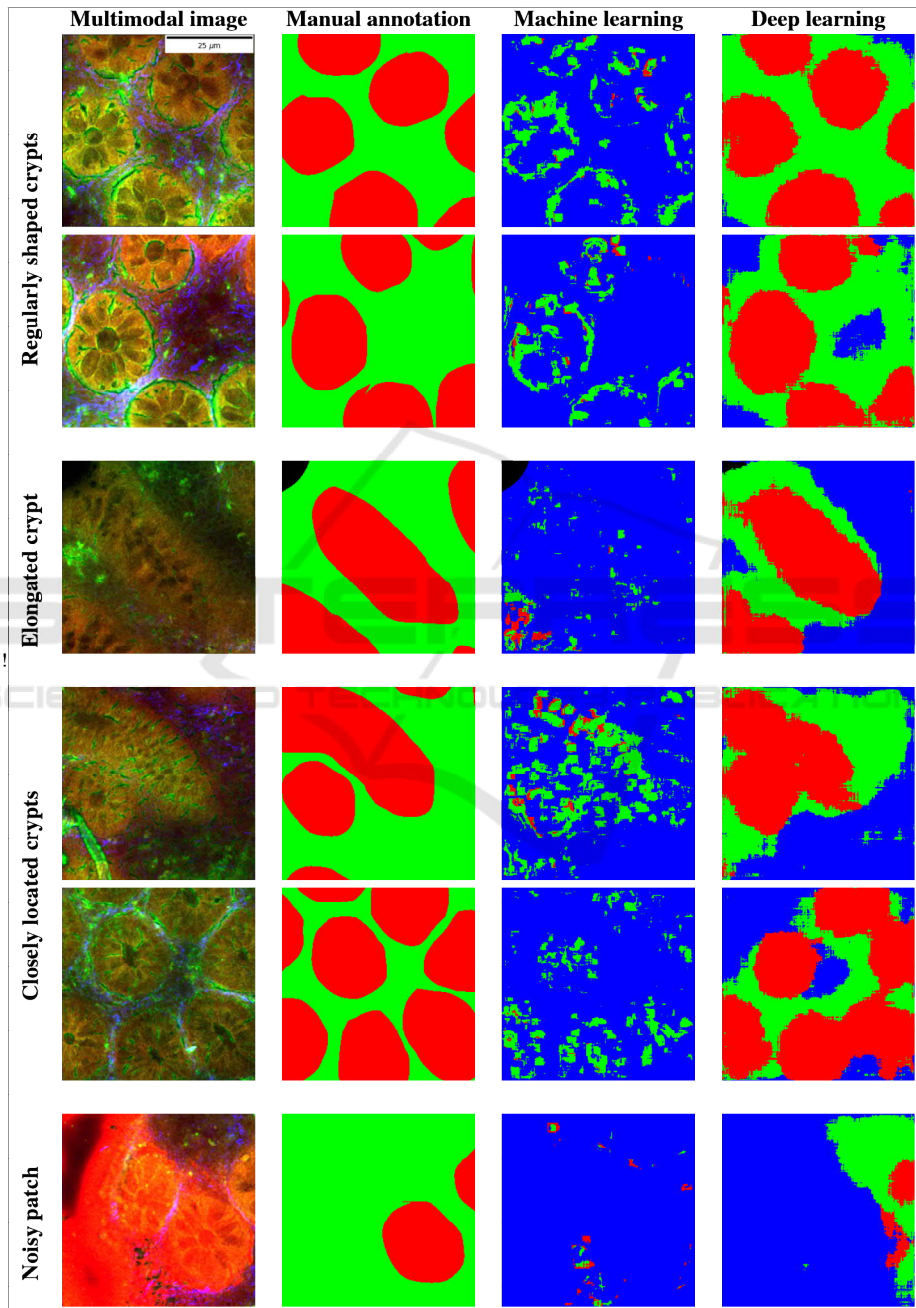
In table 3, we report the F1 scores of R_0 (mucosa without crypt) and R_1 (crypt) for the test and validation image along with its histological indexes for both learning approaches. The evaluation of the segmented regions is important as chronicity of the mucosa and the architecture of crypts serve as predictive marker for the IBD diagnosis. From the results, we observe that using deep learning the two regions for different histological index levels are efficiently segmented. Furthermore, the SegNet model (Badrinarayanan et al., 2015) shows remarkable performance on images with lower histological index (like test image 4). The F1 score for both regions R_0 (mucosa without crypt) and R_1 (crypt) were 0.75 and 0.71, respectively.

However, the trained SegNet model (Badrinarayanan et al., 2015) shows under-segmentation for some cases shown in table 4. The first column is a multimodal image patch, the second column is the manual annotation, the third and fourth column is the prediction by classical machine learning and deep learning (without post-processing), respectively. The crypt segmentation using the handcrafted features show worse performance, whereas the SegNet (Badrinarayanan et al., 2015) model can efficiently segment regularly shaped and distorted crypts. However, the SegNet model (Badrinarayanan et al., 2015) leads to under segmentation of closely located crypts shown in the fourth and fifth row. Also, a noisy patch shown in the last row can degrade the quality of the segmentation.

Table 3: F1 score of the region R_0 (mucosa without crypt) and R_1 (crypt) for validation and test images along with its histological indexes. The best performance is obtained for the image with architecture = 0, activity = 0, chronicity = 0.

Image	Architecture	Chronicity	Activity	$F_1(R_0)$	$F_1(R_1)$
Test image 1	1	1	0	0.69	0.60
Test image 2	1	1	0	0.67	0.69
Test image 3	1	0	0	0.49	0.61
Test image 4	0	0	0	0.76	0.71
Test image 5	1	0	0	0.26	0.32
Val image 6	2	2	1	0.37	0.68
Val image 7	1	0	0	0.74	0.74
Val image 8	1	1	0	0.46	0.65
Val image 9	1	0	0	0.56	0.70

Table 4: The first three rows show regularly sized and deformed crypts with a precise segmentation (without post-processing) using our trained SegNet model. The last three rows illustrate typical cases of under-segmentation, possibly due to closely located crypts or a noisy patch. The region R_0 (mucosa without crypt) is shown in green, R_1 (crypt) is shown in red, R_2 (non-mucosa) is shown in blue and R_3 (background) is shown in black.



5 DISCUSSION

In this paper, we presented a semantic segmentation of non-linear multimodal images to automatize the predictive modelling of histological indexes for characterizing inflammatory bowel disease stages. We used a SegNet (Badrinarayanan et al., 2015) model for the segmentation of multimodal images into mucosa and crypt regions. Moreover, we compared the SegNet (Badrinarayanan et al., 2015) based semantic segmentation of multimodal images with a classical machine learning approach.

For the classical machine learning approach, texture features and linear classifier (PCA-LDA) was chosen due to simplicity. In order to make a fair comparison between the two approaches, same set of training images were used and the window size in the machine learning approach was set comparable to the receptive field of the SegNet model. Optimizing the window size for the machine learning approach did not affect the performance significantly, rather smaller window size increased the computation time. It was observed that the classical machine learning approach along with the hand-crafted features lack the ability to segment the tissue regions, due to a disturbing biological variance resulting from different grades of IBD. As these hand-crafted features are calculated using the intensity at pixel-level, it failed to retain the intrinsic shape information of the crypts. While manually calculated texture features were incapable of segmenting the crypt and mucosa regions, deep neural network like SegNet (Badrinarayanan et al., 2015) achieved reasonable to good result.

Our SegNet model was trained using categorical cross entropy loss function which considers every pixel as an independent sample and asserts equal learning for all pixels. This is a drawback for images with unbalanced classes. Therefore, we believe that weighted pixel wise cross entropy and dice loss function can segment the multimodal images effectively. The weighted pixel wise cross entropy loss in the U-net (Ronneberger et al., 2015) assisted the segmentation of closely located cells in biomedical images. Similarly for closely located crypts more advanced loss functions (Hashemi et al., 2018) can be implemented.

Deep learning approach can generalize the diversity in the underlying data and learn domain-specific representations, although it manifests certain drawbacks. Firstly, it is difficult to understand the contribution of the CARS, TPEF and SHG signal intensity for the segmentation of the mucosa and the crypts. Secondly, a deep learning approach requires large amount of good quality data which is difficult to

obtain particularly in a new technique like non-linear multimodal imaging. Thus, a data augmentation was needed.

For data augmentation, the multimodal images were randomly rotated to consider arbitrary orientations of the multimodal images. This helped to construct a rotation-invariant model. The patches outside the image grid were zero-padded and were filtered by the patch selection process before training the model. However, another possibility could be to mask these zero-padded regions in the loss calculation during model training. Other augmentation techniques like zooming, shearing and resizing of the images affected the spatial resolution and the crypt architecture in the multimodal image. Therefore these techniques were not applied.

In addition to data augmentation, a patch-based DCNN was used to increase the training data and also retain the crypt architecture. The patch size 256×256 was optimized such that maximum tissue structure is retained. Smaller patch size failed to retain information between the crypts and generated more data making the training computationally expensive. The patch based DCNN worked efficiently, but due the combining of the patches to an image a “blocky effect” was generated. “Blocky effect” can also be generated due some other factors like the use of ‘same’ convolutions instead of ‘valid’ convolutions and odd number of feature maps before the pooling layer during training process.

To tackle this “blocky effect” simple post-processing methods were applied, which include morphological operations like remove blobs and region filling to eliminate false positive regions. These post-processing methods improved the segmentation results qualitatively. However, quantitative evaluation of these methods did not show significant changes in the F1 score and recall. Therefore, more complicated post-processing procedures like conditional random field (CRF) (Sutton and McCallum, 2012) are needed which can remove the false positives and improve results quantitatively. Nevertheless, these procedures increase the model complexity. Our post-processing methods led to an under segmentation of the crypts in some patches. This can be misleading in assessing the histological index as fusion of two regularly shaped crypts can be identified as one deformed crypt, leading to false prediction of IBD stage. Therefore, care must be taken while choosing appropriate post-processing procedures.

On the whole, a robust model can be constructed for segmenting the multimodal images with a large number of good quality images. Further, advanced loss function and post-processing procedures as men-

tioned above will need to be explored in future studies.

6 CONCLUSIONS

In summary, we achieved a quantitative evaluation of a semantic segmentation task of non-linear multimodal images to complement IBD diagnosis. An automatic segmentation of the crypt and mucosa region can reduce the manual diagnostic effort and can be used to predict histological indexes in real-time based on non-linear multimodal images. One limitation of our work was a small database with only a few exemplars of high histological index levels. Nevertheless, in future, a large dataset of annotated multimodal images to evaluate the model will be generated and this will improve the model presented here. In conclusion, non-linear multimodal imaging can assist the 'gold-standard' techniques and can be utilized under clinical conditions. Furthermore, incorporating a model for automatic segmentation of multimodal images into the multimodal microscope can provide a real-time histological index prediction and accelerate the start of a clinical therapy.

ACKNOWLEDGEMENTS

Financial support of the EU, the 'Thüringer Ministerium für Wirtschaft, Wissenschaft und Digitale Gesellschaft', the 'Thüringer Aufbaubank', the Federal Ministry of Education and Research, Germany (BMBF), the German Science Foundation (BO 4700/1-1, PO 563/30-1, STA 295/11-1), and Leibniz association via the ScienceCampus 'InfectoOptics' for the project 'BLOODi' are greatly acknowledged.

REFERENCES

- Awan, R., Sirinukunwattana, K., Epstein, D., Jefferyes, S., Qidwai, U., Aftab, Z., Mujeeb, I., Snead, D., and Rajpoot, N. (2017). Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. *Scientific Reports*, 7(1):16852.
- Babaie, M., Kalra, S., Sriram, A., Mitcheltree, C., Zhu, S., Khatami, A., Rahnamayan, S., and Tizhoosh, H. R. (2017). Classification and retrieval of digital pathology scans: A new dataset. *CoRR*, abs/1705.07522.
- Badrinarayanan, V., Kendall, A., and Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561.
- Chen, H., Qi, X., Yu, L., and Heng, P. (2016). DCAN: deep contour-aware networks for accurate gland segmentation. *CoRR*, abs/1604.02677.
- Chernavskaia, O., Heuke, S., Vieth, M., Friedrich, O., Schürmann, S., Atreya, R., Stallmach, A., Neurath, M. F., Waldner, M., Petersen, I., Schmitt, M., Bocklitz, T., and Popp, J. (2016). Beyond endoscopic assessment in inflammatory bowel disease: real-time histology of disease activity by non-linear multimodal imaging. *Scientific Reports*, 6:29239.
- Chollet, F. et al. (2015). Keras.
- Cicchi, R. and Pavone, F. S. (2014). Multimodal non-linear microscopy: A powerful label-free method for supporting standard diagnostics on biological tissues. *Journal of Innovative Optical Health Sciences*, 7(5):1330008.
- Doyle, S., Hwang, M., Shah, K., Madabhushi, A., Feldman, M., and Tomaszewski, J. (2007). Automated grading of prostate cancer using architectural and textural image features. In *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1284–1287.
- Farjam, R., Soltanian-Zadeh, H., Jafari-Khouzani, K., and Zoroofi, R. A. (2007). An image analysis approach for automatic malignancy determination of prostate pathological images. *Clinical Cytometry*, 72B:227–240.
- Gunduz-Demir, C., Kandemir, M., Tosun, A. B., and Sokmensuer, C. (2010). Automatic segmentation of colon glands using object-graphs. *Medical Image Analysis*, 14:1–12.
- Guo, S., Pfeifenbring, S., Meyer, T., Ernst, G., Eggeling, F., Maio, V., Massi, D., Cicchi, R., Pavone, F. S., Popp, J., and Bocklitz, T. (2018). Multimodal image analysis in tissue diagnostics for skin melanoma. *Journal of Chemometrics*, 32:e2963.
- Hashemi, S. R., Salehi, S. S. M., Erdogmus, D., Warfield, S. K., and Gholipour, A. (2018). Asymmetric similarity loss function to balance precision and recall in highly unbalanced deep medical image segmentation. *CoRR*, abs/1803.11078.
- Janowczyk, A. and Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7(1):29.
- Jaremenko, C., Maier, A., Steidl, S., Hornegger, J., Oetter, N., Knipfer, C., Stelzle, F., and Neumann, H. (2015). Classification of confocal laser endomicroscopic images of the oral cavity to distinguish pathological from healthy tissue.
- Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python. [Online; accessed ;today;].
- Kainz, P., Pfeiffer, M., and Urschler, M. (2017). Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization. *PeerJ*, 5:e3874.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012a). Imagenet classification with deep convolutional neu-

- ral networks. In *Advances in Neural Information Processing Systems*.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012b). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12*, pages 1097–1105.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition.
- Legesse, F., Chernavskaya, O., Heuke, S., Bocklitz, T., Meyer, T., Popp, J., and Heintzmann, R. (2015). Seamless stitching of tile scan microscope images. *Journal of Microscopy*, 258(3):223–232.
- Li, C., Xu, C., Gui, C., and Fox, M. D. (2005). Level set evolution without re-initialization: a new variational formulation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 430–436 vol. 1.
- Long, J., Shelhamer, E., and Darrell, T. (2014). Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038.
- Naik, S., Doyle, S., Feldman, M., Tomaszewski, J., and Madabhushi, A. (2008). Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology.
- Norouzi, A., Rahim, M. S. M., Altameem, A., Saba, T., Rad, A. E., Rehman, A., and Uddin, M. (2014). Medical image segmentation methods, algorithms, and applications. *IETE Technical Review*, 31(3):199–213.
- Pathak, A. R., Pandey, M., and Rautaray, S. (2018). Application of deep learning for object detection. *Procedia Computer Science*, 132:1706 – 1717. International Conference on Computational Intelligence and Data Science.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Peng, Y., Jiang, Y., Eisengart, L., Healy, M., Straus, F., and Yang, X. (2011). Computer-aided identification of prostatic adenocarcinoma: Segmentation of glandular structures. *Journal of Pathology Informatics*, 2(1):33.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.
- Roth, H., Farag, A., Lu, L., Turkbey, E. B., and Summers, R. M. (2015). Deep convolutional networks for pancreas segmentation in CT imaging. *CoRR*, abs/1504.03967.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Li, F. (2014). Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575.
- Schürmann, S., Foersch, S., Atreya, R., Neumann, H., Friedrich, O., Neurath, M. F., and Waldner, M. J. (2013). Label-free imaging of inflammatory bowel disease using multiphoton microscopy. *Gastroenterology*, 145(3):514 – 516.
- Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.
- Sirinukunwattana, K., Snead, D. R. J., and Rajpoot, N. M. (2015). A stochastic polygons model for glandular structures in colon histology images. *IEEE Transactions on Medical Imaging*, 34(11):2366–2378.
- Sutton, C. and McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going deeper with convolutions. *CoRR*, abs/1409.4842.
- Theano Development Team (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688.
- Travis E, O. (2006). A guide to numpy. USA: Trelgol Publishing.
- Vogler, N., Heuke, S., Bocklitz, T. W., Schmitt, M., and Popp, J. (2015). Multimodal imaging spectroscopy of tissue. *Annual Review of Analytical Chemistry*, 8:359–387.
- Waldner, M. J., Rath, T., Schürmann, S., Bojarski, C., and Atreya, R. (2017). Imaging of mucosal inflammation: Current technological developments, clinical implications, and future perspectives. *Frontiers in Immunology*, 8:1256.
- Wu, H.-S., Xu, R., Harpaz, N., Burstein, D., and Gil, J. (2005). Segmentation of intestinal gland images with iterative region growing. *Journal of Microscopy*, 220:190–204.

P5 COMPUTATIONAL TISSUE STAINING OF NON-LINEAR MULTIMODAL IMAGING USING GENERATIVE ADVERSARIAL NETWORKS

Reprinted with permission from [P. Pradhan, T. Meyer, M. Vieth, A. Stallmach, M. Waldner, M. Schmitt, J. Popp, T. Bocklitz, Computational tissue staining of non-linear multimodal imaging using Generative Adversarial Networks, *Submitted to Journal Optica*].

The declared individual contributions of the doctoral candidate and the other doctoral candidates participate as co-authors in the publications are listed below.

P. Pradhan ¹ , T. Meyer ² , M. Vieth ³ , A. Stallmach ⁴ , M. Waldner ⁵ , M. Schmitt ⁶ , J. Popp ⁷ , T. Bocklitz ⁸ , Computational tissue staining of non-linear multimodal imaging using Generative Adversarial Networks, <i>Submitted to Journal Optica</i>								
Involved in (Please tick the boxes that apply.)								
	1	2	3	4	5	6	7	8
Conceptual research design			X	X	X	X	X	X
Planning of research activities	X							X
Data collection		X						
Data analysis and interpretation	X							X
Manuscript writing	X	X	X	X	X	X	X	X
Suggested publication equivalence value	1.0							

Computational tissue staining of non-linear multimodal imaging using supervised and unsupervised deep learning

PRANITA PRADHAN,^{1,2} TOBIAS MEYER,² MICHAEL VIETH,³
ANDREAS STALLMACH,⁶ MAXIMILIAN WALDNER,^{4, 5} MICHAEL
SCHMITT,¹ JUERGEN POPP,^{1, 2} AND THOMAS BOCKLITZ^{1, 2}

¹*Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller-University, Jena, Germany.*

²*Leibniz Institute of Photonic Technology, Member of Leibniz Health Technologies Jena, Germany.*

³*Institute of Pathology, Klinikum Bayreuth, Bayreuth, Germany.*

⁴*Erlangen Graduate School in Advanced Optical Technologies (SAOT), Friedrich-Alexander University of Erlangen-Nuremberg.*

⁵*Medical Department 1, Friedrich-Alexander University of Erlangen-Nuremberg, Erlangen, Germany.*

⁶*Department of Internal Medicine IV (Gastroenterology, Hepatology, and Infectious Diseases), Jena University Hospital, Jena, Germany.*

*thomas.bocklitz@uni-jena.de

Abstract: Hematoxylin and Eosin (H&E) staining is the ‘gold-standard’ method in histopathology. However, H&E staining requires long sample preparation time, which restricts its application for ‘real-time’ disease diagnosis. Due to this reason, a label-free alternative technique like non-linear multimodal (NLM) imaging, which is the combination of three non-linear optical modalities including coherent anti-Stokes Raman scattering, two-photon excitation fluorescence and second-harmonic generation, is proposed in this work. To correlate the information of the NLM images with H&E images, this work proposes computational staining of NLM images using deep learning models in a supervised and an unsupervised approach. In the supervised and the unsupervised approach, conditional generative adversarial networks (CGANs) and cycle conditional generative adversarial networks (cycle CGANs) are used, respectively. Both CGAN and cycle CGAN models generate computationally stained H&E images, which are quantitatively analyzed based on mean squared error, structure similarity index and color shading similarity index. The mean of the three metrics are $>5 \times 10^3$, ≥ 0.55 and ≥ 0.9 , respectively. Overall, the computationally stained H&E images obtained from CGAN and cycle CGAN model shows promising results and can be utilized for diagnostic applications without performing a laboratory based staining procedure. To the author’s best knowledge, it is the first time that NLM images are computationally stained to H&E stained images using GANs in an unsupervised manner.

© 2020 Optical Society of America

1. Introduction

Conventional staining techniques, like histopathological (H&E) staining, require long staining protocols and do not exhibit functional and bio-molecular information, but is the ‘gold-standard’ for tissue diagnostics. If bio-molecular information is needed for diagnostics, it must be acquired using molecular imaging techniques. One of the molecular imaging techniques, which can complement the ‘gold-standard’ histopathological staining technique, is non-linear multimodal (NLM) imaging. NLM imaging provides not only structural but also bio-molecular information [1, 2]. Here, we define NLM imaging as the combination of three non-linear optical modalities, namely coherent anti-Stokes Raman scattering (CARS) microscopy, two-photon excitation fluorescence (TPEF) microscopy and second-harmonic generation (SHG) microscopy.

These three modalities highlight the distribution of biomolecules like collagen, NADH, proteins and lipids [2, 3]. Furthermore, these molecular imaging modalities are label-free and provide highly resolved images of biological tissues. The non-destructive nature of these modalities is suitable for *in vivo* studies. Due to these properties and the fact that NLM imaging provides morphological and functional information of a tissue sample, this imaging technique is beneficial for tissue imaging and other biomedical applications [3] like investigations of skin diseases [4–6], diagnostics of head-neck cancer [7, 8], classification of brain tumors [9], and characterization of inflammatory bowel disease samples [10].

Despite the ever-increasing use of NLM imaging, its establishment in clinics is not achieved until now. One of the reasons is the complexity to understand the multimodal images, due to its high spatial resolution and its contrast associated with various biomolecules. To understand these molecular sensitive images and link it to standard histopathological stained images, a parallel tissue section is stained with conventional staining procedures. Subsequently, the stained parallel section is compared with the NLM image. This comparison is laborious, which reduces the advantage of NLM imaging. Therefore, comparison of histopathology and NLM images require an automatic translation of both images.

An automatic translation of different modalities to histopathology can aid intraoperative histopathologic diagnosis and efficient decision-making during surgery [11]. In this context, researchers in 2016 performed the modality transfer of NLM images to histopathological stained H&E images by image analysis and machine learning methods [12]. Although their work showed comparable results, the approach had two limitations. Foremost, the colors in the generated H&E images were different compared to the histopathologically stained H&E images. Secondly, their work trained a machine learning model that required a corresponding pair of NLM images and histopathologically stained H&E images. This training procedure is time-consuming as the histopathologically stained H&E image of the same tissue section must be prepared and registered to the coordinate space of the NLM image before constructing the machine learning model. For the modality transfer using a machine learning model, the multimodal image registration is a difficult task due to tissue alterations that occur during the staining procedure.

In contrast, our work presents an improvement of the work of Bocklitz et al., 2016 in terms of the staining results and the required manual effort for modality transfer. This was achieved by utilizing deep learning models instead of conventional machine learning methods. Briefly, deep learning models, specifically generative adversarial networks [13], were utilized to translate NLM images into computationally stained H&E images. This work was performed in a supervised and an unsupervised approach, where a paired [14] and an unpaired image translation [15] of the NLM image was performed, respectively. The supervised approach or paired image translation required a corresponding pair of images measured with the two modalities (NLM imaging and H&E staining), while the unsupervised approach did not require paired images of the two modalities. Like the previous work of Bocklitz et al. 2016, the supervised approach has the limitation of registering the histopathologically stained H&E images to the corresponding NLM images. On the other hand, the unsupervised approach does not require the image registration of the two modalities.

The supervised and unsupervised approach utilized a conditional generative adversarial network (CGAN) [16] and a cycle conditional generative adversarial network (cycle CGAN) [14], respectively. CGANs are commonly used in computer vision tasks for translating images [14], but they were never used to translate an NLM image to a H&E stained image. Common applications of CGAN in computer vision are the transformation of photographs acquired in daylight into photographs of night scenes, or the transfer of horse images into images of zebras. Likewise, its application in the biomedical and optical field is gaining popularity [17–23]. Recent works transformed auto-fluorescence images [24] or hyperspectral images [25] into histopathologically stained H&E images using CGANs. A similar approach was performed for translating quantitative

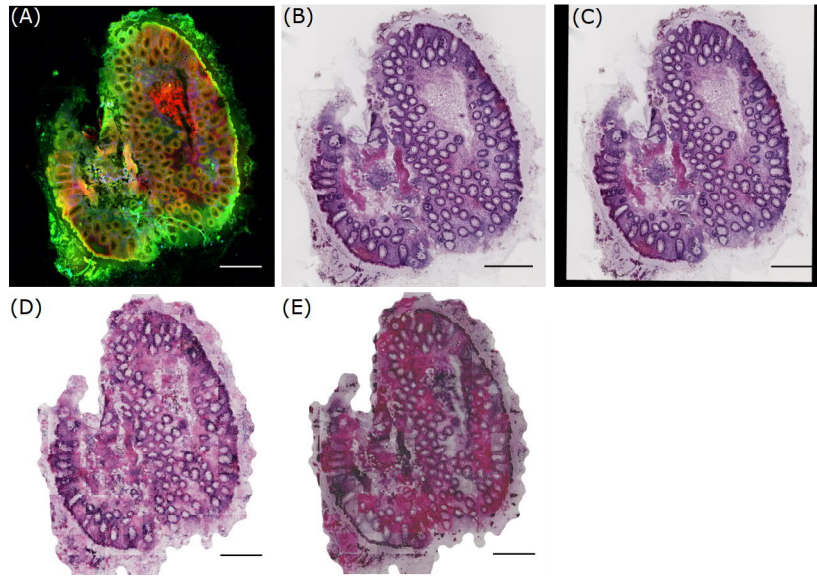


Fig. 1. (A) shows a pre-processed non-linear multimodal image with CARS, TPEF and SHG as the red, green and blue channel, respectively, (B) visualizes pathologically stained (or unregistered) H&E image used for unsupervised pseudo-stain H&E model, (C) depicts a registered H&E stained image used for supervised pseudo-stain H&E model. The image in (C) shows the registration effect, which is filled with zeros. The images in (D) and (E) are computationally stained H&E images with the supervised and unsupervised approach, respectively. All images are downscaled to 20% of the original size for clarity. The scale bar in all images represents 100 μm .

phase imaging into three different stains, namely H&E stain, Jone's stain, and Masson's trichrome stain [26]. CGANs were also employed to increase the spatial resolution [27, 28] and remove speckle noise from optical microscopic images. Similarly, the cycle CGAN were utilized to stain a H&E stained image into an immunohistochemically (IHC) stained image. The generated IHC image was used to reconstruct the original H&E stained image [29]. As mentioned earlier, unpaired image translation is advantageous as co-registration of images from different modalities is not needed, but the model for the unpaired image translation needs to be more complex as compared to the model for paired image translation.

Our work is different from the state-of-the-art methods because it is the first time that NLM images were used for an unsupervised transfer to H&E stained images. While performing the unsupervised modality transfer, the corresponding difficulties were tackled. First of all, the different contrast between the two modalities makes the image translation task complicated. Furthermore, the NLM images used in this work are measured from tissue of patients with different disease severity (namely Inflammatory bowel disease), which is reflected in the alterations of the tissue structure and changes in the pixel contrast [30]. The availability of NLM images is limited, which is problematic because the training of adversarial networks requires large datasets. Lastly, we evaluated the modelling quantitatively by considering perceptual or texture information and color information. Overall, this work is an improvement of the state-of-the-art method presented

by Bocklitz et al. in 2016 [12], based on paired and unpaired image translation of NLM images into H&E stained images.

2. Material and methods

2.1. Dataset

The dataset used in this work is published elsewhere [30]. It contains 19 pairs of non-linear multimodal images (referred to as MM) and histopathologically stained images (referred to as H&E) (see Fig. 1). The non-linear multimodal image is an RGB image where each channel represents one of the three non-linear optical modalities. Precisely, the CARS signal, the TPEF signal, and the SHG signal form the red, green and blue channel of the RGB image, respectively. The spatial (pixel) resolution of the non-linear multimodal image is $0.227 \mu\text{m}/\text{pixel}$ (see Fig. 1A). For the histopathologically stained H&E images, the corresponding tissue sections were stained by an experienced pathologist. The (digital) histopathologically stained H&E images in the form of slide scanner files were extracted using Aperio Image scope software with a spatial resolution approximately equal to the MM image. The spatial resolution of the extracted H&E stained image is $0.219 \mu\text{m}/\text{pixel}$ (see Fig. 1B). The corresponding pairs of MM and H&E stained images were used to construct a “pseudo-stain H&E model” based on the conditional generative adversarial networks in a supervised and an unsupervised approach. The pseudo-stain H&E model was trained using 13 image pairs and tested on six image pairs. For building the pseudo-stain H&E model, both images were pre-processed, and the H&E image was registered to the MM image only for the supervised approach.

2.2. Image pre-processing of histopathologically stained H&E image

The histopathologically stained H&E image was registered to the coordinate space of the corresponding MM image using the Image processing toolbox in Matlab 2018a. For the image registration purpose, the MM and H&E images were converted to grayscale, followed by contrast inversion of the H&E image. The contrast inversion was achieved by subtracting the pixel values in each channel of the H&E image by 255. The inverted H&E image (grayscale) was used as a moving image, and the corresponding MM image (grayscale) was used as a fixed image. Subsequently, a multimodal image registration [31] based on the mutual information metric was performed using the MM and the H&E images. The registered H&E image (see Fig. 1C) was used for the supervised approach or paired image translation, while the unregistered H&E image (see Fig. 1B) was utilized for the unsupervised approach or unpaired image translation. Further, patches of size 256×256 were extracted from the registered and the unregistered H&E image. All the H&E patches were scaled in the range $[-1, 1]$ before model training. The patches from the registered H&E image were used to train the CGAN model, while the patches from the unregistered H&E image were used to train the cycle CGAN model (see Fig. 2A and Fig. 2B).

2.3. Image pre-processing of non-linear multimodal image

The data acquisition and pre-processing of non-linear multimodal images were similar to Chernavskaia et al., 2016 [30]. Briefly, the pre-processing steps included median filtering, downsampling by a factor of 4, correcting the uneven illumination and adjusting the contrast of the MM images. A pre-processed MM image is shown in Fig. 1A. Subsequently, the contrast of MM images was inverted by subtracting the pixel values by 255. Further, patches of size 256×256 were extracted from the “contrast-inverted” MM image (see Fig. 2C and Fig. 2D). These patches were filtered separately for the supervised and unsupervised approach. For the supervised approach or the CGAN model, the pair of MM and H&E patch showing registration artefact were removed. The registration artefacts were seen at the borders of the registered H&E image, which were filled with zero values during registration (see Fig. 1C). For the unsupervised

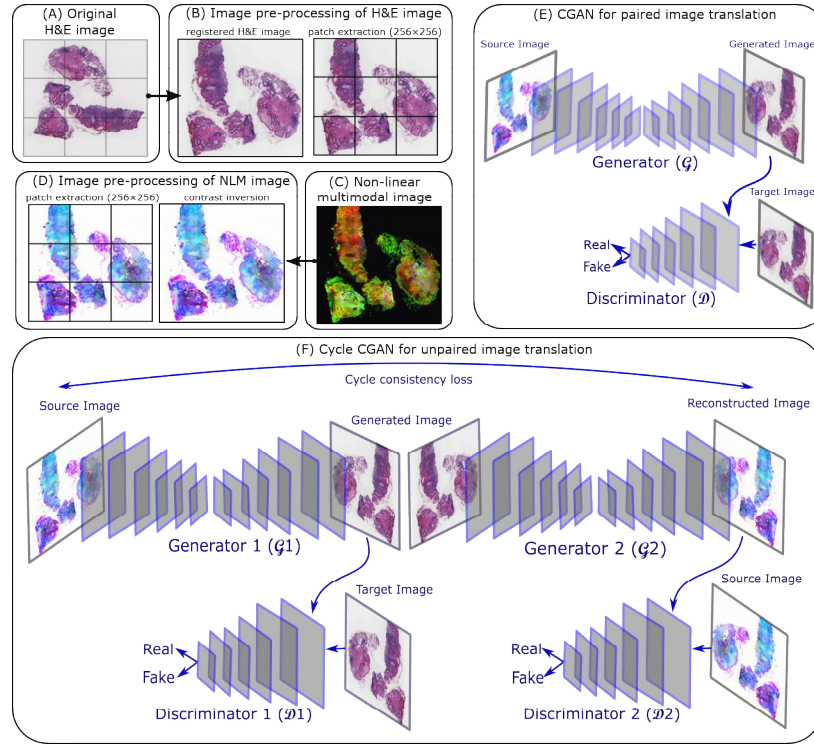


Fig. 2. (A) is a histopathologically stained H&E image, (B) shows the image pre-processing of a histopathological stained H&E image including image registration and patch extraction of size 256×256 , (C) is a corresponding non-linear multimodal image, (D) visualizes the contrast inversion of the non-linear multimodal image followed by patch extraction of size 256×256 , (E) shows a CGAN model for paired image translation which utilizes the registered H&E images and contrast inverted multimodal images, (F) depicts a cycle CGAN model for unpaired image translation using unregistered H&E images and contrast inverted multimodal images.

method or the cycle CGAN model, the MM and H&E patches belonging to the background region were removed using the homogeneity factor [32], i.e. the patches with homogeneity factor greater than 60% were removed [32]. Similar to the H&E patches, all the selected MM patches were normalized in the range $[-1, 1]$ before model training.

2.4. Conditional generative adversarial network

The conditional generative adversarial network (CGAN) used in this work was inspired by the Pix2Pix model developed by Isola et al., 2017 [14]. The Pix2Pix model comprised of a generator (\mathcal{G}) and a discriminator (\mathcal{D}) (see Fig. 2E). The generator with an autoencoder architecture [33] transforms a contrast-inverted MM patch (x_m) to a computationally stained H&E patch ($z_{generated} = \mathcal{G}(x_m)$) which looked visually similar to the histopathologically stained H&E patch (z_{target}). The input to the generator was a pre-processed MM patch (see column B

in Fig. 3) and a target (or histopathologically stained) H&E patch (see column C in Fig. 3). The computationally stained H&E patch, i.e. output of the generator, (see column D in Fig. 3) was evaluated by calculating mean absolute error with the target H&E patch and was optimized to be minimal. The discriminator model was trained to predict the plausibility of the computationally stained H&E patch ($z_{generated}$). In simpler words, the discriminator model was trained to predict if the computationally stained H&E patch was ‘fake’ (i.e. not belonging to histopathologically stained H&E patches) or ‘real’ (i.e. belonging to the original dataset of histopathologically stained H&E patches). The details of the generator and discriminator networks are elaborated below.

The generator network was inspired by the U-Net model [34] which is an autoencoder. The autoencoder model had eight blocks in the encoder and the decoder part. Each block of the encoder utilized convolution layer, batch normalization layer and Leaky ReLU activation layer. The last layer of the encoder was a bottleneck layer without batch normalization layer. The eight encoder blocks comprised of 64, 128, 256, 512, 512, 512, 512 and 512 filters, respectively. On the other hand, each decoder block comprised of a convolution layer, batch normalization layer, dropout layer with a 50% dropout rate and ReLU activation layer. Like the encoder network, the first layer of the decoder did not use a batch normalization layer. The layers in eight decoder blocks comprised of 512, 1024, 1024, 1024, 1024, 512, 256 and 128 filters, respectively (after concatenation from the encoder). All the convolutional layers in the encoder and the decoder blocks used a kernel size of 4 and stride size of 2. The encoder and the decoder models were linked through ‘skip-connections’ similar to the U-Net architecture. The output layer of the generator was a single convolutional layer with three channels and tanh activation function. The output of the generator was a computationally stained H&E image ($z_{generated} = \mathcal{G}(x_m)$) which was one part of the discriminator network’s input.

The discriminator network was a standard convolutional neural network with input as computationally stained ($z_{generated}$) and histopathologically stained H&E patch (z_{target}) of size 256×256 . The architecture of the discriminator network was inspired by the ‘PatchGAN’ discriminator given in reference [14]. The basic idea of the PatchGAN discriminator model is to classify an $N \times N$ region in the $M \times M$ input image ($N < M$) as ‘real’ or ‘fake’, instead of classifying the whole $M \times M$ input image as ‘real’ or ‘fake’. In our case, $M = 256$ and $N = 70$ i.e. a 70×70 region in the 256×256 computationally stained H&E patch was classified as ‘real’ or ‘fake’. The 70×70 region is termed as the ‘receptive field’. The output of the discriminator model was a map with 16×16 values scaled using a sigmoid activation function. In other words, each value in the 16×16 sigmoid activation map corresponded to the probability of the 70×70 region in the input patch being ‘real’ (1.0) or ‘fake’ (0.0). These values were combined to achieve a single probability value, which corresponded to the probability of the entire input patch being ‘real’ or ‘fake’. The layers of the PatchGAN discriminator model were adjusted to maintain the receptive field size to 70×70 . Specifically, the layers of the PatchGAN discriminator model used 64, 128, 256 and 512 filters respectively, and Leaky ReLU activation function with slope 0.2. The configuration of the Leaky ReLU activation function, kernel size and stride size were the same for both the generator and discriminator networks.

Before training of generator and discriminator networks, the weights of both networks were initialized using random Gaussian numbers with a standard deviation of 0.02. During the training phase, the weights of the discriminator model were updated by a set of histopathologically stained H&E patches (z_{target}) and computationally stained H&E patches ($z_{generated}$), and calculating the discriminator loss

$$\mathcal{L}_{\mathcal{D}} = \mathcal{D}(z_{generated})^2 + (1 - \mathcal{D}(z_{target}))^2. \quad (1)$$

When the discriminator network is better than the generator network, i.e. $\mathcal{D}(z_{target}) = 1$ and $\mathcal{D}(z_{generated}) = 0$, it is able to identify all the computationally stained H&E patches as ‘fake’.

To avoid the discriminator network to become better than the generator network, the training process of the discriminator network was slowed down by weighting the discriminator loss $L_{\mathcal{D}}$ by 50% for each model update [35]. The ideal case is to converge the discriminator loss to 0.5 and the generator to create H&E patches exactly similar to the target H&E patches. On the other hand, the weights of the generator network were updated by calculating the mean absolute error between $z_{generated}$ and z_{target} . Additionally, the weights of the generator network were updated through the adversarial loss obtained from the discriminator network. Thus, the total loss of the generator network $\mathcal{L}_{\mathcal{G}}$ is given by

$$\mathcal{L}_{\mathcal{G}} = \lambda \text{MAE}(z_{generated}, z_{target}) + (1 - \mathcal{D}(z_{generated}))^2, \quad (2)$$

where the mean absolute error (MAE) was weighted by a hyperparameter λ . In our case, λ was set to 10. The weights of the generator and discriminator networks were updated separately to avoid misleading updates. Furthermore, both networks were trained using the Adam optimizer [36] with learning rate and β set to 0.0002 and 0.5, respectively.

2.5. Cycle conditional generative adversarial networks

The cycle CGAN model is an extension of the conditional generative adversarial network which does not require paired images for the image translation task [15]. The cycle CGAN model involved simultaneous training of two generators ($\mathcal{G}_1, \mathcal{G}_2$) and two discriminators ($\mathcal{D}_1, \mathcal{D}_2$) (see Fig. 2F). The first generator \mathcal{G}_1 utilized a contrast-inverted MM patch (x_m) (see Fig. 3, column B) as input and generated an H&E stained patch as output ($z_{generated} = \mathcal{G}_1(x_m)$) (see Fig. 3, column E). The second generator utilized the computationally stained H&E patch (i.e. the output of the generator 1, $z_{generated}$) as input and reconstructed it to the original MM patch (similar to the input of the generator 1, $\tilde{x}_m = \mathcal{G}_2(z_{generated})$). The output of the second generator (\tilde{x}_m) was optimized to be visually similar to the input of the first generator (x_m) and was regularized by calculating the (forward) cycle consistency loss $\mathcal{L}_{(cyc_f)}$. In similar fashion, backward cycle consistency loss $\mathcal{L}_{(cyc_b)}$ regulated the second generator. Additionally, the first generator \mathcal{G}_1 was regularized by the identity loss $L_{(id_1)}$ which means that the first generator network utilized the histopathologically stained H&E patch (z_{target}) and reconstructed it at its output. We included only identity mapping loss $\mathcal{L}_{(id_1)}$ for generator \mathcal{G}_1 as we were interested in creating flawless H&E images. However, identity mapping loss $\mathcal{L}_{(id_2)}$ for generator \mathcal{G}_2 can be included in future studies when better reconstruction on MM images is desired. Furthermore, each generator had its own discriminator model, which predicted the plausibility of the generated outputs. This is like the CGAN model explained earlier where each generator-discriminator pair was trained in an adversarial process. The architecture of the two discriminator models in the cycle CGAN model was similar to the Pix2Pix model; however, the architecture of generator networks was different.

The generator networks were inspired by the architecture proposed by Isola et al., 2017 [14]. Both generator networks ($\mathcal{G}_1, \mathcal{G}_2$) used input image size 256×256, and the outputs were a computationally stained H&E patch ($z_{generated}$) and a reconstructed MM patch (\tilde{x}_m), respectively. The generator networks comprised of downsampling convolution blocks to encode the input, a sequence of six ResNet blocks, and upsampling convolution blocks that decodes the bottleneck features to an output. The shorthand notation of the generator network can be given as C7s1-64, D128, D256, R256, R256, R256, R256, U128, U64, C7s1-3 where C7s1-k denotes a 7×7 Convolution-InstanceNorm-ReLU layer with k filters and stride 1. Dk denotes a 3×3 Convolution-InstanceNorm-ReLU layer with k filters and stride 2. Uk denotes a 3×3 fractional-strided-Convolution-InstanceNorm-ReLU layer with k filters and stride 1/2. Rk denotes a ResNet block that contains two 3×3 convolutional layers with the same number of filters on both the layers. Like the CGAN model, the last layer of the generator network comprised of the tanh activation function. The weights of the generator networks were updated through adversarial loss, identity loss [29] and cycle consistency losses [29] (including forward and backward cycle).

Mathematically, the full objective function of the cycle CGAN model can be given as

$$\begin{aligned} \mathcal{L}(\mathcal{G}_1, \mathcal{G}_2, \mathcal{D}_1, \mathcal{D}_2) = & \mathcal{L}_{(\mathcal{D}_1)}(\mathcal{G}_1, \mathcal{D}_1, X, Y) + \mathcal{L}_{(id_1)}(\mathcal{G}_1, \mathcal{D}_1, X) \\ & + \lambda \mathcal{L}_{(cyc_f)}(\mathcal{G}_1, \mathcal{G}_2) + \lambda \mathcal{L}_{(cyc_b)}(\mathcal{G}_1, \mathcal{G}_2), \end{aligned} \quad (3)$$

where $\mathcal{L}_{(\mathcal{D}_1)}$ is the adversarial loss through which the generator 1 was updated. This is mean squared error instead of binary cross-entropy in the Pix2Pix model, as it provided better results in the literature [14]. In future, the adversarial loss for second generator $\mathcal{L}_{(\mathcal{D}_2)}$ can also be added. The identity loss $\mathcal{L}_{(id_1)}$ and the forward and backward cycle loss $\mathcal{L}_{(cyc_f)}$, $\mathcal{L}_{(cyc_b)}$ are the mean absolute error. The four losses were weighted by a factor of 1, 5, 10 and 10, respectively. The training of each generator-discriminator pair was similar to the CGAN model.

2.6. Model training and removal of patch-effect

The Pix2Pix model and the cycle CGAN model were trained on patches obtained from the non-linear multimodal images and histopathologically stained H&E images. Both models were trained for 100 epochs, and a batch size of one patch was used. The model training was performed using Python 3.5 on a commercially available PC system with NVIDIA GeForce GTX 1060, 6GB. The generator models were saved after every fifth epoch, and the model that generated plausible H&E images from the training dataset (on visual inspection) was used for predicting the images from the test dataset. During the testing phase, the test images were pre-processed in a similar fashion as the training dataset. Further, the prediction of the images in the test dataset was performed on patches, which were subsequently combined to a whole image. Combining the patches resulted into a ‘‘patch-effect’’ which was visible at the edge of each patch in the combined image, precisely the pixel at every 256th row or column in the whole image. For this purpose, the pixels which showed the patch-effect were linearly interpolated with its neighbouring three pixels. The generated H&E images (before and after correction of patch-effect) from both Pix2Pix and cycle CGAN models were visually inspected (Fig. S1 in Supplement 1). In addition to the visual inspection, quantitative evaluation of the computationally stained H&E images obtained from both pseudo-stain H&E models was performed. The quantitative evaluation was based on three metrics explained further.

2.7. Evaluation method

For performance quantification, the mean squared error (MSE) was utilized to calculate the error between the histopathologically stained H&E image (z_{target}) and the computationally stained H&E image ($z_{generated}$) [37]. However, MSE has a limitation caused due to arbitrarily high numbers which are difficult to standardize. Also, the MSE metric is inconsistent with the human perception ability [37]. Therefore, two other metrics namely structure similarity index (SSIM) [37, 38] and color shading similarity (CSS) [39], which are well-suited for evaluating GAN performances, were utilized.

The structure similarity index [37, 38] quantifies the perceptual similarity between the two images ($z_{target}, z_{generated}$) by considering the contrast, luminance and texture of these images. Mathematically, SSIM between two images X and Y can be given as

$$SSIM(X, Y) = \frac{(2\mu_X\mu_Y + c_1)(2\sigma_{XY} + c_2)}{(\mu_X^2 + \mu_Y^2 + c_1)(\sigma_X^2 + \sigma_Y^2 + c_2)}, \quad (4)$$

where μ is mean of an image, σ is the standard deviation of an image, σ_{XY} is the cross-covariance of the two images and $C = (c_1, c_2)$ are constants to avoid division by zero. For a multichannel image or an RGB image, the SSIM metric is calculated for each channel separately and the average SSIM value is considered. Higher values of SSIM indicate higher structural similarity between the two images.

Another metric called color shading similarity (CSS) [39] was used to quantify the similarity in the colors of the pixels in the histopathologically stained H&E image (z_{target}) and the computationally stained H&E image ($z_{generated}$). The CSS is calculated by converting both images in the CIELAB space and utilizing only the color channels A* and B*. For each color channel, the mathematical formulation of the CSS metric between two images X and Y is given by

$$CSS(X, Y) = \frac{1}{N} \sum_{i=1}^N Ind_i \cdot Sim(X_i, Y_i), \quad (5)$$

where, $Sim(X_i, Y_i)$ is the similarity between pixel X_i of the pathologically stained H&E image and pixel Y_i of the computationally stained H&E image, i is the index used for the N pixels in the image. Mathematically, Sim and Ind are given as,

$$Sim(X_i, Y_i) = 1 - \frac{dist(X_i, Y_i)}{\max(dist)}, \quad (6)$$

$$Ind_i = \begin{cases} 1; & \text{if } Sim(X_i, Y_i) > \text{threshold,} \\ 0; & \text{if } Sim(X_i, Y_i) \leq \text{threshold.} \end{cases}$$

In our case, the threshold was set to 0.5, and $dist$ was the Euclidean distance. Higher values of the CSS indicate a higher color similarity. Likewise, the three metrics were evaluated for all the computationally stained H&E images (excluding the background region) from the Pix2Pix and the cycle CGAN models. The average of the three metrics calculated for all the 19 images from the training and test dataset is reported in table 1 and Fig. 7.

3. Results

The training of the Pix2Pix model for 100 epochs required ~ 7 hours, while the cycle CGAN model required more than ~ 100 hours on our commercial PC. The training of the cycle CGAN model was terminated after 60 epochs as no significant improvement in the computationally stained H&E patches was observed visually. Furthermore, it was observed that the generator and discriminator losses for the Pix2Pix model and the cycle CGAN model fluctuated throughout the training process. The discriminator loss for both models had difficulties to remain converged at an ideal value ≥ 0.5 , which can be due to high variance [40, 41] and noise in our dataset. The computationally stained H&E patches obtained using the saved generator models were visually assessed for their quality. A detailed explanation of the assessment procedure is provided in the next section.

3.1. Visual similarity of the GAN generated images

Computational staining of MM images using CGANs achieved visually pleasing results compared to the state-of-the-art machine learning model used by Bocklitz et al. in 2016 [12]. As the visual appearance directly impacts the histopathological examination of any disease [25, 42], its qualitative evaluation is vital. For this purpose, computationally stained H&E patches using the Pix2Pix and the cycle CGAN model from the training and testing dataset were inspected.

Fig. 3 and Fig. 4 show computationally stained H&E patches in good and bad quality from the training dataset, respectively. Similarly, Fig. 5 and Fig. 6 show good and bad quality computationally stained H&E patches from the testing dataset. For the good quality patches, it can be observed that computationally stained H&E patches in columns D and E of Fig. 3 and Fig. 5 look visually similar to the histopathologically stained H&E patches in column C. Precisely, the good computationally stained H&E patches show a color contrast similar to the histopathologically stained H&E patches, i.e. the region within the crypts is light or pale pink, whereas the epithelial layer or mucosa region appears dark purple or dark pink. Furthermore,

the tissue structures in the good computationally stained H&E patches are clinically acceptable. Nevertheless, the bad quality patches as shown in columns D and E of Fig. 4 and Fig. 6 looks visually different than the histopathologically stained H&E patches shown in column C. In the bad computationally stained H&E patches, the structures within the crypts are lost and the colors in some regions are wrongly modelled. The bad quality of computationally stained H&E patches in Fig. 4 and Fig. 6 can be explained by the measurement of the images. This is because the computationally stained H&E patches are generated from MM patches, which show slight variations from its corresponding histopathologically stained H&E patches. For example, the MM patch in the top row, column A of Fig. 6 is totally different from the histopathologically stained H&E patch in column C.

The variations between the MM and its corresponding histopathologically stained H&E patch is due to the optical properties of the NLM imaging technique. The NLM imaging technique shows structures in a focal plane within the tissue section, which is approximately $\sim 5\mu\text{m}$. In contrast, the histopathological staining technique reveals the structures from the entire thickness of the tissue section ($\sim 20\mu\text{m}$). Thus, both modalities show slightly different structures, which can be seen through a fine observation of patches in columns A and C of Figs. 3-6. This is the reason that an exact correspondence between computationally stained H&E patch and histopathologically stained H&E patch cannot be achieved for all images in the dataset.

In addition to the structural differences between histopathologically and computationally stained H&E patches, there are other critical issues which need attention. Foremost, it can be

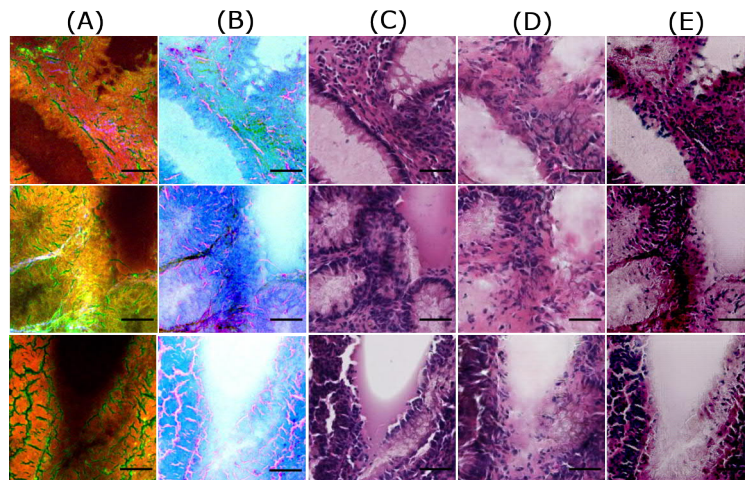


Fig. 3. Good quality predictions of patches from the training dataset. Columns (A) and (B) visualize MM patches and contrast inverted MM patches, column (C) shows histopathologically stained H&E patch, and columns (D) and (E) visualize computationally stained H&E patch generated by the Pix2Pix model and the cycle CGAN model, respectively. The scale bar represents $50\mu\text{m}$. For all patches in column C, the region within the crypts is light or pale pink, whereas the epithelial layer or mucosa region appears dark purple. Similar colors with few variations are observed in the computationally stained H&E patches generated by the Pix2Pix (column D) and the cycle CGAN (column E) model. Also, the crypt structures are efficiently generated in the computationally stained H&E patches by both models.

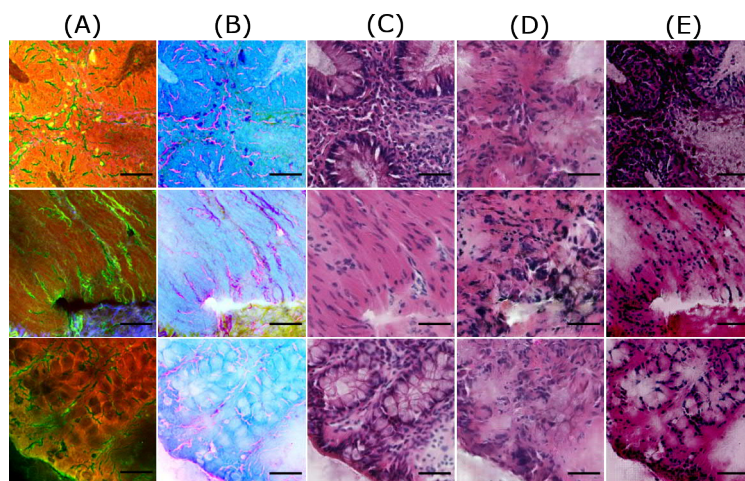


Fig. 4. Bad quality predictions of patches from the training dataset. Columns (A) and (B) visualize MM patches and contrast inverted MM patches, column (C) shows histopathologically stained H&E patch, and columns (D) and (E) visualize computationally stained H&E patch generated by the Pix2Pix model and the cycle CGAN model, respectively. The scale bar represents 50 μm . Here the patches generated by the cycle CGAN model show a promising translation of corresponding MM patches; however, the colors in some regions are not well represented. The computationally stained H&E patches generated by the Pix2Pix model have a low spatial resolution. Thus, the structures within the crypts are not visible.

seen from Figs. 3-6 that the Pix2Pix model generates low spatial resolution images as compared to images generated using the cycle CGAN model. The Pix2Pix model tends to lose detailed boundaries and edges within the crypt region and show a blurry effect, at least for images from the test dataset (see Fig. 5 and Fig. 6). One of the reasons for the loss of detailed information can be the mean absolute error which was used as the objective function while training the generator network in the Pix2Pix model. The next critical issue was that the computationally stained H&E patches generated by the cycle CGAN model showed higher color contrast, thus making the colors more vivid. The high color contrast in the computationally stained H&E patch by the cycle CGAN model can be due to unsupervised training. It is suspected that the unsupervised training of the cycle CGAN model can be sensitive to alterations in the pixel intensity of the multimodal images, staining inconsistencies in the histopathologically stained H&E image or a pre-processing effect [40, 41]. Nevertheless, the problem of high color contrast observed in the computationally stained H&E patches from the cycle CGAN model can be reduced by simple image processing methods (like contrast adjustment [43]).

Overall, from the visual appearance of the computationally stained H&E patches, it can be seen that an exact correspondence with histopathologically stained H&E patches cannot be achieved. However, the computationally stained H&E patches in column D and E of Figs. 3-6 generated using the Pix2Pix and the cycle CGAN model provide an acceptable translation of MM patches. In Fig. S2 and Fig. S3 in Supplement 1, computationally stained H&E patches combined into an image are shown. Furthermore, the computationally stained H&E image from the cycle CGAN model followed by contrast reduction with a factor of 0.7 is also visualized. The computationally

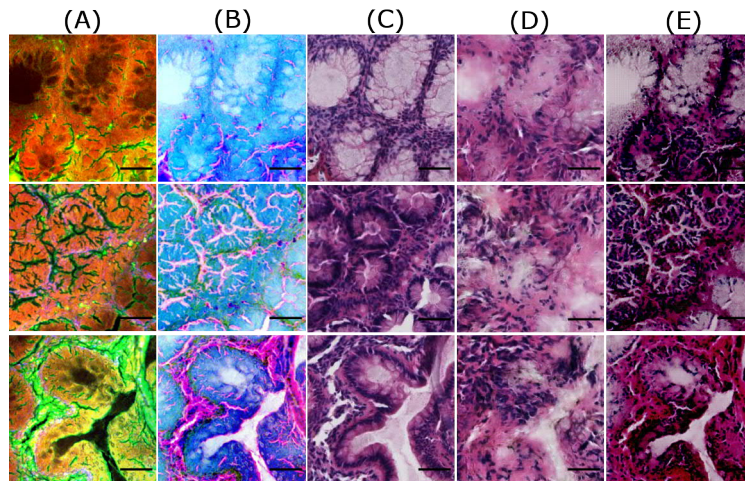


Fig. 5. Good quality predictions of patches from the test dataset. Column (A) shows MM patches, column (B) visualizes contrast inverted MM patches, column (C) shows pathologically stained H&E patch, and columns (D) and (E) depict computationally stained H&E patch by the Pix2Pix model and the cycle CGAN model, respectively. The scale bar represents 50 μm . Here, the computationally stained H&E patches by the cycle CGAN model shows a good quality translation of MM patches, whereas the translation by the Pix2Pix model produces blurry results.

stained H&E images were also examined by a pathologist for its clinical significance. According to the pathological analysis, both models show promising results for translating MM images. In addition to visual analysis, a quantitative evaluation was done and is discussed below.

3.2. Quality of computationally staining based on metrics

An evaluation of the Pix2Pix and the cycle CGAN model was performed based on three metrics: MSE, SSIM and CSS. The average values of MSE, SSIM and CSS for training and testing dataset are reported in Table 1. Here, the three metrics were calculated with the same histopathologically stained H&E image and were considered as baseline values. The aim of the pseudo-stain H&E models was to acquire values “close” to these baseline values.

From Table 1, it can be seen that the computationally stained H&E images generated from both models show very high MSE and low SSIM values as compared to the baseline values. High MSE and low SSIM values were expected as an exact correspondence of computationally stained H&E image with its histopathologically stained H&E image cannot be achieved. Thus, the interpretation of the image quality based on the MSE and SSIM metric is unfair. Despite the high MSE or low SSIM values, the computationally stained H&E images from both models shown in Fig. S2 and Fig. S3 in Supplement 1 have promising visual appearance when compared to its MM image. Furthermore, the MSE values are higher for the cycle CGAN model as compared to the Pix2Pix model. This is can be due to largely different pixel values of computationally stained H&E patches using the cycle CGAN model. On the other hand, the SSIM and CSS metrics report similar performance for the Pix2Pix and the cycle CGAN model, which implies that the overall structural and color content of the computationally stained H&E image is acceptable. Furthermore, the metric values are similar for training and testing dataset (see Table 1) which

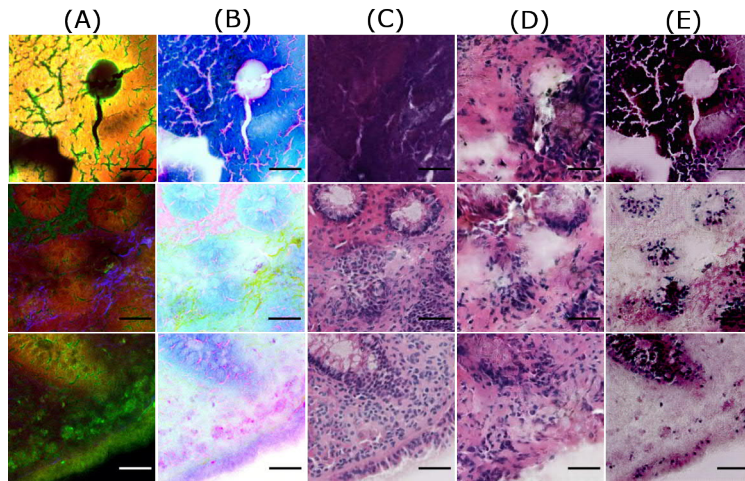


Fig. 6. Bad quality predictions of patches from the test dataset. Column (A) shows MM patches, column (B) visualizes contrast inverted MM patches, column (C) shows pathologically stained H&E patch, and columns (D) and (E) depict computationally stained H&E patch by the Pix2Pix model and the cycle CGAN model, respectively. The scale bar represents 50 μm . Here, computationally stained H&E patches (columns D and E) are not similar to histopathologically stained H&E patches (column C), as histopathologically stained H&E patches show different structures than the corresponding MM patch. Furthermore, in dark MM patches (second and third row), the computationally stained H&E patches fail to generate appropriate color contrast.

shows that the models are minimally overfitted. The mean SSIM and mean CSS metric for the training and the testing dataset using both models are > 0.50 and > 0.90 , respectively. The three metrics for all images are given in Table S1 in [Supplement 1](#).

In addition to Table S1 in [Supplement 1](#), the range of the three metrics is given in Fig. 7, which shows a large variance in the three metrics. The large variance in the three metrics was expected and can possibly be due to the large variance in the dataset. Nevertheless, the color information produced by both pseudo-stain H&E models is close to the baseline value (1.0). The three metrics for a randomly chosen H&E image generated using both models are given in Fig. S2 for the testing dataset and Fig. S3 for the training dataset in [Supplement 1](#). Lastly, there was no significant difference in the performance metrics before and after correction of patch-effect of computationally stained H&E images.

4. Discussion

The computationally stained H&E images generated by the supervised (Pix2Pix) and the unsupervised (cycle CGAN) pseudo-stain H&E model showed a substantial improvement to the state-of-the-art machine learning model [12]. Furthermore, we believe that the cycle CGAN model can be applied for multi-modality conversion, augment the non-linear multimodal images and remove noise from multimodal images. However, in all these tasks, a systematic investigation is needed. Realization of these tasks using the cycle CGAN model can cause staining protocols cost-effective and less labor intensive [11]. However, there are some important aspects considered for training both models, particularly, the training dataset, the pre-processing of the H&E and

Table 1. The average of the three evaluation metrics obtained for the 19 images using the Pix2Pix model and the cycle CGAN model are given for training and testing dataset. For reference purpose, the three metrics were also calculated with the same histopathological stained H&E image. It is seen that MSE values are very large for both models, whereas SSIM and CSS are almost similar for both models. This means that the pixel values of computationally stained H&E images are different, but the overall structural and color information is acceptable. Furthermore, the metric values for training and testing dataset do not have a large difference, which indicates that the models are minimally overfitted.

	MSE	SSIM	CSS
Training dataset			
Pathological stained H&E image	0.00	1.00	1.00
Pix2Pix stained H&E image	4.69×10^3	0.52	0.93
Cycle CGAN stained H&E image	10.26×10^3	0.49	0.91
Testing dataset			
Pathological stained H&E image	0.00	1.00	1.00
Pix2Pix stained H&E image	4.27×10^3	0.60	0.94
Cycle CGAN stained H&E image	7.79×10^3	0.59	0.93

MM image and the objective function. These aspects are discussed in more detail below.

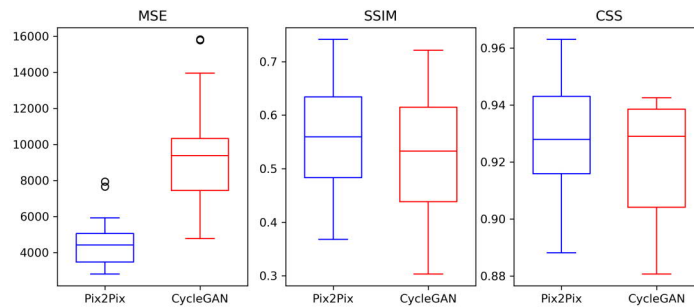


Fig. 7. The boxplot shows a quantitative comparison of the Pix2Pix and the cycle CGAN model based on the three evaluation metrics. The MSE metric is higher for the cycle CGAN model and shows larger variation. This is expected as the pixel values of computationally stained H&E images generated by the cycle CGAN model differ more than the computationally stained H&E images generated by the Pix2Pix model. Nevertheless, the CSS and SSIM metric is in a similar range for both models, which implies that the content of computationally stained H&E images generated by both models is similar.

4.1. Effect of training dataset

The first aspect is the training dataset utilized for constructing the pseudo-stain H&E models. Similar to any other deep learning networks, pseudo-stain H&E model based on GANs are also sensitive to the training dataset. It was observed that the training dataset with a large number of noisy patches or background patches affected the convergence of the generator and the discriminator network. Therefore, patch filtering was vital. Furthermore, a large number of trainable parameters in the generator and the discriminator network can easily cause overfitting on the training dataset. This was a major problem in the supervised approach, e.g. the Pix2Pix model, where the target patches were available. The overfitting on training dataset is seen in Fig. 5 and Fig. 6, column D. Here, we see the patches from the test dataset lose their spatial resolution compared to the patches from the training dataset in the Fig. 3.

Contrarily, in the unsupervised approach, the cycle CGAN model trained on unpaired image data required a quality check of the training dataset. It was observed that for the cycle CGAN model, the color of the computationally stained H&E patches was influenced by the color of the majority of patches in the training dataset. For instance, the cycle CGAN model trained with a large number of patches from the mucosa region, i.e. patches with pink color, was likely to produce pinkish H&E images. Therefore, to create a balance in the color of the generated H&E images, a manual quality check of the patches in the training dataset was crucial for the training of the cycle CGAN model. We believe with an increasing dataset and computation power, the performance of both pseudo-stain H&E models can be improved.

4.2. Effect of the objective function and performance metric

The next aspect for training the pseudo-stain H&E models is the objective function and the performance metric. We begin with the selection of the objective function. Foremost, an appropriate selection of the objective function for the generator and the discriminator network is important to generate plausible H&E images. In this regard, researchers have shown the benefits of using various objective functions like the style transfer loss [44], the perceptual loss [44], the total variation loss [24] and the image gradient loss [45]. Nevertheless, in our case, the L1-loss for the generator network of the cycle CGAN model showed acceptable results. We believe that the addition of other losses to the objective function can improve the perceptual quality of the generated H&E images yet increasing the model complexity. These losses can be applied for the Pix2Pix and the cycle CGAN model and researched in future studies.

The second aspect is the performance metric. The performance metrics used in this work were calculated on the pixel basis and are sensitive to slight variations in the computational H&E images. For instance, a histopathologically stained H&E image and a computationally stained H&E image offset by one pixel can create a major difference in these performance metrics [44]. This problem is often encountered during registration H&E image and MM image. Therefore, the high values of the MSE and low values of the SSIM metric shown in table 1 is justified. In future studies, an objective function that can evaluate the global quality of the computationally stained H&E images can be utilized.

4.3. Effect of image normalization and contrast inversion

In the end, this section discusses the aspect of image normalization and “contrast-inversion” performed while training the pseudo-stain H&E models. Foremost, the normalization methods of both MM and H&E images was essential to avoid multiplications of large numbers during the training process. During the training phase, several methods of normalizing the MM images and pathologically stained H&E images were evaluated. It was observed that the MM and H&E patches scaled in the range $[-1,1]$ generated the best results. It was also observed that scaling of MM and H&E images instead of scaling its patches did not affect the training or model performance. Furthermore, scaling the MM and/or H&E patches in the range $[0,1]$ led to the

failure of the discriminator network by immediately converging the discriminator losses to zero. The scaling of H&E patches was essential due to the tanh activation function used in the last layer of the generator network [13]. These findings coincide with the results of reference [46].

In addition to the normalization, “contrast-inversion” of the multimodal images was performed to remove the “inverse-color” effect [47]. This effect was seen when the original multimodal image (without contrast inversion) was used (see Fig. S4 in [Supplement 1](#)). This effect was especially seen in the unsupervised approach, i.e. using the cycle CGAN model. Because of this effect, the crypt region was transformed into dark purple instead of light pink and vice versa. Therefore, “contrast-inversion” was an important step for modality conversion, especially where the two modalities showed significantly different color contrasts.

5. Conclusion

Computational staining of non-linear multimodal images is beneficial from a clinical perspective as it prevents the long staining protocols and is cost-effective. This work was an improvement of the state-of-the-art method, which utilized the conventional machine learning approach for computational staining of non-linear multimodal images. On the contrary, this work presented a supervised and unsupervised approach to computationally stain non-linear multimodal images into H&E stained images. The supervised approach utilized the Pix2Pix model, and the unsupervised approach used the cycle CGAN model. For the Pix2Pix model, a corresponding pair of non-linear multimodal image and histopathologically stained H&E image was required. Therefore, image registration of the histopathologically stained H&E image was crucial. On the other hand, the cycle CGAN model did not require the corresponding pair of the non-linear multimodal image and histopathologically stained H&E image. Thus, the effort of image registration and pathological staining was reduced. The qualitative and quantitative evaluation of both models showed comparable results using evaluation metrics based on color, texture and perceptual quality. The evaluation metric like mean squared error reported values $>5 \times 10^3$ and $>8 \times 10^3$ for the Pix2Pix and the cycle CGAN model, respectively. In contrast, the evaluation metric, including SSIM and CSS reported values >0.50 and >0.90 for both models, respectively. In addition to quantitative evaluation, various pre- and post-processing procedures were explored in this work, however more advanced post-processing procedures could be investigated in future. Furthermore, a cycle CGAN model that can perform multiple staining using a non-linear multimodal image can be one of the future research directions. The cycle CGAN model can also be used for additional benefits like the artificial generation of non-linear multimodal images, increasing the spatial resolution of the computationally stained H&E images and removing fluorescence effect from the reconstructed non-linear multimodal images. However, a systematic investigation for such tasks is needed. Overall, the results showed several benefits of using computational staining of non-linear multimodal images than performing histopathological staining in laboratories. Thus, the computational staining approach should be encouraged in clinics to benefit the pathological and clinical field of science.

Funding

Funding of the German Research Foundation (DFG) for the projects BO 4700/1-1, BO 4700/4-1, PO 563/30-1 and STA 295/11-1 is highly acknowledged. This work received financial support by the Ministry for Economics, Sciences and Digital Society of Thuringia (TMWWDG), under the framework of the Landesprogramm ProDigital (DigLeben-5575/10-9).

Disclosures

The authors declare no conflicts of interest.

See Supplement 1 for supporting content.

References

1. T. Bocklitz, A. Silge, H. Bae, M. Rodewald, F. B. Legesse, T. Meyer, and J. Popp, "Non-invasive imaging techniques: From histology to in vivo imaging," in *Molecular Imaging in Oncology*, (Springer, 2020), pp. 795–812.
2. N. Vogler, S. Heuke, T. W. Bocklitz, M. Schmitt, and J. Popp, "Multimodal imaging spectroscopy of tissue," *Annu. Rev. Anal. Chem.* **8**, 359–387 (2015).
3. R. Cicchi and F. S. Pavone, "Multimodal nonlinear microscopy: A powerful label-free method for supporting standard diagnostics on biological tissues," *J. Innov. Opt. Heal. Sci.* **7**, 1330008 (2014).
4. S. Heuke, N. Vogler, T. Meyer, D. Akimov, F. Kluschke, H. R wert-Huber, J. Lademann, B. Dietzek, and J. Popp, "Multimodal mapping of human skin," *Br. J. Dermatol.* **169**, 794–803 (2013).
5. S. Heuke, N. Vogler, T. Meyer, D. Akimov, F. Kluschke, H.-J. R wert-Huber, J. Lademann, B. Dietzek, and J. Popp, "Detection and discrimination of non-melanoma skin cancer by multimodal imaging," in *Healthcare*, vol. 1 (Multidisciplinary Digital Publishing Institute, 2013), pp. 64–83.
6. S. Guo, S. Pfeifenbring, T. Meyer, G. Ernst, F. von Eggeling, V. Maio, D. Massi, R. Cicchi, F. S. Pavone, and J. Popp, "Multimodal image analysis in tissue diagnostics for skin melanoma," *J. Chemom.* **32**, e2963 (2018).
7. S. Heuke, O. Chernavskaia, T. Bocklitz, F. B. Legesse, T. Meyer, D. Akimov, O. Dirsch, G. Ernst, F. von Eggeling, and I. Petersen, "Multimodal nonlinear microscopy of head and neck carcinoma—toward surgery assisting frozen section analysis," *Head & neck* **38**, 1545–1552 (2016).
8. T. Meyer, O. Guntinas-Lichius, F. von Eggeling, G. Ernst, D. Akimov, M. Schmitt, B. Dietzek, and J. Popp, "Multimodal nonlinear microscopic investigations on head and neck squamous cell carcinoma: Toward intraoperative imaging," *Head & neck* **35**, E280–E287 (2013).
9. T. Meyer, N. Bergner, C. Krafft, D. Akimov, B. Dietzek, J. Popp, C. Bielecki, B. F. Romeike, R. Reichart, and R. Kalff, "Nonlinear microscopy, infrared, and Raman microspectroscopy for brain tumor analysis," *J. biomedical optics* **16**, 021113 (2011).
10. S. Sch rmmann, S. Foersch, R. Atreya, H. Neumann, O. Friedrich, M. F. Neurath, and M. J. Waldner, "Label-free imaging of inflammatory bowel disease using multiphoton microscopy," *Gastroenterology* **145**, 514–516 (2013).
11. D. A. Orringer, B. Pandian, Y. S. Niknafs, T. C. Hollon, J. Boyle, S. Lewis, M. Garrard, S. L. Hervey-Jumper, H. J. Garton, C. O. Maher *et al.*, "Rapid intraoperative histology of unprocessed surgical specimens via fibre-laser-based stimulated Raman scattering microscopy," *Nat. biomedical engineering* **1**, 0027 (2017).
12. T. W. Bocklitz, F. S. Salah, N. Vogler, S. Heuke, O. Chernavskaia, C. Schmidt, M. J. Waldner, F. R. Greten, R. Br uer, and M. Schmitt, "Pseudo-HE images derived from CARS/TPEF/SHG multimodal imaging in combination with Raman-spectroscopy as a pathological screening tool," *BMC cancer* **16**, 534 (2016).
13. I. J. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," *ArXiv abs/1701.00160* (2017).
14. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2017), pp. 1125–1134.
15. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, (2017), pp. 2223–2232.
16. M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784* (2014).
17. Y. Ma, X. Chen, W. Zhu, X. Cheng, D. Xiang, and F. Shi, "Speckle noise reduction in optical coherence tomography images based on edge-sensitive cGAN," *Biomed. Opt. Express* **9**, 5129–5146 (2018).
18. R. Zheng, L. Liu, S. Zhang, C. Zheng, F. Bunyak, R. Xu, B. Li, and M. Sun, "Detection of exudates in fundus photographs with imbalanced learning using conditional generative adversarial network," *Biomed. Opt. Express* **9**, 4863–4878 (2018).
19. C. Zhang, K. Wang, Y. An, K. He, T. Tong, and J. Tian, "Improved generative adversarial networks using the total gradient loss for the resolution enhancement of fluorescence images," *Biomed. Opt. Express* **10**, 4742–4756 (2019).
20. H. Zhang, C. Fang, X. Xie, Y. Yang, W. Mei, D. Jin, and P. Fei, "High-throughput, high-resolution deep learning microscopy based on registration-free generative adversarial network," *Biomed. Opt. Express* **10**, 1044–1063 (2019).
21. J. Ouyang, T. S. Mathai, K. Lathrop, and J. Galeotti, "Accurate tissue interface segmentation via adversarial pre-segmentation of anterior segment OCT images," *Biomed. Opt. Express* **10**, 5291–5324 (2019).
22. H. Jiang, X. Chen, F. Shi, Y. Ma, D. Xiang, L. Ye, J. Su, Z. Li, Q. Chen, Y. Hua, X. Xu, W. Zhu, and Y. Fan, "Improved cGAN based linear lesion segmentation in high myopia ICGA images," *Biomed. Opt. Express* **10**, 2355–2366 (2019).
23. K. J. Halupka, B. J. Antony, M. H. Lee, K. A. Lucy, R. S. Rai, H. Ishikawa, G. Wollstein, J. S. Schuman, and R. Garnavi, "Retinal optical coherence tomography image enhancement via deep learning," *Biomed. Opt. Express* **9**, 6205–6221 (2018).
24. Y. Rivenson, H. Wang, Z. Wei, Y. Zhang, H. Gunaydin, and A. Ozcan, "Deep learning-based virtual histology staining using auto-fluorescence of label-free tissue," *Nat. Biomed. Eng.* pp. 466–477 (2018).
25. N. Bayramoglu, M. Kaakinen, L. Eklund, and J. Heikkil , "Towards virtual H&E staining of hyperspectral lung histology images using conditional generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, (2017), pp. 64–71.
26. T. Liu, Z. Wei, Y. Rivenson, K. de Haan, Y. Zhang, Y. Wu, and A. Ozcan, "Deep learning-based color holographic microscopy," *J. biophotonics* **12**, e201900107 (2019).

27. E. Nehme, L. E. Weiss, T. Michaeli, and Y. Shechtman, "Deep-storm: super-resolution single-molecule microscopy by deep learning," *Optica* **5**, 458–464 (2018).
28. H. Wang, Y. Rivenson, Y. Jin, Z. Wei, R. Gao, H. Günaydin, L. A. Bentolila, C. Kural, and A. Ozcan, "Cross-modality deep learning achieves super-resolution in fluorescence microscopy," in *2019 Conference on Lasers and Electro-Optics (CLEO)*, (IEEE, 2019), pp. 1–2.
29. Z. Xu, C. F. Moro, B. Bozóky, and Q. Zhang, "GAN-based virtual re-staining: a promising solution for whole slide image analysis," arXiv preprint arXiv:1901.04059 (2019).
30. O. Chernavskaia, S. Heuke, M. Vieth, O. Friedrich, S. Schürmann, R. Atreya, A. Stallmach, M. F. Neurath, M. Waldner, and I. Petersen, "Beyond endoscopic assessment in inflammatory bowel disease: real-time histology of disease activity by non-linear multimodal imaging," *Sci. reports* **6**, 29239 (2016).
31. A. Roche, G. Malandain, X. Pennec, and N. Ayache, "The correlation ratio as a new similarity measure for multimodal image registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (Springer, 1998), pp. 1115–1124.
32. P. Pradhan, T. Meyer, M. Vieth, A. Stallmach, M. Waldner, M. Schmitt, J. Popp, and T. Bocklitz, "Semantic segmentation of non-linear multimodal images for disease grading of inflammatory bowel disease: A segnet-based application," in *International Conference on Pattern Recognition Applications and Methods 2019*, (2019).
33. P. Pradhan, S. Guo, O. Ryabchykov, J. Popp, and T. W. Bocklitz, "Deep learning a boon for biophotonics?" *J. Biophotonics* p. e201960186 (2020).
34. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, (Springer, 2015), pp. 234–241.
35. H. Cho, S. Lim, G. Choi, and H. Min, "Neural stain-style transfer learning using GAN for histopathological images," arXiv preprint arXiv:1710.08543 (2017).
36. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 (2014).
37. H. Ren, J. Li, and N. Gao, "Automatic sketch colorization with tandem conditional adversarial networks," in *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 01 (2018), pp. 11–15.
38. Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *Proceedings (ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 2 (IEEE, 2005), pp. ii/573–ii/576 Vol. 2.
39. D. Wang, L. Shi, Y. J. Wang, G. C. Man, P. A. Heng, J. F. Griffith, and A. T. Ahuja, "Color quantification for evaluation of stained tissues," *Cytom. Part A* **79**, 311–316 (2011).
40. M. T. McCann, J. A. Ozolek, C. A. Castro, B. Parvin, and J. Kovacevic, "Automated histology analysis: Opportunities for signal processing," *IEEE Signal Process. Mag.* **32**, 78–87 (2014).
41. N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," in *2016 23rd International conference on pattern recognition (ICPR)*, (IEEE, 2016), pp. 2440–2445.
42. A. BenTaieb and G. Hamarneh, "Adversarial stain transfer for histopathology image analysis," *IEEE transactions on medical imaging* **37**, 792–802 (2017).
43. R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using MATLAB* (Prentice-Hall, Inc., USA, 2003).
44. J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, (Springer, 2016), pp. 694–711.
45. D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (Springer, 2017), pp. 417–425.
46. A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *CoRR abs/1511.06434* (2016).
47. T. Wang and Y. Lin, "CycleGAN with better cycles," (2018).

Computational tissue staining of non-linear multimodal imaging using supervised and unsupervised deep learning

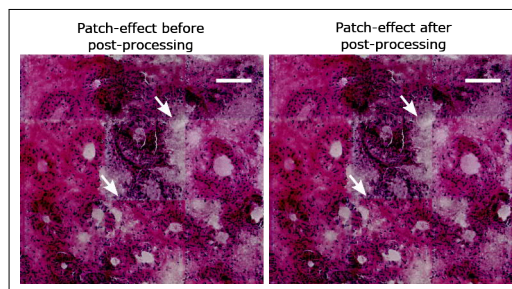


Fig. S1. This figure shows the effect of post-processing for removing 'patch-effect'. The 'patch-effect' was removed by interpolating pixel values of three neighbouring pixels at the end of every patch (256th pixel). This effect (shown in white arrows) was visible for few images and its removal did not significantly affect the performance metrics. The scale bar represents 100 μm

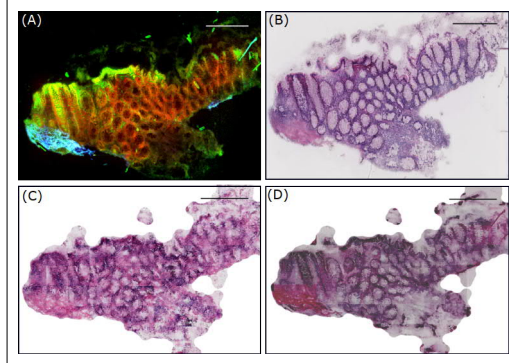


Fig. S2. (A) shows a non-linear multimodal image from test dataset, (B) visualizes corresponding histopathologically stained H&E image (unregistered), (C) shows the computational H&E image by the Pix2Pix ($MSE = 4.4 \times 10^3$, $SSIM = 0.65$, $CSS = 0.94$) and (D) depicts computational H&E image by the cycle CGAN model ($MSE = 8.4 \times 10^3$, $SSIM = 0.63$, $CSS = 0.94$). The contrast of the computational H&E image in (D) is reduced by a factor of 0.7. The images here are downsampled to 20% of original size for clarity. The scale bar represents $100 \mu m$.

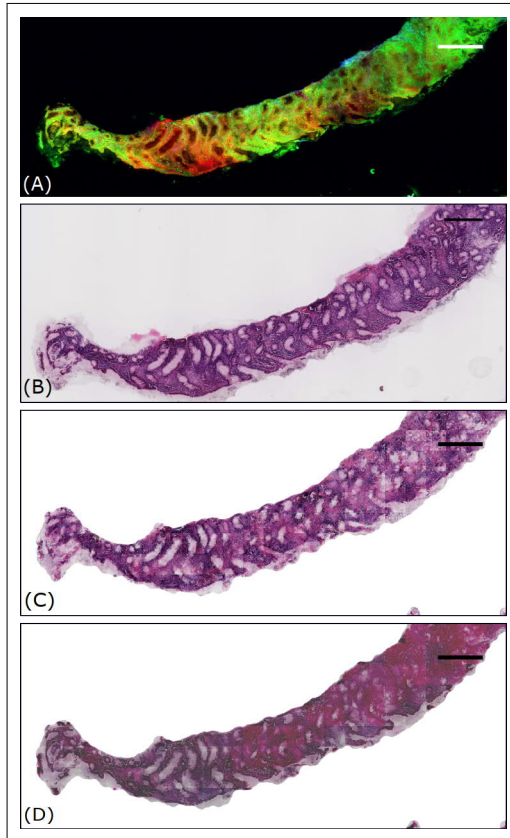


Fig. S3. (A) shows a non-linear multimodal image from training dataset, (B) visualizes corresponding histopathologically stained H&E image (unregistered), (C) shows the computational H&E image by the Pix2Pix ($MSE = 2.8 \times 10^3$, $SSIM = 0.74$, $CSS = 0.96$) and (D) depicts computational H&E image by the cycle CGAN model ($MSE = 5.9 \times 10^3$, $SSIM = 0.72$, $CSS = 0.94$). The contrast of the computational H&E image in (D) is reduced by a factor of 0.7. The images here are downsampled to 20% of original size for clarity. The scale bar represents $100 \mu m$.

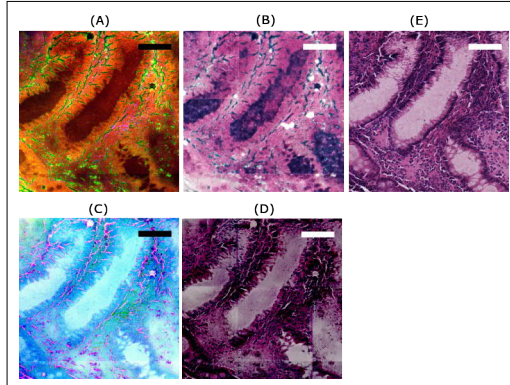


Fig. S4. This figure shows effect of ‘contrast inversion’ on the computational H&E images of the cycle CGAN model. (A) shows an original MM patch given to the cycle CGAN model and (B) visualizes the generated H&E patch. Here, we see that the colors in the generated H&E image are inverted when compared to the pathologically stained H&E patch shown in (E). Contrarily, (C) shows a contrast inverted MM patch and (D) depicts the generated H&E patch by the cycle CGAN model. Here, the color contrast is similar to the pathologically stained H&E patch shown in (E). The scale bar represents 100 μm .

Table S1. The quantitative metrics namely mean squared error (MSE), structure similarity index (SSIM) and color similarity index (CSS) evaluated for 19 images from the training and testing dataset is given for the Pix2Pix and the cycle CGAN models, respectively.

Image	Pix2Pix			Cycle CGAN		
	MSE	SSIM	CSS	MSE	SSIM	CSS
Train 1	3601.43	0.61	0.94	9373.09	0.57	0.92
Train 2	2800.83	0.74	0.96	5890.62	0.72	0.94
Train 3	4984.19	0.38	0.89	13955.34	0.33	0.89
Train 4	3337.39	0.56	0.93	8157.08	0.53	0.93
Train 5	4451.48	0.51	0.92	11509.74	0.47	0.90
Train 6	5100.78	0.49	0.92	9629.29	0.46	0.89
Train 7	4956.57	0.41	0.91	10389.17	0.38	0.90
Train 8	4167.94	0.52	0.92	8376.58	0.49	0.90
Train 9	7938.24	0.46	0.91	15823.56	0.41	0.91
Train 10	7657.41	0.39	0.88	9725.03	0.36	0.88
Train 11	3066.22	0.67	0.94	6749.37	0.65	0.93
Train 12	3057.30	0.65	0.94	8129.79	0.62	0.93
Train 13	5918.03	0.36	0.90	15783.79	0.30	0.90
Test 1	5617.82	0.60	0.92	6600.03	0.59	0.91
Test 2	3582.64	0.61	0.94	6633.59	0.60	0.94
Test 3	3827.09	0.65	0.94	4767.95	0.63	0.94
Test 4	3292.67	0.60	0.94	10118.23	0.58	0.93
Test 5	4416.22	0.64	0.94	8408.46	0.63	0.94
Test 6	4937.68	0.49	0.91	10265.39	0.48	0.93





P6 DATA FUSION OF HISTOLOGICAL AND IMMUNOHISTOCHEMICAL IMAGE DATA FOR BREAST CANCER DIAGNOSTICS USING TRANSFER LEARNING

Reprinted with permission from [P. Pradhan, K. Köhler, S. Guo, O. Rosin, J. Popp, A. Niendorf and T. Bocklitz, Data fusion of histological and immunohistochemical image data for breast cancer diagnostics using Transfer learning, *Accepted as conference proceeding for 10th International Conference on Pattern Recognition Applications and Methods*, 2021]. Copyright 2021 by SCITEPRESS – Science and Technology Publications, Lda.

The declared individual contributions of the doctoral candidate and the other doctoral candidates participate as co-authors in the publications are listed below.

P. Pradhan ¹ , K. Köhler ² , S. Guo ³ , O. Rosin ⁴ , J. Popp ⁵ , A. Niendorf ⁶ , T. Bocklitz ⁷ , Data fusion of histological and immunohistochemical image data for breast cancer diagnostics using Transfer learning, <i>Accepted as conference proceeding for 10th International Conference on Pattern Recognition Applications and Methods</i> , 2021.							
Involved in (Please tick the boxes that apply.)							
	1	2	3	4	5	6	7
Conceptual research design					X	X	X
Planning of research activities	X	X	X			X	X
Data collection		X		X			
Data analysis and interpretation	X	X	X			X	X
Manuscript writing	X	X	X	X	X	X	X
Suggested publication equivalence value	1.0						

Data fusion of histological and immunohistochemical image data for breast cancer diagnostics using transfer learning

Pranita Pradhan^{1,2}^a, Katharina Köhler^{3,4}, Shuxia Guo^{1,2}^b, Olga Rosin^{3,4}, Jürgen Popp^{1,2}^c, Axel Niendorf^{3,4} and Thomas Wilhelm Bocklitz^{*,1,2}^d

¹*Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University, Helmholtzweg 4, Jena, 07743, Thüringen, Germany*

²*Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, Jena, 07745, Thüringen, Germany*

³*MVZ Prof. Dr. med. A. Niendorf Pathologie Hamburg-West GmbH, Lornsenstraße 4-6, Hamburg, 22767, Hamburg, Germany*

⁴*Institute for Histology, Cytology and Molecular Diagnostics, Lornsenstraße 4, Hamburg, 22767, Hamburg, Germany*
**thomas.bocklitz@uni-jena.de*


Keywords: Breast cancer, transfer learning, histology, immunohistochemistry


Abstract: A combination of histological and immunohistochemical tissue features can offer better breast cancer diagnosis as compared to histological tissue features alone. However, manual identification of histological and immunohistochemical tissue features for cancerous and healthy tissue requires an enormous human effort which delays the breast cancer diagnosis. In this paper, breast cancer detection using the fusion of histological (H&E) and immunohistochemical (PR, ER, Her2 and Ki-67) imaging data based on deep convolutional neural networks (DCNN) was performed. DCNNs, including the VGG network, the residual network and the inception network were comparatively studied. The three DCNNs were trained using two transfer learning strategies. In transfer learning strategy 1, a pre-trained DCNN was used to extract features from the images of five stain types. In transfer learning strategy 2, the images of the five stain types were used as inputs to a pre-trained multi-input DCNN, and the last layer of the multi-input DCNN was optimized. The results showed that data fusion of H&E and IHC imaging data could increase the mean sensitivity at least by 2% depending on the DCNN model and the transfer learning strategy. Specifically, the pre-trained inception and residual networks with transfer learning strategy 1 achieved the best breast cancer detection.


1 INTRODUCTION


Breast cancer is one of the most prevalent cancers among women. It is diagnosed by a routine procedure which is based on morphological tissue features in hematoxylin and eosin (H&E) stained tissue sections (figure 1a). The morphological tissue features include tumour size and type, which are regularly documented to assess the histological grade of breast cancer tissue (Webster et al., 2005). These morphological tissue features are also used to prevent recurrence risk of breast cancer and prescribe personalized therapies. Breast cancer is additionally verified by other staining technique called the immunohistochemical

(IHC) staining technique. The IHC staining technique uses antibodies to highlight specific antigens in the tissue region (Veta et al., 2014), and includes estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor-2 (Her2) (figure 1b-d). Studies have shown that the IHC examination with ER, PR, Her2 and Ki-67 can detect five molecular breast cancer sub-types to provide adequate personalized therapies (Perou et al., 2000; Sørliet et al., 2001; Cheang et al., 2009). However, none of the studies report a combination of histology (H&E) and IHC staining techniques (ER, PR, Her2 and Ki-67) for breast cancer diagnosis. Therefore, in this work, an integration of IHC imaging technique i.e. hormone receptors including ER, PR, Her2 and Ki-67 nuclear protein stained images with H&E stained images is proposed to gain new insights into breast cancer biology (Elledge et al., 2000; Damodaran and Olson, 2012). The combination of histology and IHC stain-

^a <https://orcid.org/0000-0002-0558-2914>

^b <https://orcid.org/0000-0001-8237-8936>

^c <https://orcid.org/0000-0003-4257-593X>

^d <https://orcid.org/0000-0003-2778-6624>

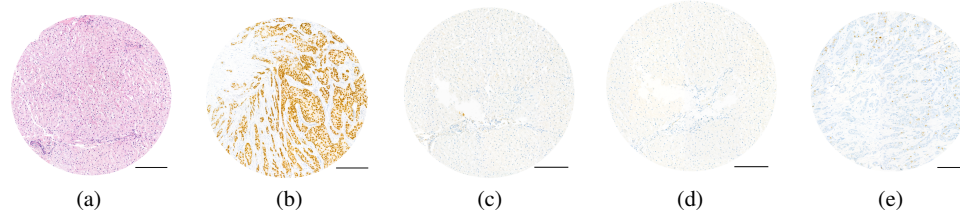


Figure 1: Five stain type images. (a) Hematoxylin and eosin (H&E), (b) Estrogen receptor (ER), (c) Progesterone receptor (PR), (d) Human epidermal growth factor-2 (Her2) and (e) Ki-67 protein are shown. Scale bar is 200 μm .

ing technique is referred to as ‘Data fusion’ approach.

Data fusion approach by combining the histological and IHC stained images can provide various tissue features associated with the disease stage and relapse of breast cancer. However, visual inspection of all five stained images is a tedious process which can prolong the diagnosis. Therefore, automation of breast cancer detection based on the combination of histological and IHC imaging data is needed. In this regard, researchers (Pham et al., 2007; Dobson et al., 2010) used computer-assisted image analysis techniques to automatically monitor changes in the tissue features of histological and IHC stained images separately. However, computer-assisted image analysis can be limited due to the need for specific software systems or the need for user-specific input to analyze the images. This slows down the process of analyzing images and providing personalized therapies to the patients. To increase the analysis speed and reduce human intervention, this work proposes machine learning (ML) instead of computer-assisted image analysis techniques.

Conventional ML methods can automatize breast cancer detection based on the fusion of histological and IHC imaging data in the following way. First, the features (e.g. color, shape and texture features) from the five stain type of imaging data (H&E, ER, PR, Her2 and Ki-67) can be extracted using image analysis methods. The feature extraction step in the conventional ML method is subjective and requires the effort of an image analyst. Based on the extracted features, a classification, or a regression model can be constructed. Subsequently, the classification or the regression model can be used to make ‘predictions’ (i.e. to predict a class like tumour or normal) on a new or unseen dataset. Thus, the extracted features affect the predictions made by the ML model. However, recently developed ML methods are capable of performing automatic feature extraction for classification or regression purpose. These self-learning methods are categorized into a broad family of ML called ‘Deep learning’ (DL). The DL models can have

many types of network architectures. Widely used DL model for images is the deep convolutional neural network (DCNN) and its numerous applications are reported in the field of digital pathology (Liu et al., 2017); for example, cell segmentation or detection (Chen and Chef’ Hotel, 2014), tumour classification (Cireřan et al., 2013; Wang et al., 2016) and carcinoma localization (Janowczyk and Madabhushi, 2016; Coudray et al., 2018; Khosravi et al., 2018; Sheikhzadeh et al., 2018). Nevertheless, a bottleneck for DL models is the requirement of huge dataset during training, which is difficult to acquire, particularly in the medical imaging field. In such cases, ‘transfer learning’ methods for DCNNs can be applied for improving the model performance (Tajbakhsh et al., 2016).

Transfer learning is the transfer of knowledge learned on a source task using a source dataset to improve the performance on a target task using the target dataset (Torrey and Shavlik, 2010). Transfer learning using any DL model like DCNN can be performed by three strategies. First, a pre-trained DCNN can be used as a feature extractor. In this strategy, features for the target dataset are extracted using a DCNN trained on different or similar source dataset. The second strategy is fine-tuning the weights of the last layers of a pre-trained DCNN, and the third strategy is fine-tuning the weights of all layers of a pre-trained DCNN. In the second and third fine-tuning strategies, the weights of specific layers of a DCNN trained on a source dataset are further optimized based on the target dataset. The three transfer learning strategies like using a DCNN as a feature extractor or fine-tuning of a DCNN, requires adequate knowledge of the size and type of the source and the target dataset (Pan and Yang, 2010). Transfer learning, if used appropriately, can improve the initial and final performance of the DL model on the target dataset. It can also reduce the total training time of the DL model on the target dataset. Different transfer learning strategies acquire different results based on the source and target dataset which is evident in the next section.

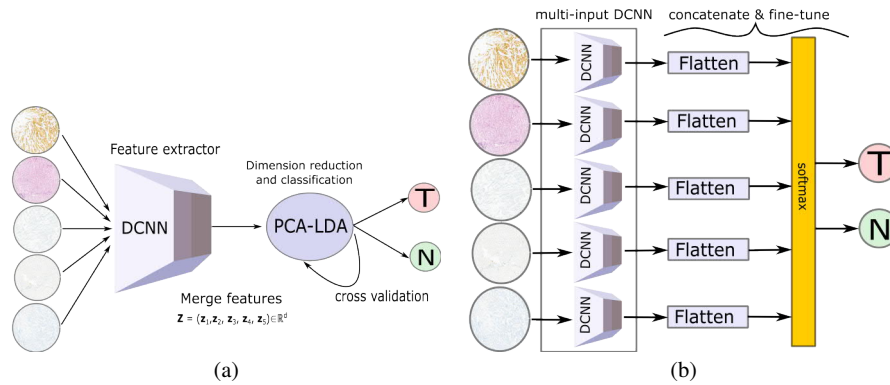


Figure 2: (a) visualizes transfer learning strategy 1 for data fusion approach where a pre-trained DCNN is used as a feature extractor. The features extracted from a pre-trained DCNN for all five stain type images are merged and classified into tumour and normal using the PCA-LDA model. (b) shows transfer learning strategy 2 for data fusion approach where fine-tuning of the last layer of a pre-trained multi-input DCNN is performed. The five DCNNs are pre-trained models like the VGG16, the Inceptionv3 or the ResNet50, each having a stain type as its input.

2 RELATED WORK

Transfer learning in medical imaging can be achieved by training a DCNN on a large medical or non-medical dataset, and transferring its knowledge to the target medical dataset (Bayramoglu and Heikkilä, 2016; Tajbakhsh et al., 2016). A recent study used a large non-medical dataset like the ImageNet dataset (Russakovsky et al., 2015) to pre-train a DCNN and transfer its off-the-shelf features to investigate two computer-aided detection (CADs) problems namely thoracoabdominal lymph node detection and interstitial lung disease detection (Shin et al., 2016). In their work, three different DCNNs including the CifarNet (Krizhevsky and Hinton, 2009), the AlexNet (Krizhevsky et al., 2012) and the GoogleNet (Szegedy et al., 2015) were evaluated with three transfer learning strategies. Similarly, a recent publication (Mormont et al., 2018) compared various transfer learning strategies based on pre-trained DCNNs using eight classification datasets in digital pathology. Their results showed that fine-tuning the ResNet (He et al., 2016) and the DenseNet (Huang et al., 2017) models outperformed the other tested models in the morphological classification task. Similar findings were observed in other references (Antony et al., 2016; Kieffer et al., 2017; Ravishankar et al., 2016).

In contrast to the previously mentioned applications where fine-tuning of a DCNN achieved the best performance, several other applications using a DCNN as feature extractor achieved significant performance on binary and multi-class classification tasks. These applications included prediction of morphological changes in cells in microscopic images

(Kensert et al., 2018), classification of colon polyps in endoscopic images (Ribeiro et al., 2016), identification of mammographic tumours (Huynh et al., 2016) and detection of pulmonary nodules in computed tomography scans (Van Ginneken et al., 2015). It is clear from the previous researches that transfer learning techniques are data-dependent, and a generalization of the above-mentioned results is not feasible, especially in the medical imaging field (Litjens et al., 2017). Therefore, no consensus of the proper application of transfer learning in the medical imaging field is established. Likewise, the application of transfer learning, especially for medical imaging data requires utmost care and further investigations.

In this contribution, data fusion of histological and immunohistochemical imaging data for classifying breast cancer is presented for the first time. Due to our small dataset size, the classification task is performed using two transfer learning strategies. From previous experience, the third transfer learning strategy i.e. the training of a DCNN from scratch is avoided, as it is computationally expensive and may lead to overfitting in the absence of large datasets. The performance of the two transfer learning strategies for the data fusion approach is compared with histological imaging data. Moreover, the two transfer learning strategies are performed using three pre-trained DCNN models like the VGG16 (He et al., 2016), the Inceptionv3 (Szegedy et al., 2016) and the ResNet50 network (Simonyan and Zisserman, 2014). The goal of this study was to verify whether the data fusion approach along with transfer learning improves the breast cancer diagnosis based on the sensitivity and F1 score metric.

3 MATERIAL AND METHODS

3.1 Sample Preparation

A Tissue Microarray (TMA) with 97 cores representing 23 breast cancer cases (78 tumour cores, 18 non-cancerous tissue cores or the normal breast tissue and one control core of liver tissue) was produced using the Manual Tissue Arrayer MTA-1 by Estigen. The cases were randomly selected out of the daily routine of MVZ Prof. Dr. med. A. Niendorf Pathologie Hamburg-West GmbH and anonymized according to a statement of the ethics committee of the Hamburg Medical Chamber. Core tissue biopsies (1.0 mm in diameter) were taken from individual FFPE (formalin-fixed paraffin-embedded) blocks and arranged within a new recipient block. From the block, 2 μm sections were cut, placed on glass microscope slides and H&E staining (figure 1a) following a standard protocol was performed. Digital images of histology (H&E) slides were obtained at 40 \times magnification using the 3DHistech Panoramic 1000 Flash IV slide scanner with a spatial resolution of 0.24 $\mu\text{m}/\text{pixel}$ (.mrxs image file). Subsequently, immunohistochemistry staining (ER, PR, Her2 and Ki-67) (figure 1b-e) was performed on super frost charged glass slides.

3.2 Image preprocessing

For the analysis, 96 TMAs or scans (78 tumour scans and 18 normal scans) from 23 patients were used, and each TMA had five stain types (H&E, PR, ER, Her2 and Ki-67). The pixel intensity I of each TMA was standardized using a min-max scaling $(I - I_{min}) / (I_{max} - I_{min})$, where I_{min} and I_{max} is the minimum and maximum intensity of a pixel in a TMA. The background pixels were cropped manually and non-overlapping patches of size 1024 \times 1024 were extracted from a standardized TMA. This led to 9 patches per TMA (702 tumour and 162 normal patches). The four corner patches including a large number of background pixels were removed, leading to 390 tumour and 90 normal patches. Based on the 480 selected patches, three pre-trained models were used with two transfer learning strategies.

3.3 DCNN architectures

To check the robustness of the data fusion approach, three DCNNs: the VGG network, the Inception network and the residual network, with unique architectures were chosen. The VGG network is a DCNN that has acquired state-of-the-art performances for image classification tasks. However, the VGG network can

exhibit the problem of vanishing gradients with an increasing number of layers (Hanin, 2018). Thus, the residual network which can solve the problem of vanishing gradients by adding the 'shortcut connections' was explored in this work. Furthermore, the inception network that provides width in addition to the depth to a conventional DCNN was utilized. A detailed explanation of the architecture of the three models is given further.

3.3.1 VGG network

A VGG network is a DCNN with different configurations from 11 to 16 convolutional layers followed by three fully connected layers. The number of convolutional layers increases the depth of the VGG network. It is shown that an increase in the depth of the VGG network decreases the top-5 validation error (He et al., 2016). However, the decrease in the error for the VGG network from 16 to 19 convolutional layers is not significant. Thus, the VGG network with 16 convolutional layers referred to as VGG16 from Keras was used (Chollet et al., 2015). The input to the VGG16 network was an RGB image of size 224 \times 224, and each image was preprocessed by subtracting the mean RGB values computed over the training dataset.

3.3.2 Inception network

Deep networks like VGG network require an appropriate selection of the number of convolution filters and filter sizes. For this reason, the inception network concatenates convolutional layers of different filter size, including the spatial dimension of 1 \times 1, 3 \times 3 and 5 \times 5. This captures information at various scales while increasing the computational complexity. In order to reduce the computational cost, a convolutional layer of 1 \times 1 filter size is applied before each convolutional layer of filter size 3 \times 3 and 5 \times 5. These two salient features of the Inception network reduce the dimensionality in the feature space and thereby allows the network to be deeper and wider. Moreover, the inception network replaces the fully connected layer with global averaging layers which reduces the number of trainable weights, thus reducing over-fitting on the training dataset (Szegedy et al., 2016). The Inceptionv3 implementation from Keras, which has 95 layers and requires an RGB image as input with size 299 \times 299 was used.

3.3.3 Residual network

The configurations of the VGG network show that deep neural networks achieve good top-5 accuracy

Table 1: This table shows confusion matrices, mean sensitivities and mean F1 scores for the VGG, the Inception and the residual networks using transfer learning strategy 1. Here, two feature sets extracted from pre-trained models are used; one feature set is extracted from H&E images only, while the other feature set is extracted from all the five stain types. All metrics are computed for 96 TMAs by taking majority voting of the predictions acquired for the patches using the PCA-LDA model. N: normal scans, T: tumour scans.

Data fusion (H&E+IHC imaging data)					Only histological imaging data						
DCNN		N	T	Sens (%)	F1 (%)	DCNN		N	T	Sens (%)	F1 (%)
VGG16	N	13	5	79.06	76.24	VGG16	N	14	4	80.56	76.61
	T	11	67				T	13	65		
Inceptionv3	N	16	2	89.32	85.47	Inceptionv3	N	15	3	88.46	86.97
	T	8	70				T	5	75		
ResNet50	N	14	4	86.97	87.80	ResNet50	N	14	4	85.68	84.96
	T	3	75				T	5	73		

Table 2: This table shows confusion matrices, mean sensitivities and mean F1 scores for the VGG, the Inception and the residual networks using transfer learning strategy 2. Data fusion approach used multi-input DCNN with the five stain type images as input, whereas a single-input DCNN was used only the H&E image as input. The last layers of both single-input and multi-input DCNNs were fine-tuned. The mean sensitivities are computed for 96 TMAs by taking majority voting of the predictions obtained for the patches. N: normal scans, T: tumour scans.

Data fusion (H&E+IHC imaging data)					Only histological imaging data						
DCNN		N	T	Sens (%)	F1 (%)	DCNN		N	T	Sens (%)	F1 (%)
VGG16	N	7	11	66.88	70.86	VGG16	N	3	15	55.13	57.57
	T	4	74				T	5	73		
Inceptionv3	N	0	18	50.00	44.83	Inceptionv3	N	9	9	72.44	75.66
	T	0	78				T	4	74		
ResNet50	N	0	18	50.00	44.83	ResNet50	N	12	6	81.41	83.78
	T	0	78				T	3	75		

Table 3: This table shows confusion matrices, mean sensitivities and mean F1 scores for the VGG, the Inception and the residual network using the two transfer learning strategies. All metrics are computed for 96 TMAs by taking majority voting of the predictions acquired by the models for patches.

Transfer learning strategy 1					Transfer learning strategy 2						
DCNN		N	T	Sens (%)	F1 (%)	DCNN		N	T	Sens (%)	F1 (%)
VGG16	N	13	5	79.06	76.24	VGG16	N	7	11	66.88	70.86
	T	11	67				T	4	74		
Inceptionv3	N	16	2	89.32	85.47	Inceptionv3	N	0	18	50.00	44.83
	T	8	70				T	0	78		
ResNet50	N	14	4	86.97	87.80	ResNet50	N	0	18	50.00	44.83
	T	3	75				T	0	78		

until a certain depth limit (He et al., 2016). An increase in the network depth causes a problem of vanishing or exploding gradients (Hanin, 2018) which affects the network convergence and degrades the performance (Simonyan and Zisserman, 2014). Therefore, the residual networks are built to solve this degradation problem by adding activations of the top layers into the deeper layers of the network. For instance, in a deep neural network the activation a of the $(l+2)^{th}$ layer with weight w and bias b is given as

$$a_{(l+2)} = f[(w_{(l+2)} \times a_{(l+1)}) + b_{(l+2)}], \quad (1)$$

where f is an activation function like linear rectified unit ($f = \max(a_{(l+2)}, 0)$). However, in a residual block the activation a of the l^{th} layer (or an identity mapping) is added via the ‘skip or shortcut connections’ (Bishop et al., 1995; Venables and Ripley, 2013) to the $(l+2)^{th}$ layer of the network. Therefore, the activation of the $(l+2)^{th}$ layer in a residual block can be given as

$$a_{(l+2)} = f[(w_{(l+2)} \times a_{(l+1)}) + b_{(l+2)} + a_{(l)}]. \quad (2)$$

This implies that in worse cases when the network fails to learn representative features, i.e. $w_{(l+2)} = 0$

and $b_{(l+2)} = 0$, the output still remains an identity mapping of the input a_l . In residual networks, a series of residual blocks along with intermediate normalization layers was used; thus improving the learning of the deep neural networks. In this work, the ResNet50 implementation from Keras, which has 152 layers and requires an RGB image as an input with size 224×224 , was used.

The above explained three DCNN models were trained using two transfer learning strategies which are discussed in the next section.

3.4 Transfer learning strategies

The above-mentioned DCNNs were utilized for two transfer learning strategies. For the first strategy, a pre-trained DCNN model to extract off-the-shelf features followed by a linear classifier was used. In the second strategy, a multi-input pre-trained DCNN model followed by a softmax classifier was used. Both strategies were performed on a commercially available PC system intel® Core™ with NVIDIA GeForce GTX 1060, 6GB with python packages: Keras(Chollet et al., 2015), Tensorflow(Abadi et al., 2015), Scikit-learn (Pedregosa et al., 2011), Scipy (Jones et al., 2001) and Numpy (Oliphant, 2006).

3.4.1 DCNN as feature extractor

In the first strategy (figure 2a), features $\mathbf{z}_i \in \mathbb{R}^m, i = (1, 2, 3, 4, 5)$ were extracted for patches of each stain type i using the pre-trained VGG16, Inceptionv3 and ResNet50 networks. The patches were resized according to the model's input size requirement. For a patch of a single stain type, 25,088 features were extracted by the VGG16 (feature shape: 1, 7, 7, 512), 51,200 features were calculated by the Inceptionv3 (feature shape: 1, 5, 5, 2048) and 2048 features were obtained by the ResNet50 (feature shape: 1, 1, 1, 2048). For data fusion approach, the features from all five stain types were concatenated, $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4, \mathbf{z}_5) \in \mathbb{R}^d (d \gg m)$ resulting in ~ 0.12 million features by the VGG16 model, ~ 0.25 million features by the Inceptionv3 model and 10,240 features by the ResNet50 model per patch. For histological imaging data, i.e. without the data fusion approach, the features extracted only from the H&E images were used. In both cases, the large feature dimension of each patch was reduced by principal component analysis (PCA) model, and classified as normal or tumour using linear discriminant analysis (LDA) model (Hastie et al., 2009). The PCA-LDA model was evaluated using internal and external cross-validation scheme explained elsewhere (Guo et al., 2017). Shortly, the internal cross-validation

was used to optimize the number of PC's of the PCA-LDA model. The external cross-validation was used to predict an independent test dataset based on the PCA-LDA model. The external cross-validation used leave-one-patient-out cross-validation, such that the patches acquired from TMAs of 23 patients were used at least once as an independent test dataset. The internal cross-validation used 10 fold cross-validation. The predictions by the PCA-LDA model acquired for the patches from the external cross-validation step were voted to assign each TMA into a tumour or normal class. Based on the predicted TMA labels (obtained after majority voting of the patches) and true TMA labels, metrics like confusion matrix, mean sensitivity and mean F1 score were reported. The mean sensitivity and the mean F1 score were calculated using an average of the mean sensitivities and the mean F1 scores for the tumour and normal class, respectively. Lastly, the transfer learning strategy 1 was performed for all the three DCNNs and their classification performance based on TMAs was compared.

3.4.2 Fine-tuning of DCNN

In the second strategy (figure 2b), for histological imaging data, a single-input DCNN was used; whereas for the data fusion approach, a multi-input DCNN was used. The multi-input DCNN model \mathcal{N} was constructed using five pre-trained models of the same architecture; for instance, five pre-trained VGG networks each using a stain type image as an input. The input to the multi-input DCNN model was the five stained images (H&E, ER, Her2, Ki-67 and PR). The last layer of the multi-input DCNN models was concatenated and followed by a dense layer with two outputs (corresponding to the normal and tumour class) with a softmax activation layer. The softmax activation layer mapped the non-normalized output of the model \mathcal{N} to the distribution of K probabilities and is defined as

$$P(\mathbf{r})_i = \frac{\exp(r_i)}{\sum_{j=1}^K \exp(r_j)}, \quad (3)$$

where $\mathbf{r} = (r_1, \dots, r_K)$ and $K = 2$ for a binary classification task. During the training process, the last two layers were fine-tuned using Adam optimizer (Kingma and Ba, 2014) with a learning rate 0.001 and mini-batch size of 5 patches. To allocate higher class weight for the minority class (here, the normal class), the weighted binary cross-entropy loss function

$$\mathcal{L} = - \sum_i^K \alpha_i y_i \log(P(\mathbf{r})_i) \quad (4)$$

was used, where $\alpha_i = \frac{1}{\#K_i}$, y_i , $P(\mathbf{r})_i$ are the weight, ground truth and the probability from the softmax ac-

tivation layer of the l^{th} class in K , respectively. The model was evaluated using the mean sensitivity and the mean F1 score similar to transfer learning strategy 1.

For the evaluation of the single and multi-input DCNN, the dataset was divided into three parts: training, validation and testing. In every iteration, patches of one patient were used as an independent test dataset and the patches of remaining patients were used as training and validation dataset. To avoid any training bias, the training and validation datasets were randomly split patient-wise such that patches from 30% patients were used as validation dataset and the rest as the training dataset. In other words, during each iteration, patches of one patient were used as the test dataset, patches of 16 patients formed the training dataset and patches of remaining 6 patients belonged to the validation dataset. The combination of 16 and 6 patients in training and validation datasets were chosen randomly. The iterations were repeated until all 23 patients were used as an independent test dataset. Further, every iteration was executed for ten epochs, and validation sensitivity was monitored for early stopping of the model training. The model with best validation sensitivity was used for predicting the independent test dataset in that iteration. In this way, the patches of all 23 patients were used individually as an independent test dataset, and majority voting of the patches similar to transfer learning strategy 1 was performed. The confusion matrices and average of the mean sensitivities for the normal and tumour classes were evaluated using the independent test dataset. Subsequently, transfer learning strategy 2 was performed for all the three pre-trained DCNN models with the same hyper-parameter setting.

3.4.3 ROC curve analysis for TMAs

The results of the two transfer learning strategies were obtained as ROC curves showing the true and the false positive rate for the tumour class. The ROC curves were evaluated for TMAs based on the majority voting of the selected patches. To achieve ROC curves for TMAs, the model output in the form of probabilities of each patch for the tumour class was thresholded using 100 different values in the range [0, 1]. This led to predictions for patches with different threshold values. Subsequently, the predictions for patches obtained for each threshold value were majority voted to obtain a prediction for a TMA. The predictions for TMAs were used to calculate the true positive rate, the false positive rate and the ROC curve, as shown in figure 3 and 4. The predictions for the TMAs obtained with 0.5 threshold were used to obtain the confusion matrix, mean sensitivities and

mean F1 scores as reported in table 1, 2 and 3.

4 RESULTS

The main aim of this work was to confirm that the data fusion approach can achieve better breast cancer diagnosis than histological imaging data based on performance metrics. This was confirmed by one of the two transfer learning strategies. The results are divided in three parts as shown in table 1, 2 and 3. Table 1 and 2 report performance metrics obtained for transfer learning strategy 1 and transfer learning strategy 2, with and without data fusion approach, respectively. Table 3 shows a comparison of the two transfer learning strategies using only the data fusion approach. In table 1, 2 and 3 report values for the VGG16, the Inceptionv3 and the ResNet50 models. These values were evaluated for 96 TMAs acquired by majority voting of the five patches extracted from each TMA.

The results in table 1 show that the pre-trained features acquired from the data fusion approach yield slightly higher mean sensitivities and mean F1 scores in comparison to the pre-trained features extracted from the histological imaging data. Higher mean sensitivities using the data fusion approach were seen for at least two of the three DCNNs. Higher mean F1 score using the data fusion approach was seen only for the ResNet50 model. Specifically, the pre-trained features obtained from the data fusion approach using the Inceptionv3 and the ResNet50 models showed mean sensitivities 89.32% and 86.97%, respectively. Similarly, the mean F1 scores for the two models were 85.47% and 87.80%, respectively. In comparison, the pre-trained features from the histological data using the same DCNN model showed mean sensitivities 88.46% and 85.68%, respectively. Thus, there was approximately 2% increase in the model performance by data fusion approach based on the mean sensitivity, which is significant from a clinical perspective. However, the VGG16 model showed higher mean sensitivity (80.56%) using histological imaging data compared to the mean sensitivity calculated for the data fusion approach (79.06%). Overall, it can be seen that transfer learning using pre-trained DCNN features and a linear classification model (PCA-LDA)

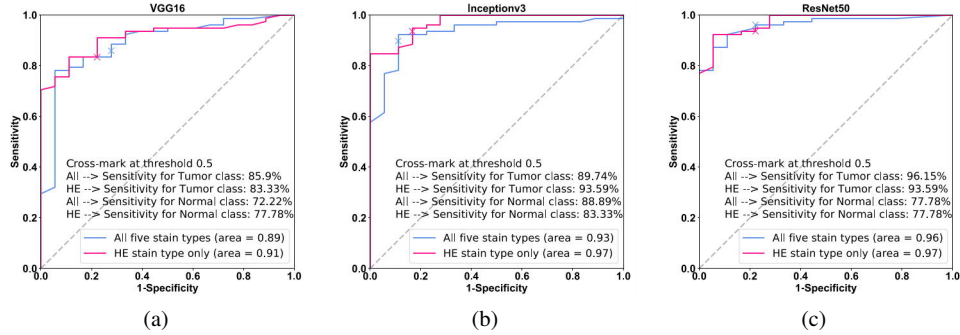


Figure 3: (a-c) show ROC curves for the VGG16, the Inceptionv3 and the ResNet50 networks using the transfer learning strategy 1 based on TMAs. The blue line shows ROC curve for the PCA-LDA model trained using the pre-trained DCNN features obtained from the histological and IHC imaging data, whereas the pink line shows ROC curve for the PCA-LDA model trained using pre-trained DCNN features extracted from the histological imaging data only. The cross-mark shows the true and the false positive rate on the ROC curve with 0.5 threshold.

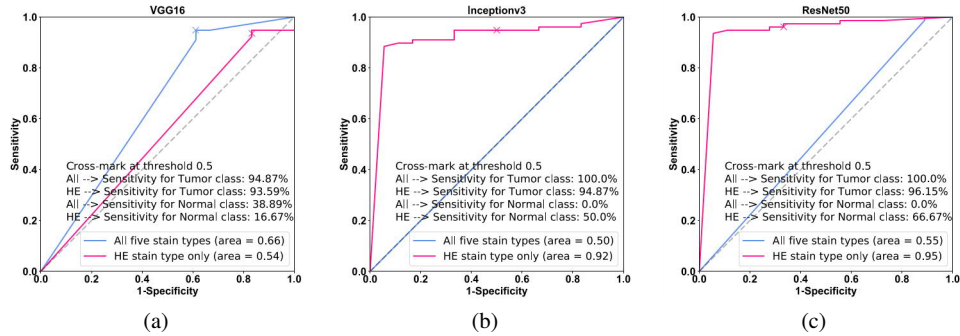


Figure 4: (a-c) show ROC curves for the VGG16, the Inceptionv3 and the ResNet50 networks using the transfer learning strategy 2 based on TMAs. The blue line shows ROC curve for the multi-input DCNN model fine-tuned using the histological and IHC imaging data, whereas the pink line shows the ROC curve for the single-input DCNN model fine-tuned using only the histological imaging data. The cross-mark shows the true and the false positive rate at 0.5 threshold.

based on data fusion approach show a slight improvement in breast cancer detection in some cases for a small dataset as in our study.

Contrarily, table 2 obtained by the transfer learning strategy 2 shows lower mean sensitivities for the data fusion approach in comparison to the performance achieved by using histological imaging data alone. Except for the multi-input VGG16 network, the multi-input Inceptionv3 and the multi-input ResNet50 network trained with a combination of histological and IHC imaging data predicted all normal patches as tumour patches. Thus, the multi-input Inceptionv3 and the multi-input ResNet50 model achieved mean sensitivity of 50% and mean F1 score of 44.83%; whereas, the multi-input VGG16 network showed mean sensitivity of 66.88% and mean F1 score of 70.86% for the data fusion approach. The mean sensitivity of the single-input VGG16 network

declined to 55.13% when only histological imaging data was used. On the other hand, the single-input Inceptionv3 and the single-input ResNet50 models using histological imaging data showed an opposite trend with comparatively higher mean sensitivities of 72.44% and 81.41%, and higher mean F1 scores of 75.66% and 83.78%, respectively. Overall, it was observed that transfer learning performed by fine-tuning the last layer of the pre-trained multi-input DCNNs result in lower mean sensitivities for the data fusion approach. This behaviour can be a consequence of the small sample size. It is clear from the results that fine-tuning the last layer of DCNNs is not the best approach for our small breast cancer dataset. Thus, it is suspected that the fine-tuning of all the layers of a DCNN will decrease the model performance further. However, fine-tuning of all layers for large breast cancer dataset should be investigated in the future.

Lastly, the performance of the two transfer learning strategies for the data fusion approach is summarized in table 3, where higher mean sensitivities are reported for strategy 1, i.e. using pre-trained features from the VGG16, the Inceptionv3 and the ResNet50 model. The training of the PCA-LDA model based on pre-trained features of the Inceptionv3 and the ResNet50 network yield promising results. The results from the VGG16 network are lower in comparison to the other two models for transfer learning strategy 1, but higher for transfer learning strategy 2.

The performance of the two transfer learning strategies based on TMAs is summarized in the form of ROC curves in figure 3 and 4. The ROC curve calculated for the data fusion approach and histological imaging data at various thresholds is depicted in blue and pink, respectively. The AUC values given in the figure legend show lower values for the data fusion approach in comparison to the AUC values calculated using histological imaging data. This trend is observed for both the transfer learning strategies. From figure 3 and 4, it can be inferred that the overall performance of DCNN models trained using an H&E image is better for both transfer learning strategies. However, the final performance of the models in terms of mean sensitivities evaluated at 0.50 threshold is better for the data fusion approach in some cases. The mean sensitivities cross-marked in each subplot of figure 3 and 4 are calculated at 0.50 threshold coincide with the values reported in table 1, 2 and 3. These values are evaluated for TMA's by performing majority voting of the five patches in each TMA. The ROC curves at threshold 0.50 which is mostly used to evaluate the model performance, show higher mean sensitivities for data fusion approach than using histological data, at least for the Inceptionv3 and the ResNet50 model in transfer learning strategy 1 (figure 3). Nevertheless, the AUC derived from the ROC curves for transfer learning strategy 2 (figure 4) show low mean sensitivities for all the DCNN networks. The inconsistency in the results of two transfer learning strategies can be due to various reasons discussed below.

5 DISCUSSION

Based on the results, three critical findings can be discussed.

5.1 Data fusion vs. histological imaging

The results showed that the data fusion approach, i.e. combining histological and IHC imaging data, increases the model performance by $\sim 2\%$. However,

the increase in model performance was achieved only for transfer learning strategy 1, where features were extracted from a pre-trained DCNN followed by binary classification using the PCA-LDA model. It is important to mention that the analysis was performed on a limited number of TMAs and it is suspected that the results can improve with an increasing number of TMAs, at least for the transfer learning strategy 1. Furthermore, the data fusion approach can largely increase the feature dimension of the data, thus increasing computational complexity. Nevertheless, these limitations are the cost of performing reliable and early breast cancer diagnosis. In future studies, feature dimension can be reduced by extracting features from the last layers and a comparative study can be performed.

5.2 Strategy 1 vs. Strategy 2

From the results shown in table 3 it is clear that transfer learning strategy 1 outperforms the transfer learning strategy 2 for our breast cancer dataset. For transfer learning strategy 2, the misclassification of the under-represented normal class as tumour class is higher. This means that transfer learning strategy 2 performed by merging and fine-tuning the last layer of the pre-trained multi-input model causes 'negative transfer learning' showing lower binary classification performance. Although the past studies (Kensert et al., 2018; Mormont et al., 2018) have shown that transfer learning strategy 2 for medical imaging data can provide good classification performances, these studies used a single-input DCNN for fine-tuning; whereas, in this study a multi-input DCNN was used. Thus, training a large multi-input network on a small dataset can cause the model to overfit and degrade its performance. Degradation in model performance can also be a consequence of transferring features of top layers from two different domains (Yosinski et al., 2014). Specifically, the transferability of features can be negatively affected when the source task (e.g. classification of the ImageNet dataset) is different from the target task (e.g. breast cancer detection). Thus, transfer learning of features for different domains should be performed cautiously (Yosinski et al., 2014). Further, merging and fine-tuning only the last layer and initializing the weights of the whole network based on the ImageNet dataset transferred the specific features (learned in top layers) of the non-medical domain to the medical domain, thus decreasing the classification performance in the strategy 2. To improve the performance of a DCNN model by the transfer learning strategy 2, initializing and fine-tuning weights of the top and intermediate layers of

the multi-input DCNN model should be investigated in future studies.

So far, limitations of the transfer learning strategy 2 were discussed, now it is important to discuss few limitations of the transfer learning strategy 1. One of the limitations is the need for an aggressive downsampling of the pathological images according to the input size of the pre-trained DCNN, ignoring the essential information. Although it is also possible to use a desired input image size by removing the fully connected layers of a pre-trained DCNN, downsampling our patches of size 1024×1024 to the model's input size facilitated the best classification performance. Extracting smaller size patches to increase the number of patches were also evaluated during the analysis. However, it was observed that small size patches increased the dataset size but decreased the biologically significant tissue features in each patch. Irrespective of our acceptable results using the pre-trained DCNNs as feature extractors, the interpretability of the transferred features is questionable. It is difficult to obtain an intuitive understanding of the transferability of non-medical features obtained from the ImageNet dataset to the medical domain. Thus, it is important to investigate transferring features from the medical domain to improve the breast cancer classification rate in future.

5.3 Effect of DCNN architecture

It was clear from the results that acquiring a good classification rate using data fusion approach is dependent on the DCNN model. For transfer learning strategy 1, the Inceptionv3 and the ResNet50 network achieved better classification performances. While for transfer learning strategy 2, the multi-input VGG16 network achieved good classification performance. Furthermore, for transfer learning strategy 1, the Inceptionv3 and the VGG16 provided a large number of features (as they were combined from multiple modalities) in comparison to the ResNet50 network. Large feature dimension not only increased the dataset size but also increased the memory requirement. However, large feature dimension obtained by large DCNNs like the Inceptionv3 and the ResNet50 proved to be beneficial for training the PCA-LDA model in transfer learning strategy 1. While for transfer learning strategy 2, it was seen that large DCNN like the multi-input Inceptionv3 and the multi-input ResNet50 networks easily overfit and degrade model performance. It is suspected that large networks in multi-input fashion like the Inceptionv3 and the ResNet50 network generates a large number of trainable parameters which degrades model performance

during fine-tuning. Furthermore, the time required to fine-tune the last layers of networks increases with network size.

6 CONCLUSION

The results show that combining histological imaging data along with IHC imaging data (estrogen receptor, progesterone receptor, human epidermal growth factor-2 and Ki-67) can improve breast cancer classification rate as compared to histological imaging data alone. The improvement in the classification performance was approximately 2% when deep convolutional neural networks (DCNN) were used as feature extractors (i.e. transfer learning strategy 1). However, the classification performance degraded when fine-tuning of the last layer of the multi-input DCNN (i.e. transfer learning strategy 2) was performed. Out of all three pre-trained networks, the pre-trained residual network and inception network as feature extractor outperformed the binary classification task (tumour vs normal), while the pre-trained VGG network as feature extractor obtained reasonable results. On the other hand, the VGG network showed better performances than the residual network and the inception network when fine-tuning of last layers was performed. The increase in performance by 2% for diagnosing breast cancer is explainable, because this task is normally performed using H&E, so the advancement is limited. Nevertheless, the data fusion approach can substantially improve differential diagnosis, which is important from a clinical perspective. Therefore, combining histology and IHC staining technique should be encouraged in future for more complicated tasks like a differential diagnosis or the prognosis of breast cancer patients. Overall, this comparative study showed that transfer learning could be utilized to diagnose breast cancer based on the combined histological and IHC imaging data with acceptable results. However, it is important to perform this study on a larger dataset in future. On large dataset, transfer learning strategy 3 i.e. training a DCNN from scratch can also be investigated. Furthermore, the data fusion approach can be performed to characterize stages of breast cancer in future.

ACKNOWLEDGEMENTS

Financial support of the German Science Foundation (BO 4700/1-1, PO 563/30-1 and STA 295/11-19) and funding by the BMBF for the project Uro-MDD (FKZ 03ZZ0444J) are highly acknowledged.

REFERENCES

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
- Antony, J., McGuinness, K., O'Connor, N. E., and Moran, K. (2016). Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1195–1200. IEEE.
- Bayramoglu, N. and Heikkilä, J. (2016). Transfer learning for cell nuclei classification in histopathology images. In *European Conference on Computer Vision*, pages 532–539. Springer.
- Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.
- Cheang, M. C., Chia, S. K., Voduc, D., Gao, D., Leung, S., Snider, J., Watson, M., Davies, S., Bernard, P. S., Parker, J. S., et al. (2009). Ki67 index, her2 status, and prognosis of patients with luminal b breast cancer. *JNCI: Journal of the National Cancer Institute*, 101(10):736–750.
- Chen, T. and Chefd'Hotel, C. (2014). Deep learning based automatic immune cell detection for immunohistochemistry images. In *International workshop on machine learning in medical imaging*, pages 17–24. Springer.
- Chollet, F. et al. (2015). Keras. <https://github.com/fchollet/keras>.
- Cireşan, D. C., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 411–418. Springer.
- Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567.
- Damodaran, S. and Olson, E. M. (2012). Targeting the human epidermal growth factor receptor 2 pathway in breast cancer. *Hospital Practice*, 40(4):7–15.
- Dobson, L., Conway, C., Hanley, A., Johnson, A., Costello, S., O'Grady, A., Connolly, Y., Magee, H., O'Shea, D., Jeffers, M., et al. (2010). Image analysis as an adjunct to manual her-2 immunohistochemical review: a diagnostic tool to standardize interpretation. *Histopathology*, 57(1):27–38.
- Elledge, R. M., Green, S., Pugh, R., Allred, D. C., Clark, G. M., Hill, J., Ravdin, P., Martino, S., and Osborne, C. K. (2000). Estrogen receptor (er) and progesterone receptor (pgr), by ligand-binding assay compared with er, pgr and ps2, by immuno-histochemistry in predicting response to tamoxifen in metastatic breast cancer: A southwest oncology group study. *International journal of cancer*, 89(2):111–117.
- Guo, S., Bocklitz, T., Neugebauer, U., and Popp, J. (2017). Common mistakes in cross-validating classification models. *Analytical Methods*, 9(30):4410–4417.
- Hanin, B. (2018). Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems*, pages 582–591.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, Germany.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.
- Huynh, B. Q., Li, H., and Giger, M. L. (2016). Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501.
- Janowczyk, A. and Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7.
- Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python.
- Kensert, A., Harrison, P. J., and Spjuth, O. (2018). Transfer learning with deep convolutional neural networks for classifying cellular morphological changes. *SLAS DISCOVERY: Advancing Life Sciences R&D*, page 2472555218818756.
- Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O., and Hajirasouliha, I. (2018). Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. *EBioMedicine*, 27:317–328.
- Kieffer, B., Babaie, M., Kalra, S., and Tizhoosh, H. R. (2017). Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, Citeseer.
- 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.
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.
- Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26.
- Mormont, R., Geurts, P., and Marée, R. (2018). Comparison of deep transfer learning strategies for digital pathology. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2262–2271.
- Oliphant, T. (2006). NumPy: A guide to NumPy. USA: Trelgol Publishing.
- Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000). Molecular portraits of human breast tumours. *nature*, 406(6797):747–752.
- Pham, N.-A., Morrison, A., Schwock, J., Aviel-Ronen, S., Iakovlev, V., Tsao, M.-S., Ho, J., and Hedley, D. W. (2007). Quantitative image analysis of immunohistochemical stains using a cmyk color model. *Diagnostic pathology*, 2(1):1–10.
- Ravishankar, H., Sudhakar, P., Venkataramani, R., Thiruvankadam, S., Annangi, P., Babu, N., and Vaidya, V. (2016). Understanding the mechanisms of deep transfer learning for medical images. In *Deep Learning and Data Labeling for Medical Applications*, pages 188–196. Springer.
- Ribeiro, E., Uhl, A., Wimmer, G., and Häfner, M. (2016). Exploring deep learning and transfer learning for colonic polyp classification. *Computational and mathematical methods in medicine*, 2016.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.
- Sheikhzadeh, F., Ward, R. K., van Niekerk, D., and Guillaud, M. (2018). Automatic labeling of molecular biomarkers of immunohistochemistry images using fully convolutional networks. *PloS one*, 13(1):e0190783.
- Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Sørli, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.
- Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312.
- Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global.
- Van Ginneken, B., Setio, A. A., Jacobs, C., and Ciompi, F. (2015). Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *2015 IEEE 12th International symposium on biomedical imaging (ISBI)*, pages 286–289. IEEE.
- Venables, W. N. and Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.
- Veta, M., Pluim, J. P., Van Diest, P. J., and Viergever, M. A. (2014). Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering*, 61(5):1400–1411.
- Wang, D., Khosla, A., Gargeya, R., Irshad, H., and Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.
- Webster, L., Bilous, A., Willis, L., Byth, K., Burgemeister, F., Salisbury, E., Clarke, C., and Balleine, R. (2005). Histopathologic indicators of breast cancer biology: insights from population mammographic screening. *British journal of cancer*, 92(8):1366–1371.
- Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

9

Appendix

A1 COMPUTATIONAL TISSUE STAINING OF NON-LINEAR MULTIMODAL IMAGING
USING SUPERVISED AND UNSUPERVISED DEEP LEARNING

Submitted on: 30 November 2020

Revised on: 28 January 2021

Accepted on: 17 February 2021

Published on: 23 March 2021



Computational tissue staining of non-linear multimodal imaging using supervised and unsupervised deep learning

PRANITA PRADHAN,^{1,2}  TOBIAS MEYER,² MICHAEL VIETH,³
ANDREAS STALLMACH,⁴ MAXIMILIAN WALDNER,^{5,6} MICHAEL
SCHMITT,¹ JUERGEN POPP,^{1,2} AND THOMAS BOCKLITZ^{1,2,*} 

¹*Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller-University, Jena, Germany*

²*Leibniz Institute of Photonic Technology, Member of Leibniz Health Technologies Jena, Germany*

³*Institute of Pathology, Klinikum Bayreuth, Bayreuth, Germany*

⁴*Department of Internal Medicine IV (Gastroenterology, Hepatology, and Infectious Diseases), Jena University Hospital, Jena, Germany*

⁵*Erlangen Graduate School in Advanced Optical Technologies (SAOT), Friedrich-Alexander University of Erlangen-Nuremberg, 91052 Erlangen, Germany*

⁶*Medical Department I, Friedrich-Alexander University of Erlangen-Nuremberg, Erlangen, Germany*

*thomas.bocklitz@uni-jena.de

Abstract: Hematoxylin and Eosin (H&E) staining is the 'gold-standard' method in histopathology. However, standard H&E staining of high-quality tissue sections requires long sample preparation times including sample embedding, which restricts its application for 'real-time' disease diagnosis. Due to this reason, a label-free alternative technique like non-linear multimodal (NLM) imaging, which is the combination of three non-linear optical modalities including coherent anti-Stokes Raman scattering, two-photon excitation fluorescence and second-harmonic generation, is proposed in this work. To correlate the information of the NLM images with H&E images, this work proposes computational staining of NLM images using deep learning models in a supervised and an unsupervised approach. In the supervised and the unsupervised approach, conditional generative adversarial networks (CGANs) and cycle conditional generative adversarial networks (cycle CGANs) are used, respectively. Both CGAN and cycle CGAN models generate pseudo H&E images, which are quantitatively analyzed based on mean squared error, structure similarity index and color shading similarity index. The mean of the three metrics calculated for the computationally generated H&E images indicate significant performance. Thus, utilizing CGAN and cycle CGAN models for computational staining is beneficial for diagnostic applications without performing a laboratory-based staining procedure. To the author's best knowledge, it is the first time that NLM images are computationally stained to H&E images using GANs in an unsupervised manner.

Published by The Optical Society under the terms of the [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

1. Introduction

Conventional staining technique like histopathological (H&E) staining is the 'gold-standard' technique for tissue diagnostics. High quality H&E staining requires an embedding of the sample in paraffin to generate FFPE sections, which is time-consuming. Due to the time requirement, conventional H&E staining (using FFPE material) cannot be used for real-time disease diagnosis like in a cryosection setting, where low quality cryosections are stained. Additionally, this technique can only show limited biomolecular information. If bio-molecular information is

needed for diagnostics, it must be acquired using molecular imaging techniques. Thus, in our work, one of the molecular imaging techniques like non-linear multimodal (NLM) imaging is used, which can complement the 'gold-standard' histopathological staining technique. The NLM imaging used here is based on cryosection material which means that the sample is not embedded in FFPE and thus have a time advantage. In that way, biomolecular information can be extracted, and this technique can be applied for real-time disease diagnosis [1,2]. As NLM images used in our study are based on cryo material that were followed by H&E staining, we do not show images of H&E stained FFPE sections (instead histopathologically stained H&E images).

The NLM imaging presented here is a combination of three non-linear optical modalities, namely coherent anti-Stokes Raman scattering (CARS) microscopy, two-photon excitation fluorescence (TPEF) microscopy and second-harmonic generation (SHG) microscopy. These three modalities highlight the distribution of biomolecules like collagen, NADH, proteins and lipids [2,3]. Furthermore, NLM imaging is label-free and provide highly resolved images of biological tissues. The non-destructive nature of NLM imaging is suitable for *in vivo* studies. Due to these properties and the fact that NLM imaging provides morphological and functional information of a tissue sample, this imaging technique is beneficial for tissue imaging and other biomedical applications [3] like investigations of skin diseases [4–6], diagnostics of head-neck cancer [7,8], classification of brain tumors [9], and characterization of inflammatory bowel disease samples [10].

Despite the ever-increasing use of NLM imaging, its establishment in clinics is not achieved until now. In situations where cryo sections are analyzed, NLM would be a great technique, because the computational staining is of higher quality as H&E stains in a cryosection analysis. The NLM images have a higher resolution compared to the histopathologically stained H&E images and exhibit color contrast which is unfamiliar to physicians. For diagnostics, physicians tend to screen histopathologically stained H&E images and then zoom into suspicious regions for diagnostics. This is problematic as the contrast of NLM images is different from corresponding histopathologically stained H&E images, and NLM images can be zoomed to higher tissue levels. Thus, interpretation of NLM images with its corresponding histopathologically stained H&E images is challenging. To interpret the NLM images and link it to standard histopathologically stained H&E images, a parallel tissue section and afterwards the cryosection are stained with conventional staining procedures. Subsequently, the stained images are compared with the corresponding NLM images. This comparison is laborious, which reduces the advantage of NLM imaging. Therefore, comparison of histopathologically stained H&E image and NLM image requires an automatic translation of both images. Furthermore, an automatic translation of different modalities to H&E stained image can aid intraoperative histopathologic diagnosis and efficient decision-making during surgery [11]. Such an automatic model can also generate awareness and trust in the new NLM imaging technique.

In this context, researchers in 2016 performed the modality transfer of NLM images to histopathologically stained H&E images by image analysis and machine learning methods [12]. Although their work showed comparable results (see Fig. 1), the approach had two limitations. Foremost, the colors in the computationally stained H&E images were different compared to the original histopathologically stained H&E images. Secondly, their work trained a machine learning model that required a corresponding pair of NLM images and histopathologically stained H&E images. This training procedure is time-consuming as the histopathologically stained H&E image of the same tissue section must be prepared and registered to the coordinate space of the NLM image before constructing the machine learning model. In most cases, the multimodal image registration is a difficult task due to tissue alterations that occur during the staining procedure.

In contrast, our work presents an improvement of the work of Bocklitz et al., 2016 in terms of the staining results and the required manual effort for modality transfer. This was achieved by utilizing deep learning models instead of conventional machine learning methods. Briefly, deep

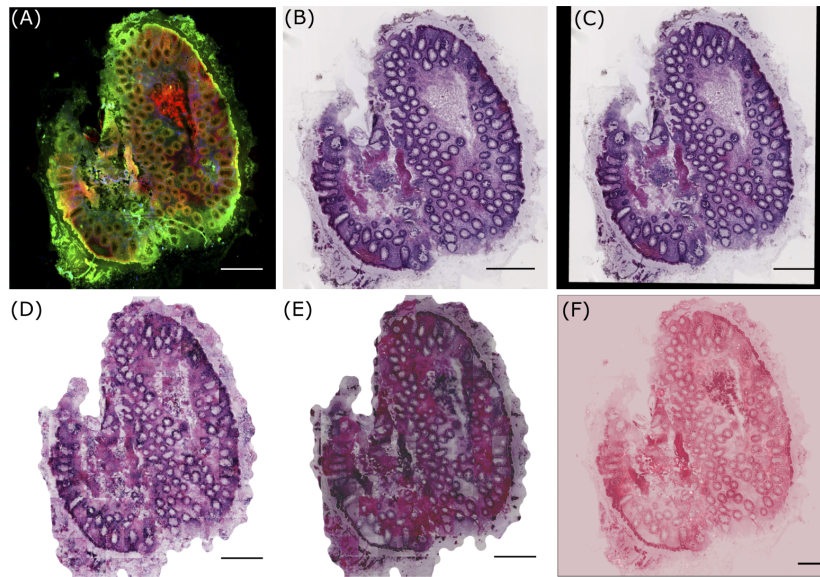


Fig. 1. (A) shows a pre-processed NLM image with CARS, TPEF and SHG as the red, green and blue channel, respectively. (B) visualizes histopathologically stained H&E image (or unregistered H&E image) used for unsupervised pseudo-stain H&E model, (C) depicts a registered H&E image used for supervised pseudo-stain H&E model. The image in (C) shows the registration effect, which is filled with zeros. The images in (D), (E) and (F) are computationally stained H&E images with the supervised, unsupervised approach and method used in reference [12], respectively. All images are downscaled to 20% of the original size for clarity. The scale bar in all images represents 100 μm .

learning models, specifically generative adversarial networks [13], were utilized to translate NLM images into computationally stained H&E images. This work was performed in a supervised and an unsupervised approach, where a paired [14] and an unpaired image translation [15] of the NLM image was performed, respectively. The supervised approach or paired image translation required a corresponding pair of images measured with the two modalities (NLM imaging and H&E staining), while the unsupervised approach did not require paired images of the two modalities. Like the previous work of Bocklitz et al. 2016, the supervised approach has the limitation of registering the histopathologically stained H&E images to the corresponding NLM images. On the other hand, the unsupervised approach does not require the image registration of the two modalities. Moreover, the unsupervised approach offer additional advantages like the artificial generation of images from both modalities, the translation of images from multiple modalities and minimal requirement of stained images.

The supervised and unsupervised approach utilized a conditional generative adversarial network (CGAN) [16] and a cycle conditional generative adversarial network (cycle CGAN) [14], respectively. CGANs are commonly used in computer vision tasks for translating images [14], but they were never used to translate an NLM image to a H&E stained image. Common applications of CGAN in computer vision are the transformation of photographs acquired in daylight into photographs of night scenes, or the transfer of horse images into images of zebras. Likewise, its application in the biomedical and optical field is gaining popularity [17–23]. Recent works

transformed auto-fluorescence images [24] or hyperspectral images [25] into histopathologically stained H&E images using CGANs. A similar approach was performed for translating quantitative phase imaging into three different stains, namely H&E stain, Jone's stain, and Masson's trichrome stain [26]. CGANs were also employed to increase the spatial resolution [27,28] and remove speckle noise from optical microscopic images. Similarly, the cycle CGAN were utilized to stain a H&E stained image into an immunohistochemically (IHC) stained image. The generated IHC image was used to reconstruct the original H&E stained image [29]. As mentioned earlier, unpaired image translation is advantageous as co-registration of images from different modalities is not needed, but the model for the unpaired image translation needs to be more complex as compared to the model for paired image translation.

Our work is different from the state-of-the-art methods because it is the first time that NLM images were used for an unsupervised transfer to histopathologically stained H&E images. While performing the unsupervised modality transfer, the corresponding difficulties were tackled. First of all, the different contrast between the two modalities makes the image translation task complicated. Furthermore, the NLM images used in this work are measured from tissue of patients with different disease severity (namely Inflammatory bowel disease), which is reflected in the alterations of the tissue structure and changes in the pixel contrast [30]. The availability of NLM images is limited, which is problematic because the training of adversarial networks requires large datasets. Lastly, we evaluated the modelling quantitatively by considering perceptual or texture information and color information. Overall, this work is an improvement of the state-of-the-art method presented by Bocklitz et al. in 2016 [12], based on paired and unpaired image translation of NLM images into histopathologically stained H&E images.

2. Material and methods

2.1. Dataset

The dataset used in this work is published elsewhere [30]. Briefly, it consists of tissue samples from biopsies of patients with Crohn's disease, ulcerative colitis or infectious colitis obtained during colonoscopy or surgical resections. The dataset has 19 pairs of NLM images and histopathologically stained H&E images (see Fig. 1). The NLM image is an RGB image where each channel represents one of the three non-linear optical modalities. Precisely, the CARS signal, the TPEF signal, and the SHG signal form the red, green and blue channel of the RGB image, respectively. The spatial (pixel) resolution of the NLM image is $0.227 \mu\text{m}/\text{pixel}$ (see Fig. 1(A)). For the histopathologically stained H&E images, the corresponding tissue sections were stained in the pathology department. The (digital) histopathologically stained H&E images in the form of slide scanner files were extracted using Aperio Image scope software with a spatial resolution approximately equal to the NLM image. The spatial resolution of the extracted histopathologically stained H&E image is $0.219 \mu\text{m}/\text{pixel}$ (see Fig. 1(B)). This spatial resolution setting was favourable for image registration step. The corresponding pairs of NLM and histopathologically stained H&E images were used to construct a "pseudo-stain H&E model" based on the conditional generative adversarial networks in a supervised and an unsupervised approach. The pseudo-stain H&E model was trained using 13 image pairs and tested on six image pairs. For building the pseudo-stain H&E model, both images were pre-processed, and the histopathologically stained H&E image was registered to the NLM image only for the supervised approach.

2.2. Image pre-processing of the histopathologically stained H&E image

The histopathologically stained H&E image was registered to the coordinate space of the corresponding NLM image using the Image processing toolbox in Matlab 2018a. For the image registration purpose, the NLM and histopathologically stained H&E images were converted to

grayscale, followed by contrast inversion of the histopathologically stained H&E image. The contrast inversion was achieved by subtracting the pixel values in each channel of the H&E image by 255. The contrast inversion of histopathologically stained H&E image was performed only for image registration purpose (not for model training). The inverted histopathologically stained H&E image (grayscale) was used as a moving image, and the corresponding NLM image (grayscale) was used as a fixed image. Subsequently, a multimodal image registration [31] based on the mutual information metric was performed using the NLM and the histopathologically stained H&E images. The registered histopathologically stained H&E image (see Fig. 1(C)) was used for the supervised approach or paired image translation, while the unregistered histopathologically stained H&E image (see Fig. 1(B)) was utilized for the unsupervised approach or unpaired image translation. Further, patches of size 256×256 were extracted from the registered and the unregistered histopathologically stained H&E image. All the histopathologically stained H&E patches were scaled in the range $[-1, 1]$ before model training. The patches from the registered histopathologically stained H&E image were used to train the CGAN model, while the patches from the unregistered histopathologically stained H&E image were used to train the cycle CGAN model (see Fig. 2(A) and Fig. 2(B)).

2.3. Image pre-processing of the non-linear multimodal image

The data acquisition and pre-processing of NLM images were similar to Chernavskaia et al., 2016 [30]. Briefly, the pre-processing steps included median filtering, downsampling by a factor of 4, correcting the uneven illumination and adjusting the contrast of the NLM images. A pre-processed NLM image is shown in Fig. 1(A). Subsequently, the contrast of NLM images was inverted by subtracting the pixel values by 255. Contrast-inversion of NLM images was performed only for GAN model training. Further, patches of size 256×256 were extracted from the “contrast-inverted” NLM image (see Fig. 2(C) and Fig. 2(D)). These patches were filtered separately for the supervised and unsupervised approach. For the supervised approach or the CGAN model, the pair of NLM and histopathologically stained H&E patch showing registration artefact were removed. The registration artefacts were seen at the borders of the registered histopathologically stained H&E image, which were filled with zero values during registration (see Fig. 1(C)). For the unsupervised method or the cycle CGAN model, the NLM and histopathologically stained H&E patches belonging to the background region were removed using the homogeneity factor [32], i.e. the patches with homogeneity factor greater than 60% were removed [32]. Similar to the H&E patches, all the selected NLM patches were normalized in the range $[-1, 1]$ before model training.

2.4. Conditional generative adversarial network

The conditional generative adversarial network (CGAN) used in this work was inspired by the Pix2Pix model developed by Isola et al., 2017 [14]. The Pix2Pix model comprised of a generator (\mathcal{G}) and a discriminator (\mathcal{D}) (see Fig. 2(E)). The generator with an autoencoder architecture [33] transforms a contrast-inverted NLM patch (x_m) to a computationally stained H&E patch ($z_{generated} = \mathcal{G}(x_m)$) which looked visually similar to the histopathologically stained H&E patch (z_{target}). The input to the generator was a pre-processed NLM patch (see column B in Fig. 3) and a target or histopathologically stained H&E patch (see column C in Fig. 3). The computationally stained H&E patch, i.e. output of the generator, (see column D in Fig. 3) was evaluated by calculating mean absolute error with the target histopathologically stained H&E patch and was optimized to be minimal. The discriminator model was trained to predict the plausibility of the computationally stained H&E patch ($z_{generated}$). In simpler words, the discriminator model was trained to predict if the computationally stained H&E patch was ‘fake’ (i.e. not belonging to histopathologically stained H&E patches) or ‘real’ (i.e. belonging to the original dataset

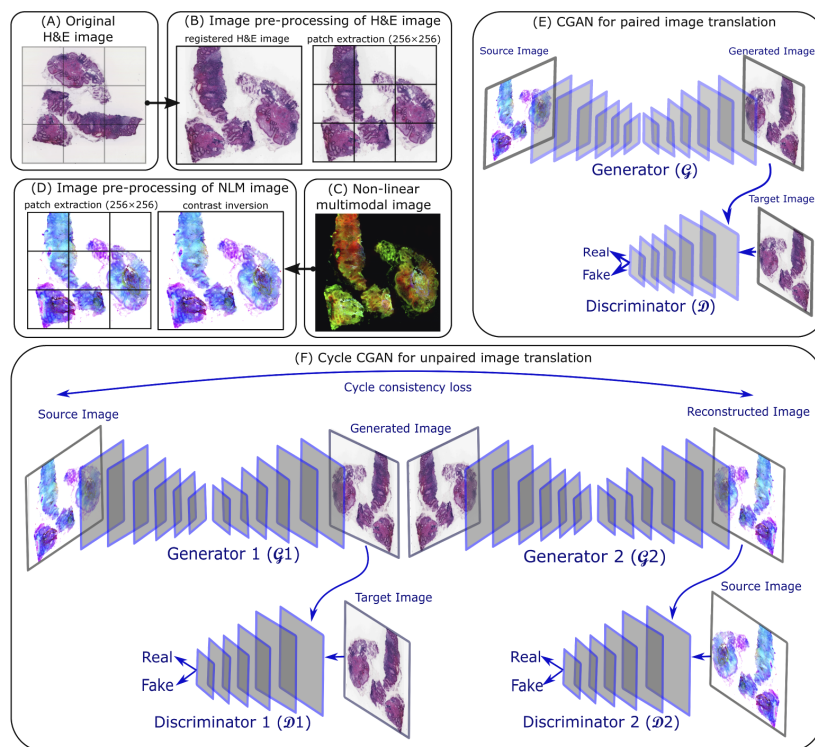


Fig. 2. (A) is a histopathologically stained H&E image, (B) shows the image pre-processing of a histopathologically stained H&E image including image registration and patch extraction of size 256x256, (C) is a corresponding NLM image, (D) visualizes the contrast inversion of the NLM image followed by patch extraction of size 256x256, (E) shows a CGAN model for paired image translation which utilizes the registered histopathologically stained H&E images and contrast inverted NLM images, (F) depicts a cycle CGAN model for unpaired image translation using unregistered histopathologically stained H&E images and contrast inverted NLM images.

of histopathologically stained H&E patches). The details of the generator and discriminator networks are elaborated below.

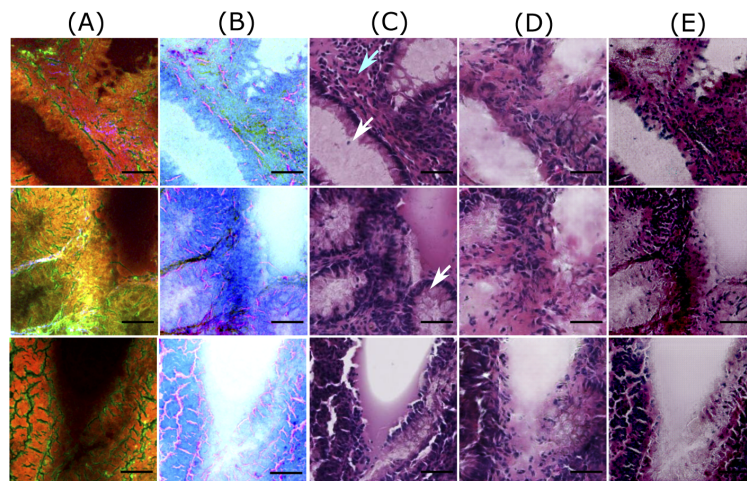


Fig. 3. Good quality predictions of patches from the training dataset. Columns (A) and (B) visualize NLM patches and contrast inverted NLM patches, column (C) shows histopathologically stained H&E patch, and columns (D) and (E) visualize computationally stained H&E patch generated by the Pix2Pix model and the cycle CGAN model, respectively. The scale bar represents $50 \mu\text{m}$. For all patches in column C, the region within the crypts (pointed by white arrows) is light or pale pink, whereas the epithelial layer or mucosa region outlining the crypts appears dark purple. Similar colors with few variations are observed in the computationally stained H&E patches generated by the Pix2Pix (column D) and the cycle CGAN (column E) model. Also, the crypt structures are efficiently generated in the computationally stained H&E patches by both models. It is also observed that nucleus (pointed by cyan arrow) are not generated by the cycle CGAN model (column E).

The generator network was inspired by the U-Net model [34] which is an autoencoder. The autoencoder model had eight blocks in the encoder and the decoder part. Each block of the encoder utilized convolution layer, batch normalization layer and Leaky ReLU activation layer. The last layer of the encoder was a bottleneck layer without batch normalization layer. The eight encoder blocks comprised of 64, 128, 256, 512, 512, 512, 512 and 512 filters, respectively. On the other hand, each decoder block comprised of a convolution layer, batch normalization layer, dropout layer with a 50% dropout rate and ReLU activation layer. Like the encoder network, the first layer of the decoder did not use a batch normalization layer. The layers in eight decoder blocks comprised of 512, 1024, 1024, 1024, 1024, 512, 256 and 128 filters, respectively (after concatenation from the encoder). All the convolutional layers in the encoder and the decoder blocks used a kernel size of 4 and stride size of 2. The encoder and the decoder models were linked through 'skip-connections' similar to the U-Net architecture. The output layer of the generator was a single convolutional layer with three channels and tanh activation function. The output of the generator was a computationally stained H&E patch ($z_{generated} = \mathcal{G}(x_m)$) which was one part of the discriminator network's input.

The discriminator network was a standard convolutional neural network with input as computationally stained H&E patch ($z_{generated}$) and histopathologically stained H&E patch (z_{target}) of size 256×256 . The architecture of the discriminator network was inspired by the 'PatchGAN'

discriminator given in Ref. [14]. The basic idea of the PatchGAN discriminator model is to classify an $N \times N$ region in the $M \times M$ input image ($N < M$) as 'real' or 'fake', instead of classifying the whole $M \times M$ input image as 'real' or 'fake'. In our case, $M = 256$ and $N = 70$ i.e. a 70×70 region in the 256×256 computationally stained H&E patch was classified as 'real' or 'fake'. The 70×70 region is termed as the 'receptive field'. The output of the discriminator model was a map with 16×16 values scaled using a sigmoid activation function. In other words, each value in the 16×16 sigmoid activation map corresponded to the probability of the 70×70 region in the input patch being 'real' (1.0) or 'fake' (0.0). These values were combined to achieve a single probability value, which corresponded to the probability of the entire input patch being 'real' or 'fake'. The layers of the PatchGAN discriminator model were adjusted to maintain the receptive field size to 70×70 . Specifically, the layers of the PatchGAN discriminator model used 64, 128, 256 and 512 filters respectively, and Leaky ReLU activation function with slope 0.2. The configuration of the Leaky ReLU activation function, kernel size and stride size were the same for both the generator and discriminator networks.

Before training of generator and discriminator networks, the weights of both networks were initialized using random Gaussian numbers with a standard deviation of 0.02. During the training phase, the weights of the discriminator model were updated by a set of histopathologically stained H&E patches (z_{target}) and computationally stained H&E patches ($z_{generated}$), and calculating the discriminator loss

$$\mathcal{L}_{\mathcal{D}} = \mathcal{D}(z_{generated})^2 + (1 - \mathcal{D}(z_{target}))^2. \quad (1)$$

When the discriminator network is better than the generator network, i.e. $\mathcal{D}(z_{target}) = 1$ and $\mathcal{D}(z_{generated}) = 0$, it is able to identify all the computationally stained H&E patches as 'fake'. To avoid the discriminator network to become better than the generator network, the training process of the discriminator network was slowed down by weighting the discriminator loss $\mathcal{L}_{\mathcal{D}}$ by 50% for each model update [35]. The ideal case is to converge the discriminator loss to 0.5 and the generator to create H&E patches exactly similar to the target histopathologically stained H&E patches. On the other hand, the weights of the generator network were updated by calculating the mean absolute error between $z_{generated}$ and z_{target} . Additionally, the weights of the generator network were updated through the adversarial loss obtained from the discriminator network. Thus, the total loss of the generator network $\mathcal{L}_{\mathcal{G}}$ is given by

$$\mathcal{L}_{\mathcal{G}} = \lambda \text{MAE}(z_{generated}, z_{target}) + (1 - \mathcal{D}(z_{generated}))^2, \quad (2)$$

where the mean absolute error (MAE) was weighted by a hyperparameter λ . In our case, λ was set to 10. The weights of the generator and discriminator networks were updated separately to avoid misleading updates. Furthermore, both networks were trained using the Adam optimizer [36] with learning rate and β set to 0.0002 and 0.5, respectively.

2.5. Cycle conditional generative adversarial networks

The cycle CGAN model is an extension of the conditional generative adversarial network which does not require paired images for the image translation task [15]. The cycle CGAN model involved simultaneous training of two generators ($\mathcal{G}_1, \mathcal{G}_2$) and two discriminators ($\mathcal{D}_1, \mathcal{D}_2$) (see Fig. 2(F)). The first generator \mathcal{G}_1 utilized a contrast-inverted NLM patch (x_m) (see Fig. 3, column B) as input and generated an H&E stained patch as output ($z_{generated} = \mathcal{G}_1(x_m)$) (see Fig. 3, column E). The second generator utilized the computationally stained H&E patch (i.e. the output of the generator 1, $z_{generated}$) as input and reconstructed it to the original NLM patch (similar to the input of the generator 1, $\tilde{x}_m = \mathcal{G}_2(z_{generated})$). The output of the second generator (\tilde{x}_m) was optimized to be visually similar to the input of the first generator (x_m) and was regularized by calculating the (forward) cycle consistency loss $\mathcal{L}_{(cyc_f)}$. In similar fashion, backward cycle consistency loss $\mathcal{L}_{(cyc_b)}$ regulated the second generator. Additionally, the first generator \mathcal{G}_1 was

regularized by the identity loss $L_{(id_1)}$ which means that the first generator network utilized the histopathologically stained H&E patch (z_{target}) and reconstructed it at its output. We included only identity mapping loss $\mathcal{L}_{(id_1)}$ for generator \mathcal{G}_1 as we were interested in creating flawless H&E images. However, identity mapping loss $\mathcal{L}_{(id_2)}$ for generator \mathcal{G}_2 can be included in future studies when better reconstruction on NLM images is desired. Furthermore, each generator had its own discriminator model, which predicted the plausibility of the generated outputs. This is like the CGAN model explained earlier where each generator-discriminator pair was trained in an adversarial process. The architecture of the two discriminator models in the cycle CGAN model was similar to the Pix2Pix model; however, the architecture of generator networks was different.

The generator networks were inspired by the architecture proposed by Isola 2017 [14]. Both generator networks ($\mathcal{G}_1, \mathcal{G}_2$) used input image size 256×256 , and the outputs were a computationally stained H&E patch ($z_{generated}$) and a reconstructed NLM patch (\tilde{x}_m), respectively. The generator networks comprised of downsampling convolution blocks to encode the input, a sequence of six ResNet blocks, and upsampling convolution blocks that decodes the bottleneck features to an output. The shorthand notation of the generator network can be given as C7s1-64, D128, D256, R256, R256, R256, R256, R256, U128, U64, C7s1-3 where C7s1-k denotes a 7×7 Convolution-InstanceNorm-ReLU layer with k filters and stride 1. Dk denotes a 3×3 Convolution-InstanceNorm-ReLU layer with k filters and stride 2. Uk denotes a 3×3 fractional-strided-Convolution-InstanceNorm-ReLU layer with k filters and stride 1/2. Rk denotes a ResNet block that contains two 3×3 convolutional layers with the same number of filters on both the layers. Like the CGAN model, the last layer of the generator network comprised of the tanh activation function. The weights of the generator networks were updated through adversarial loss, identity loss [29] and cycle consistency losses [29] (including forward and backward cycle). Mathematically, the full objective function of the cycle CGAN model can be given as

$$\begin{aligned} \mathcal{L}(\mathcal{G}_1, \mathcal{G}_2, \mathcal{D}_1, \mathcal{D}_2) = & \mathcal{L}_{(\mathcal{D}_1)}(\mathcal{G}_1, \mathcal{D}_1, X, Y) + \mathcal{L}_{(id_1)}(\mathcal{G}_1, \mathcal{D}_1, X) \\ & + \lambda \mathcal{L}_{(cyc_f)}(\mathcal{G}_1, \mathcal{G}_2) + \lambda \mathcal{L}_{(cyc_b)}(\mathcal{G}_1, \mathcal{G}_2), \end{aligned} \quad (3)$$

where $\mathcal{L}_{(\mathcal{D}_1)}$ is the adversarial loss through which the generator 1 was updated. This is mean squared error instead of binary cross-entropy in the Pix2Pix model, as it provided better results in the literature [14]. In future, the adversarial loss for second generator $\mathcal{L}_{(\mathcal{D}_2)}$ can also be added. The identity loss $\mathcal{L}_{(id_1)}$ and the forward and backward cycle loss $\mathcal{L}_{(cyc_f)}$, $\mathcal{L}_{(cyc_b)}$ are the mean absolute error. The four losses were weighted by a factor of 1, 5, 10 and 10, respectively. The training of each generator-discriminator pair was similar to the CGAN model.

2.6. Model training and removal of patch-effect

The Pix2Pix model and the cycle CGAN model were trained on patches obtained from the NLM images and histopathologically stained H&E images. Both models were trained for 100 epochs, and a batch size of one patch was used. The model training was performed using Python 3.5 on a commercially available PC system with NVIDIA GeForce GTX 1060, 6GB. The generator models were saved after every fifth epoch, and the model that generated clinically acceptable H&E images from the training dataset (on visual inspection) was used for predicting the images from the test dataset. During the testing phase, the test images were pre-processed in a similar fashion as the training dataset. Further, the prediction of the images in the test dataset was performed on patches, which were subsequently combined to a whole image. Combining the patches resulted into a “patch-effect” which was visible at the edge of each patch in the combined image, precisely the pixel at every 256^{th} row or column in the whole image. For this purpose, the pixels which showed the patch-effect were linearly interpolated with its neighbouring three pixels. The generated H&E images (before and after correction of patch-effect) from both Pix2Pix and cycle CGAN models were visually inspected (Fig. S1 in Supplement 1). In addition to the visual inspection, quantitative evaluation of the computationally stained H&E images obtained from

both pseudo-stain H&E models was performed. The quantitative evaluation was based on three metrics explained further.

2.7. Evaluation method

For performance quantification, the mean squared error (MSE) was utilized to calculate the error between the histopathologically stained H&E image (z_{target}) and the computationally stained H&E image ($z_{generated}$) [37]. However, MSE has a limitation caused due to arbitrarily high numbers which are difficult to standardize. Also, the MSE metric is inconsistent with the human perception ability [37]. Therefore, two other metrics namely structure similarity index (SSIM) [37,38] and color shading similarity (CSS) [39], which are well-suited for evaluating GAN performances, were utilized.

The structure similarity index [37,38] quantifies the perceptual similarity between the two images (z_{target} , $z_{generated}$) by considering the contrast, luminance and texture of these images. Mathematically, SSIM between two images X and Y can be given as

$$SSIM(X, Y) = \frac{(2\mu_X\mu_Y + c_1)(2\sigma_{XY} + c_2)}{(\mu_X^2 + \mu_Y^2 + c_1)(\sigma_X^2 + \sigma_Y^2 + c_2)}, \quad (4)$$

where μ is mean of an image, σ is the standard deviation of an image, σ_{XY} is the cross-covariance of the two images and $C = (c_1, c_2)$ are constants to avoid division by zero. For a multichannel image or an RGB image, the SSIM metric is calculated for each channel separately and the average SSIM value is considered. Higher values of SSIM indicate higher structural similarity between the two images.

Another metric called color shading similarity (CSS) [39] was used to quantify the similarity in the colors of the pixels in the histopathologically stained H&E image (z_{target}) and the computationally stained H&E image ($z_{generated}$). The CSS is calculated by converting both images in the CIELAB space and utilizing only the color channels A* and B*. For each color channel, the mathematical formulation of the CSS metric between two images X and Y is given by

$$CSS(X, Y) = \frac{1}{N} \sum_{i=1}^N Ind_i \cdot Sim(X_i, Y_i), \quad (5)$$

where, $Sim(X_i, Y_i)$ is the similarity between pixel X_i of the histopathologically stained H&E image and pixel Y_i of the computationally stained H&E image, i is the index used for the N pixels in the image. Mathematically, Sim and Ind are given as,

$$Sim(X_i, Y_i) = 1 - \frac{dist(X_i, Y_i)}{\max(dist)}, \quad (6)$$

$$Ind_i = \begin{cases} 1; & \text{if } Sim(X_i, Y_i) > \text{threshold,} \\ 0; & \text{if } Sim(X_i, Y_i) \leq \text{threshold.} \end{cases}$$

In our case, the threshold was set to 0.5, and $dist$ was the absolute distance. Higher values of the CSS indicate a higher color similarity. Likewise, the three metrics were evaluated for all the computationally stained H&E images (excluding the background region) from the Pix2Pix and the cycle CGAN models. The average of the three metrics calculated for all the 19 images from the training and test dataset is reported in Table 1 and Fig. 7.

Table 1. The average of the three evaluation metrics obtained for the 19 images using the Pix2Pix model and the cycle CGAN model are given for training and testing dataset. For reference purpose, the three metrics were also calculated with the same histopathologically stained H&E image. It is seen that MSE values are very large for both models, whereas SSIM and CSS are almost similar for both models. This means that the pixel values of computationally stained H&E images are different, but the overall structural and color information is acceptable. Furthermore, the metric values for training and testing dataset do not have a large difference, which indicates that the models are minimally overfitted.

	MSE	SSIM	CSS
Training dataset			
Pathological stained H&E image	0.00	1.00	1.00
Pix2Pix stained H&E image	4.69×10^3	0.52	0.93
Cycle CGAN stained H&E image	10.26×10^3	0.49	0.91
Testing dataset			
Pathological stained H&E image	0.00	1.00	1.00
Pix2Pix stained H&E image	4.27×10^3	0.60	0.94
Cycle CGAN stained H&E image	7.79×10^3	0.59	0.93

3. Results

The training of the Pix2Pix model for 100 epochs required ~7 hours, while the cycle CGAN model required more than ~100 hours on our commercial PC. The training of the cycle CGAN model was terminated after 60 epochs as no significant improvement in the computationally stained H&E patches was observed visually. Furthermore, it was observed that the generator and discriminator losses for the Pix2Pix model and the cycle CGAN model fluctuated throughout the training process. The discriminator loss for both models had difficulties to remain converged at an ideal value ≥ 0.5 , which can be due to high variance [40,41] and noise in our dataset. The computationally stained H&E patches obtained using the saved generator models were visually assessed for their quality. A detailed explanation of the assessment procedure is provided in the next section.

3.1. Visual similarity of the GAN generated images

Computational staining of NLM images using CGANs achieved visually pleasing results compared to the state-of-the-art machine learning model used by Bocklitz et al. in 2016 [12] (see Fig. 1). As the visual appearance directly impacts the histopathological examination of any disease [25,42], its qualitative evaluation is vital. For this purpose, computationally stained H&E patches using the Pix2Pix and the cycle CGAN model from the training and testing dataset were inspected for different tissue regions.

Figure 3 and Fig. 4 show computationally stained H&E patches in good and bad quality from the training dataset, respectively. Similarly, Fig. 5 and Fig. 6 show good and bad quality computationally stained H&E patches from the testing dataset. For the good quality patches, it can be observed that computationally stained H&E patches in columns D and E of Fig. 3 and Fig. 5 look visually similar to the histopathologically stained H&E patches in column C. Precisely, the good computationally stained H&E patches show a color contrast similar to the histopathologically stained H&E patches, i.e. the region within the crypts (marked by white arrows) is light or pale pink, whereas the epithelial layer or mucosa region outlining the crypts appears dark purple or dark pink. Furthermore, the regions showing nuclei (marked by cyan

arrow) are generated better by the Pix2Pix model compared to the cycle CGAN model. Overall, the tissue structures in the good computationally stained H&E patches are clinically acceptable. Nevertheless, the bad quality patches as shown in columns D and E of Fig. 4 and Fig. 6 look visually different than the histopathologically stained H&E patches shown in column C. In the bad computationally stained H&E patches, the structures within the crypts (marked by white arrows) are lost and the colors in stroma regions (marked with green arrow) are wrongly modelled. Furthermore, the nuclei signals present in Fig. 4, column C (marked with cyan arrow) are not efficiently generated in column D and E. Inconsistent modelling of nuclei signals in computationally stained H&E images is expected as NLM images show negative contrast for the cell nucleus. Sometimes the nuclei are out of focus in NLM images and due to this fact, there is no nuclei contrast. It is observed that nuclei signals and stroma region are occasionally generated in column C which can be a systematic error of GAN based model. Nevertheless, in our application (in contrast to oncology) crypt structures are more important rather than the shape of the cell nuclei in the stroma.

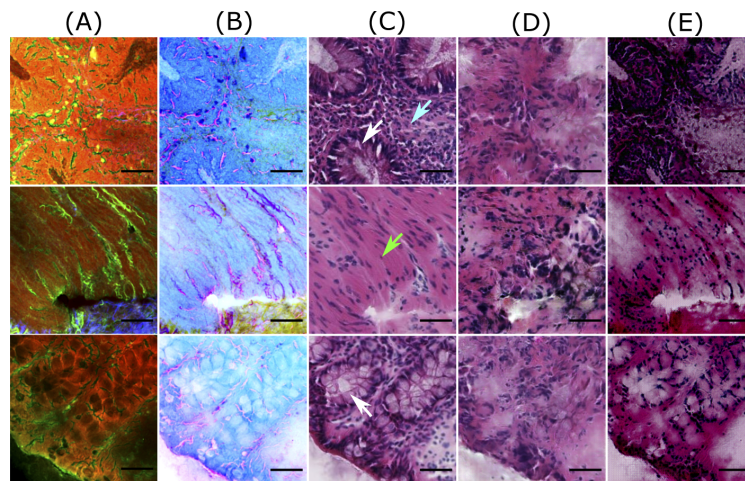


Fig. 4. Bad quality predictions of patches from the training dataset. Columns (A) and (B) visualize NLM patches and contrast inverted NLM patches, column (C) shows histopathologically stained H&E patch, and columns (D) and (E) visualize computationally stained H&E patch generated by the Pix2Pix model and the cycle CGAN model, respectively. The scale bar represents $50 \mu\text{m}$. Here the patches generated by the cycle CGAN model show a promising translation of corresponding NLM patches; however, the colors in stroma regions (marked by green arrows) and nuclei signals (marked by cyan arrows) are not well represented. The computationally stained H&E patches generated by the Pix2Pix model have a low spatial resolution, as the structures within the crypts (pointed by white arrows) are not visible.

The variations between the NLM and its corresponding histopathologically stained H&E patch is due to the optical properties of the NLM imaging technique. The NLM imaging technique shows structures in a focal plane within the tissue section, which is approximately $\sim 5 \mu\text{m}$. In contrast, the histopathological staining technique reveals the structures from the entire thickness of the tissue section ($\sim 20 \mu\text{m}$). Thus, both modalities show slightly different structures, which can be seen through a fine observation of patches in columns A and C of Figs. 3–6. This

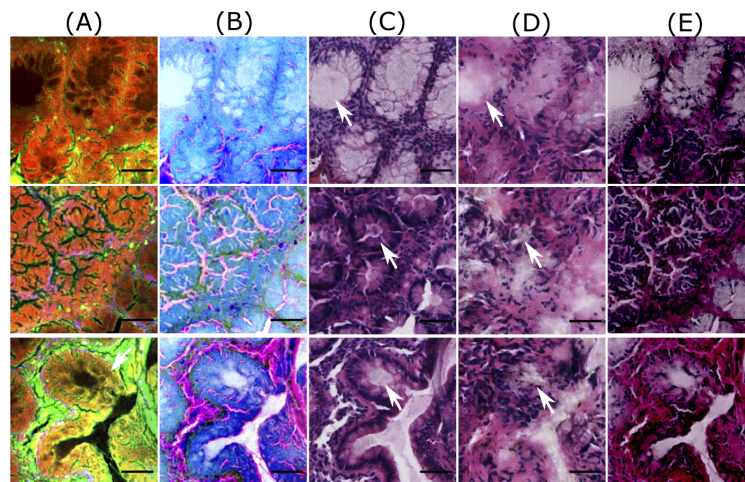


Fig. 5. Good quality predictions of patches from the test dataset. Column (A) shows NLM patches, column (B) visualizes contrast inverted NLM patches, column (C) shows histopathologically stained H&E patch, and columns (D) and (E) depict computationally stained H&E patch by the Pix2Pix model and the cycle CGAN model, respectively. The scale bar represents 50 μm . Here, the computationally stained H&E patches by the cycle CGAN model shows a good quality translation of NLM patches, whereas the translation by the Pix2Pix model produces blurry results within the crypt regions (marked by white arrows).

is the reason that an exact correspondence between computationally stained H&E patch and histopathologically stained H&E patch cannot be achieved for all images in the dataset.

In addition to the structural differences between histopathologically stained H&E patches and computationally stained H&E patches, there are other critical issues which need attention. Foremost, it can be seen from Figs. 3–6 that the Pix2Pix model generates low spatial resolution images as compared to images generated using the cycle CGAN model. The Pix2Pix model tends to lose detailed boundaries and edges within the crypt region (marked by white arrows) and show a blurry effect, at least for images from the test dataset (see Fig. 5 and Fig. 6). One of the reasons for the loss of detailed information can be the mean absolute error which was used as the objective function while training the generator network in the Pix2Pix model. The next critical issue was that the computationally stained H&E patches generated by the cycle CGAN model showed higher color contrast, thus making the colors more vivid. The high color contrast in the computationally stained H&E patch by the cycle CGAN model can be due to unsupervised training. It is suspected that the unsupervised training of the cycle CGAN model can be sensitive to alterations in the pixel intensity of the NLM images, staining inconsistencies in the histopathologically stained H&E image or a pre-processing effect [40,41]. Nevertheless, the problem of high color contrast observed in the computationally stained H&E patches from the cycle CGAN model can be reduced by simple image processing methods like contrast adjustment [43]. Overall, from the visual appearance of the computationally stained H&E patches, it can be seen that an exact correspondence with histopathologically stained H&E patches cannot be achieved. However, the computationally stained H&E patches in column D and E of Figs. 3–6 generated using the Pix2Pix and the cycle CGAN model provide an acceptable translation of

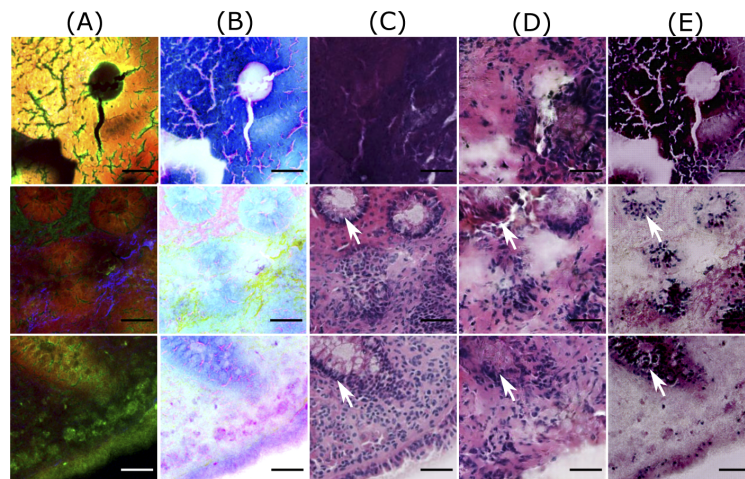


Fig. 6. Bad quality predictions of patches from the test dataset. Column (A) shows NLM patches, column (B) visualizes contrast inverted NLM patches, column (C) shows histopathologically stained H&E patch, and columns (D) and (E) depict computationally stained H&E patch by the Pix2Pix model and the cycle CGAN model, respectively. The scale bar represents 50 μm . Here, computationally stained H&E patches (columns D and E) are not similar to histopathologically stained H&E patches (column C), as histopathologically stained H&E patches show different structures than the corresponding NLM patch. Furthermore, in dark NLM patches (second and third row), the computationally stained H&E patches fail to generate appropriate color contrast in crypt regions (marked by white arrows).

NLM patches. In Fig. S2 and Fig. S3 in [Supplement 1](#), computationally stained H&E patches combined into an image, and computationally stained H&E images from the method presented in [12] are shown. Furthermore, the computationally stained H&E image from the cycle CGAN model followed by contrast reduction with a factor of 0.7 is also visualized. The computationally stained H&E images were also examined by a histologist for its clinical significance. According to the expert analysis, both models show promising results for translating NLM images. In addition to visual analysis, a quantitative evaluation was done and is discussed below.

3.2. Quality of computationally staining based on metrics

An evaluation of the Pix2Pix and the cycle CGAN model was performed based on three metrics: MSE, SSIM and CSS. The average values of MSE, SSIM and CSS for training and testing dataset are reported in [Table 1](#). Here, the three metrics were calculated with the same histopathologically stained H&E image and were considered as baseline values. The aim of the pseudo-stain H&E models was to acquire values “close” to these baseline values.

From [Table 1](#), it can be seen that the computationally stained H&E images generated from both models show very high MSE and low SSIM values as compared to the baseline values. High MSE and low SSIM values were expected as an exact correspondence of computationally stained H&E image with its histopathologically stained H&E image cannot be achieved. Thus, the interpretation of the image quality based on the MSE and SSIM metric is unfair. Despite the high MSE or low SSIM values, the computationally stained H&E images from both models shown in [Fig. S2](#) and [Fig. S3](#) in [Supplement 1](#) have promising visual appearance when compared

to its NLM image. Furthermore, the MSE values are higher for the cycle CGAN model as compared to the Pix2Pix model. Higher MSE values using the cycle CGAN model are suspected due to largely different pixel values of computationally stained H&E patches. On the other hand, the SSIM and CSS metrics report similar performance for the Pix2Pix and the cycle CGAN model, which implies that the overall structural and color content of the computationally stained H&E image is acceptable. Furthermore, the metric values are similar for training and testing dataset (see Table 1) which shows that the models are minimally overfitted. The mean SSIM and mean CSS metric for the training and the testing dataset using both models are >0.50 and >0.90 , respectively. The three metrics for all images are given in Table S1 in Supplement 1.

In addition to Table S1 in Supplement 1, the range of the three metrics is given in Fig. 7, which shows a large variance in the three metrics. The large variance in the three metrics was expected and can possibly be due to the large variance in the dataset. Nevertheless, the color information produced by both pseudo-stain H&E models is close to the baseline value (1.0). The three metrics for a randomly chosen H&E image generated using both models are given in Fig. S2 for the testing dataset and Fig. S3 for the training dataset in Supplement 1. Lastly, there was no significant difference in the performance metrics before and after correction of patch-effect of computationally stained H&E images.

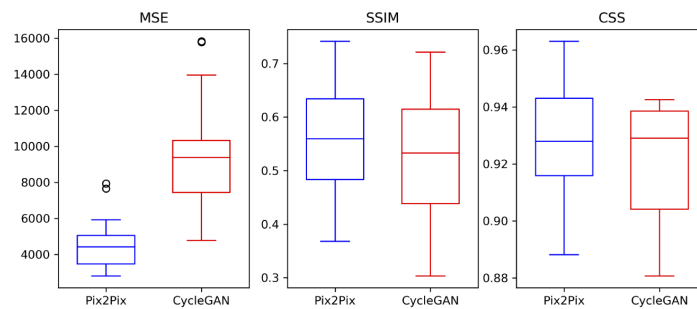


Fig. 7. The boxplot shows a quantitative comparison of the Pix2Pix and the cycle CGAN model based on the three evaluation metrics. The MSE metric is higher for the cycle CGAN model and shows larger variation. This is expected as the pixel values of computationally stained H&E images generated by the cycle CGAN model differ more than the computationally stained H&E images generated by the Pix2Pix model. Nevertheless, the CSS and SSIM metric is in a similar range for both models, which implies that the content of computationally stained H&E images generated by both models is similar.

4. Discussion

The computationally stained H&E images generated by the supervised (Pix2Pix) and the unsupervised (cycle CGAN) pseudo-stain H&E model showed a substantial improvement to the state-of-the-art machine learning model [12] based on color contrast. However, it was observed that the training time of the machine learning model was less (~4 hours) than the training time of CGAN models presented here. Nevertheless, we believe that the cycle CGAN model can provide better computationally stained H&E images. Furthermore, pseudo-stain H&E model can be applied for multi-modality conversion, augment the NLM images and remove noise from multimodal images. In all these tasks, a systematic investigation is needed. Realization of these tasks using the cycle CGAN model can cause staining protocols less labor intensive [11]. However, there are some important aspects considered for training both models, particularly,

the training dataset, the pre-processing of the histopathologically stained H&E image and NLM image and the objective function. These aspects are discussed in more detail below.

4.1. Effect of training dataset

The first aspect is the training dataset utilized for constructing the pseudo-stain H&E models. Similar to any other deep learning networks, pseudo-stain H&E model based on CGANs are also sensitive to the training dataset. It was observed that the training dataset with a large number of noisy patches or background patches affected the convergence of the generator and the discriminator network. Therefore, patch filtering was vital. Furthermore, a large number of trainable parameters in the generator and the discriminator network can easily cause overfitting on the training dataset. This was a major problem in the supervised approach, e.g. the Pix2Pix model, where the target patches were available. The overfitting on training dataset is seen in Fig. 5 and Fig. 6, column D. Here, we see the patches from the test dataset lose their spatial resolution compared to the patches from the training dataset in the Fig. 3.

Contrarily, in the unsupervised approach, the cycle CGAN model trained on unpaired image data required a quality check of the training dataset. It was observed that for the cycle CGAN model, the color of the computationally stained H&E patches was influenced by the color of the majority of patches in the training dataset. For instance, the cycle CGAN model trained with a large number of patches from the stroma region, i.e. patches with pink color, was likely to produce pinkish H&E images. Therefore, to create a balance in the color of the generated H&E images, a manual quality check of the patches in the training dataset was crucial for the training of the cycle CGAN model. We believe with an increasing dataset and computation power, the performance of both pseudo-stain H&E models can be improved.

4.2. Effect of the objective function and performance metric

The next aspect for training the pseudo-stain H&E models is the objective function and the performance metric. We begin with the selection of the objective function. Foremost, an appropriate selection of the objective function for the generator and the discriminator network is important to generate clinically acceptable H&E images. In this regard, researchers have shown the benefits of using various objective functions like the style transfer loss [44], the perceptual loss [44], the total variation loss [24] and the image gradient loss [45]. Nevertheless, in our case, the L1-loss for the generator network of the cycle CGAN model showed acceptable results. We believe that the addition of other losses to the objective function can improve the perceptual quality of the generated H&E images yet increasing the model complexity. These losses can be applied for the Pix2Pix and the cycle CGAN model and researched in future studies.

The second aspect is the performance metric. The performance metrics used in this work were calculated on the pixel basis and are sensitive to slight variations in the computational H&E images. For instance, a histopathologically stained H&E image and a computationally stained H&E image offset by one pixel can create a major difference in these performance metrics [44]. This problem is often encountered during registration of H&E image and NLM image. Therefore, the high values of the MSE and low values of the SSIM metric shown in Table 1 is justified. In future studies, an objective function that can evaluate the global quality of the computationally stained H&E images can be utilized.

4.3. Effect of image normalization and contrast inversion

In the end, this section discusses the aspect of image normalization and “contrast-inversion” performed while training the pseudo-stain H&E models. Foremost, the normalization methods of both NLM and histopathologically stained H&E images was essential to avoid multiplications of large numbers during the training process. During the training phase, several methods of normalizing the NLM images and histopathologically stained H&E images were evaluated. It was

observed that the NLM and histopathologically stained H&E patches scaled in the range $[-1,1]$ generated the best results. It was also observed that scaling of NLM and histopathologically stained H&E images instead of scaling its patches did not affect the training or model performance. Furthermore, scaling the NLM and/or histopathologically stained H&E patches in the range $[0,1]$ led to the failure of the discriminator network by immediately converging the discriminator losses to zero. The scaling of histopathologically stained H&E patches was essential due to the tanh activation function used in the last layer of the generator network [13]. These findings coincide with the results of Ref. [46].

In addition to the normalization, “contrast-inversion” of the NLM images was performed to remove the “inverse-color” effect [47]. This effect was seen when the original NLM image (without contrast inversion) was used (see Fig. S4 in Supplement 1). This effect was especially seen in the unsupervised approach, i.e. using the cycle CGAN model. Because of this effect, the crypt region was transformed into dark purple instead of light pink and vice versa. Therefore, “contrast-inversion” was an important step for modality conversion, especially where the two modalities showed significantly different color contrasts.

5. Conclusion

Computational staining of NLM images is beneficial from a clinical perspective as it prevents the staining procedures and reduces manual effort. This work was an improvement of the state-of-the-art method, which utilized the conventional machine learning approach for computational staining of NLM images. On the contrary, this work presented a supervised and unsupervised approach to computationally stain NLM images into H&E stained images. The supervised approach utilized the Pix2Pix model, and the unsupervised approach used the cycle CGAN model. For the Pix2Pix model, a corresponding pair of NLM image and histopathologically stained H&E image was required. Therefore, image registration of the histopathologically stained H&E image was crucial. On the other hand, the cycle CGAN model did not require the corresponding pair of the NLM image and histopathologically stained H&E image. Thus, the effort of image registration and pathological staining was reduced. The qualitative and quantitative evaluation of both models showed comparable results using evaluation metrics based on color, texture and perceptual quality. The evaluation metric like mean squared error reported values $>5 \times 10^3$ and $>8 \times 10^3$ for the Pix2Pix and the cycle CGAN model, respectively. In contrast, the evaluation metric, including SSIM and CSS reported values >0.50 and >0.90 for both models, respectively. In addition to quantitative evaluation, various pre- and post-processing procedures were explored in this work, however more advanced post-processing procedures could be investigated in future. Furthermore, a cycle CGAN model that can perform multiple staining using a NLM image can be one of the future research directions. The cycle CGAN model can also be investigated for additional benefits like the artificial generation of NLM images, increasing the spatial resolution of the computationally stained H&E images and removing fluorescence effect from the reconstructed NLM images. Overall, the results showed several benefits of using computational staining of NLM images than performing histopathological staining in laboratories. Thus, the computational staining approach should be encouraged in clinics to benefit the pathological and clinical field of science.

Code availability

https://github.com/Bocklitz-Lab/Pseudo_HE_modelling.git

Funding. Thüringer Ministerium für Wirtschaft, Wissenschaft und Digitale Gesellschaft (DigLeben-5575/10-9); Deutsche Forschungsgemeinschaft (BO 4700/1-1, BO 4700/4-1, PO 563/30-1, STA 295/11-1); Leibniz Association (Open Access Fund).

Disclosures. The authors declare no conflicts of interest.



Supplemental document. See [Supplement 1](#) for supporting content.

References

1. T. Bocklitz, A. Silge, H. Bae, M. Rodewald, F. B. Legesse, T. Meyer, and J. Popp, "Non-invasive imaging techniques: From histology to in vivo imaging," in *Molecular Imaging in Oncology* (Springer, 2020), pp. 795–812.
2. N. Vogler, S. Heuke, T. W. Bocklitz, M. Schmitt, and J. Popp, "Multimodal imaging spectroscopy of tissue," *Annu. Rev. Anal. Chem.* **8**(1), 359–387 (2015).
3. R. Cicchi and F. S. Pavone, "Multimodal nonlinear microscopy: A powerful label-free method for supporting standard diagnostics on biological tissues," *J. Innovative Opt. Health Sci.* **07**(05), 1330008 (2014).
4. S. Heuke, N. Vogler, T. Meyer, D. Akimov, F. Kluschke, H. R wert-Huber, J. Lademann, B. Dietzek, and J. Popp, "Multimodal mapping of human skin," *Br. J. Dermatol.* **169**(4), 794–803 (2013).
5. S. Heuke, N. Vogler, T. Meyer, D. Akimov, F. Kluschke, H.-J. R wert-Huber, J. Lademann, B. Dietzek, and J. Popp, "Detection and discrimination of non-melanoma skin cancer by multimodal imaging," in *Healthcare*, vol. 1(1) (Multidisciplinary Digital Publishing Institute, 2013), pp. 64–83.
6. S. Guo, S. Pfeifenbring, T. Meyer, G. Ernst, F. von Eggeling, V. Maio, D. Massi, R. Cicchi, F. S. Pavone, J. Popp, and T. Bocklitz, "Multimodal image analysis in tissue diagnostics for skin melanoma," *J. Chemom.* **32**(1), e2963 (2018).
7. S. Heuke, O. Chernavskaia, T. Bocklitz, F. B. Legesse, T. Meyer, D. Akimov, O. Dirsch, G. Ernst, F. von Eggeling, I. Petersen, O. Guntinas-Lichius, M. Schmitt, and J. Popp, "Multimodal nonlinear microscopy of head and neck carcinoma – toward surgery assisting frozen section analysis," *Head Neck* **38**(10), 1545–1552 (2016).
8. T. Meyer, O. Guntinas-Lichius, F. von Eggeling, G. Ernst, D. Akimov, M. Schmitt, B. Dietzek, and J. Popp, "Multimodal nonlinear microscopic investigations on head and neck squamous cell carcinoma: Toward intraoperative imaging," *Head Neck* **35**(9), E280–E287 (2013).
9. T. Meyer, N. Bergner, C. Krafft, D. Akimov, B. Dietzek, J. Popp, C. Bielecki, B. F. Romeike, R. Reichart, and R. Kalf, "Nonlinear microscopy, infrared, and Raman microspectroscopy for brain tumor analysis," *J. Biomed. Opt.* **16**(2), 021113 (2011).
10. S. Sch rmann, S. Foersch, R. Atreya, H. Neumann, O. Friedrich, M. F. Neurath, and M. J. Waldner, "Label-free imaging of inflammatory bowel disease using multiphoton microscopy," *Gastroenterology* **145**(3), 514–516 (2013).
11. D. A. Orringer, B. Pandian, Y. S. Niknafs, T. C. Hollon, J. Boyle, S. Lewis, M. Garrard, S. L. Hervey-Jumper, H. J. L. Garton, C. O. Maher, J. A. Heth, O. Sagher, D. A. Wilkinson, M. Snuderl, S. Venneti, S. H. Ramkissoon, K. A. McFadden, A. Fisher-Hubbard, A. P. Lieberman, T. D. Johnson, X. S. Xie, J. K. Trautman, C. W. Freudiger, and S. Camelo-Piragua, "Rapid intraoperative histology of unprocessed surgical specimens via fibre-laser-based stimulated Raman scattering microscopy," *Nat. Biomed. Eng.* **1**(2), 0027 (2017).
12. T. W. Bocklitz, F. S. Salah, N. Vogler, S. Heuke, O. Chernavskaia, C. Schmidt, M. J. Waldner, F. R. Greten, R. Br uer, and M. Schmitt, "Pseudo-HE images derived from CARS/TPEF/SHG multimodal imaging in combination with Raman-spectroscopy as a pathological screening tool," *BMC Cancer* **16**(1), 534 (2016).
13. I. J. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," *ArXiv abs/1701.00160*, (2017).
14. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2017), pp. 1125–1134.
15. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, (2017), pp. 2223–2232.
16. M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784* (2014).
17. Y. Ma, X. Chen, W. Zhu, X. Cheng, D. Xiang, and F. Shi, "Speckle noise reduction in optical coherence tomography images based on edge-sensitive cGAN," *Biomed. Opt. Express* **9**(11), 5129–5146 (2018).
18. R. Zheng, L. Liu, S. Zhang, C. Zheng, F. Bunyak, R. Xu, B. Li, and M. Sun, "Detection of exudates in fundus photographs with imbalanced learning using conditional generative adversarial network," *Biomed. Opt. Express* **9**(10), 4863–4878 (2018).
19. C. Zhang, K. Wang, Y. An, K. He, T. Tong, and J. Tian, "Improved generative adversarial networks using the total gradient loss for the resolution enhancement of fluorescence images," *Biomed. Opt. Express* **10**(9), 4742–4756 (2019).
20. H. Zhang, C. Fang, X. Xie, Y. Yang, W. Mei, D. Jin, and P. Fei, "High-throughput, high-resolution deep learning microscopy based on registration-free generative adversarial network," *Biomed. Opt. Express* **10**(3), 1044–1063 (2019).
21. J. Ouyang, T. S. Mathai, K. Lathrop, and J. Galeotti, "Accurate tissue interface segmentation via adversarial pre-segmentation of anterior segment OCT images," *Biomed. Opt. Express* **10**(10), 5291–5324 (2019).
22. H. Jiang, X. Chen, F. Shi, Y. Ma, D. Xiang, L. Ye, J. Su, Z. Li, Q. Chen, Y. Hua, X. Xu, W. Zhu, and Y. Fan, "Improved cGAN based linear lesion segmentation in high myopia ICGA images," *Biomed. Opt. Express* **10**(5), 2355–2366 (2019).
23. K. J. Halupka, B. J. Antony, M. H. Lee, K. A. Lucy, R. S. Rai, H. Ishikawa, G. Wollstein, J. S. Schuman, and R. Garnavi, "Retinal optical coherence tomography image enhancement via deep learning," *Biomed. Opt. Express* **9**(12), 6205–6221 (2018).
24. Y. Rivenson, H. Wang, Z. Wei, Y. Zhang, H. Gunaydin, and A. Ozcan, "Deep learning-based virtual histology staining using auto-fluorescence of label-free tissue," *Nat. Biomed. Eng.* **3**, 466–477 (2018).

25. N. Bayramoglu, M. Kaakinen, L. Eklund, and J. Heikkilä, "Towards virtual H&E staining of hyperspectral lung histology images using conditional generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, (2017), pp. 64–71.
26. T. Liu, Z. Wei, Y. Rivenson, K. de Haan, Y. Zhang, Y. Wu, and A. Ozcan, "Deep learning-based color holographic microscopy," *J. Biophotonics* **12**(11), e201900107 (2019).
27. E. Nehme, L. E. Weiss, T. Michaeli, and Y. Shechtman, "Deep-storm: super-resolution single-molecule microscopy by deep learning," *Optica* **5**(4), 458–464 (2018).
28. H. Wang, Y. Rivenson, Y. Jin, Z. Wei, R. Gao, H. Günaydin, L. A. Bentolila, C. Kural, and A. Ozcan, "Cross-modality deep learning achieves super-resolution in fluorescence microscopy," in *2019 Conference on Lasers and Electro-Optics (CLEO) (IEEE, 2019)*, pp. 1–2.
29. Z. Xu, C. F. Moro, B. Bozóky, and Q. Zhang, "GAN-based virtual re-staining: a promising solution for whole slide image analysis," arXiv preprint arXiv:1901.04059 (2019).
30. O. Chernavskaia, S. Heuke, M. Vieth, O. Friedrich, S. Schürmann, R. Atreya, A. Stallmach, M. F. Neurath, M. Waldner, and I. Petersen, "Beyond endoscopic assessment in inflammatory bowel disease: real-time histology of disease activity by non-linear multimodal imaging," *Sci. Rep.* **6**(1), 29239 (2016).
31. A. Roche, G. Malandain, X. Pennec, and N. Ayache, "The correlation ratio as a new similarity measure for multimodal image registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (Springer, 1998), pp. 1115–1124.
32. P. Pradhan, T. Meyer, M. Vieth, A. Stallmach, M. Waldner, M. Schmitt, J. Popp, and T. Bocklitz, "Semantic segmentation of non-linear multimodal images for disease grading of inflammatory bowel disease: A segnet-based application," in *International Conference on Pattern Recognition Applications and Methods 2019*, (2019).
33. P. Pradhan, S. Guo, O. Ryabchykov, J. Popp, and T. W. Bocklitz, "Deep learning a boon for biophotonics?" *J. Biophotonics* **13**, e201960186 (2020).
34. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, (Springer, 2015), pp. 234–241.
35. H. Cho, S. Lim, G. Choi, and H. Min, "Neural stain-style transfer learning using GAN for histopathological images," arXiv preprint arXiv:1710.08543 (2017).
36. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 (2014).
37. H. Ren, J. Li, and N. Gao, "Automatic sketch colorization with tandem conditional adversarial networks," in *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 1 (IEEE, 2018), pp. 11–15.
38. Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *Proceedings (ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 2 (IEEE, 2005), pp. ii/573–ii/576 Vol. 2.
39. D. Wang, L. Shi, Y. J. Wang, G. C. Man, P. A. Heng, J. F. Griffith, and A. T. Ahuja, "Color quantification for evaluation of stained tissues," *Cytometry, Part A* **79A**(4), 311–316 (2011).
40. M. T. McCann, J. A. Ozolek, C. A. Castro, B. Parvin, and J. Kovacevic, "Automated histology analysis: Opportunities for signal processing," *IEEE Signal Process. Mag.* **32**(1), 78–87 (2015).
41. N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," in *2016 23rd International conference on pattern recognition (ICPR)*, (IEEE, 2016), pp. 2440–2445.
42. A. BenTaieb and G. Hamarneh, "Adversarial stain transfer for histopathology image analysis," *IEEE Trans. Med. Imaging* **37**(3), 792–802 (2018).
43. R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using MATLAB* (Prentice-Hall, Inc., USA, 2003).
44. J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision* (Springer, 2016), pp. 694–711.
45. D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2017), pp. 417–425.
46. A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," CoRR [abs/1511.06434](https://arxiv.org/abs/1511.06434) (2016).
47. T. Wang and Y. Lin, "CycleGAN with better cycles," (2018).

Computational tissue staining of non-linear multimodal imaging using supervised and unsupervised deep learning: supplement

PRANITA PRADHAN,^{1,2}  TOBIAS MEYER,² MICHAEL VIETH,³
ANDREAS STALLMACH,⁴ MAXIMILIAN WALDNER,^{5,6} MICHAEL
SCHMITT,¹ JUERGEN POPP,^{1,2} AND THOMAS BOCKLITZ^{1,2,*} 

¹*Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller-University, Jena, Germany*

²*Leibniz Institute of Photonic Technology, Member of Leibniz Health Technologies Jena, Germany*

³*Institute of Pathology, Klinikum Bayreuth, Bayreuth, Germany*

⁴*Department of Internal Medicine IV (Gastroenterology, Hepatology, and Infectious Diseases), Jena University Hospital, Jena, Germany*

⁵*Erlangen Graduate School in Advanced Optical Technologies (SAOT), Friedrich-Alexander University of Erlangen-Nuremberg, 91052 Erlangen, Germany*

⁶*Medical Department 1, Friedrich-Alexander University of Erlangen-Nuremberg, Erlangen, Germany*

**thomas.bocklitz@uni-jena.de*

This supplement published with The Optical Society on 23 March 2021 by The Authors under the terms of the [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/) in the format provided by the authors and unedited. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Supplement DOI: <https://doi.org/10.6084/m9.figshare.14046467>

Parent Article DOI: <https://doi.org/10.1364/BOE.415962>

Computational tissue staining of non-linear multimodal imaging using supervised and unsupervised deep learning

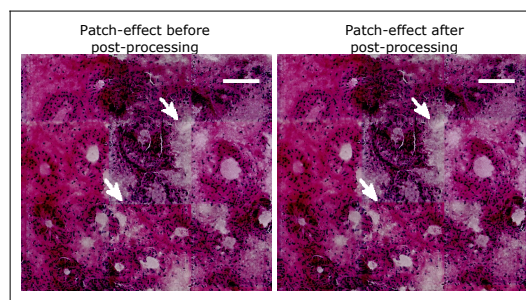


Fig. S1. This figure shows the effect of post-processing for removing 'patch-effect'. The 'patch-effect' was removed by interpolating pixel values of three neighbouring pixels at the end of every patch (256th pixel). This effect (shown in white arrows) was visible for few images and its removal did not significantly affect the performance metrics. The scale bar represents 100 μm

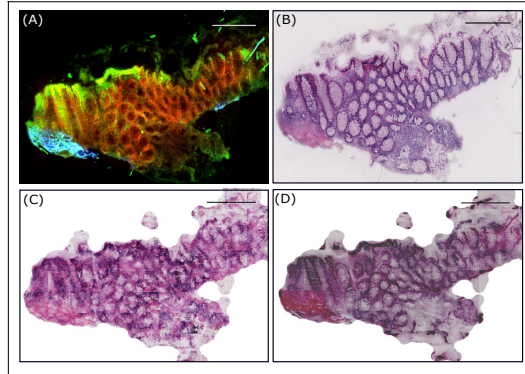


Fig. S2. (A) shows a non-linear multimodal image from test dataset, (B) visualizes corresponding histopathologically stained H&E image (unregistered), (C) shows the computational H&E image by the Pix2Pix (MSE = 4.4×10^3 , SSIM = 0.65, CSS = 0.94) and (D) depicts computational H&E image by the cycle CGAN model (MSE = 8.4×10^3 , SSIM = 0.63, CSS = 0.94). The contrast of the computational H&E image in (D) is reduced by a factor of 0.7. The images here are downsampled to 20% of original size for clarity. The scale bar represents $100 \mu\text{m}$.

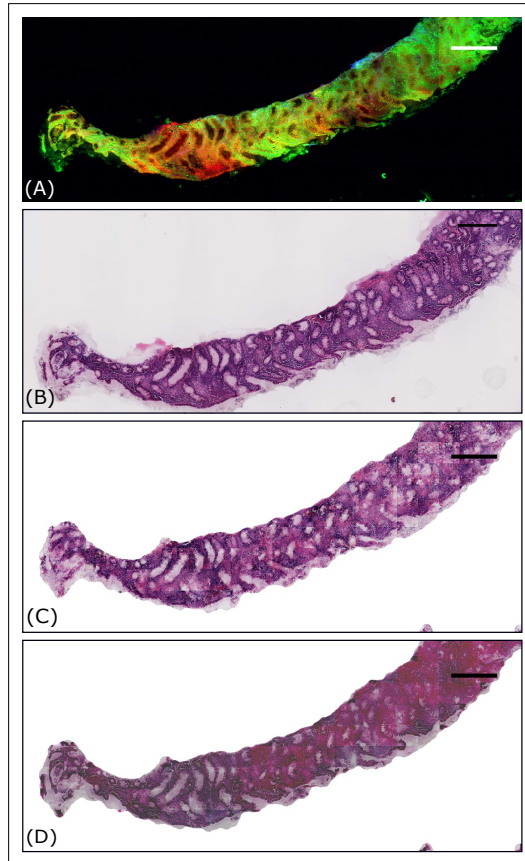


Fig. S3. (A) shows a non-linear multimodal image from training dataset, (B) visualizes corresponding histopathologically stained H&E image (unregistered), (C) shows the computational H&E image by the Pix2Pix (MSE = 2.8×10^3 , SSIM = 0.74, CSS = 0.96) and (D) depicts computational H&E image by the cycle CGAN model (MSE = 5.9×10^3 , SSIM = 0.72, CSS = 0.94). The contrast of the computational H&E image in (D) is reduced by a factor of 0.7. The images here are downsampled to 20% of original size for clarity. The scale bar represents 100 μm .

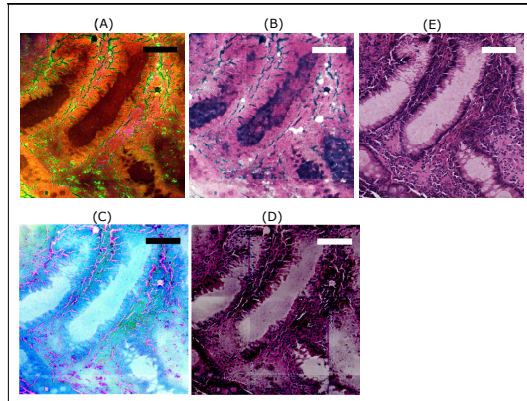


Fig. S4. This figure shows effect of 'contrast inversion' on the computational H&E images of the cycle CGAN model. (A) shows an original MM patch given to the cycle CGAN model and (B) visualizes the generated H&E patch. Here, we see that the colors in the generated H&E image are inverted when compared to the pathologically stained H&E patch shown in (E). Contrarily, (C) shows a contrast inverted MM patch and (D) depicts the generated H&E patch by the cycle CGAN model. Here, the color contrast is similar to the pathologically stained H&E patch shown in (E). The scale bar represents 100 μm .

Table S1. The quantitative metrics namely mean squared error (MSE), structure similarity index (SSIM) and color similarity index (CSS) evaluated for 19 images from the training and testing dataset is given for the Pix2Pix and the cycle CGAN models, respectively.

Image	Pix2Pix			Cycle CGAN		
	MSE	SSIM	CSS	MSE	SSIM	CSS
Train 1	3601.43	0.61	0.94	9373.09	0.57	0.92
Train 2	2800.83	0.74	0.96	5890.62	0.72	0.94
Train 3	4984.19	0.38	0.89	13955.34	0.33	0.89
Train 4	3337.39	0.56	0.93	8157.08	0.53	0.93
Train 5	4451.48	0.51	0.92	11509.74	0.47	0.90
Train 6	5100.78	0.49	0.92	9629.29	0.46	0.89
Train 7	4956.57	0.41	0.91	10389.17	0.38	0.90
Train 8	4167.94	0.52	0.92	8376.58	0.49	0.90
Train 9	7938.24	0.46	0.91	15823.56	0.41	0.91
Train 10	7657.41	0.39	0.88	9725.03	0.36	0.88
Train 11	3066.22	0.67	0.94	6749.37	0.65	0.93
Train 12	3057.30	0.65	0.94	8129.79	0.62	0.93
Train 13	5918.03	0.36	0.90	15783.79	0.30	0.90
Test 1	5617.82	0.60	0.92	6600.03	0.59	0.91
Test 2	3582.64	0.61	0.94	6633.59	0.60	0.94
Test 3	3827.09	0.65	0.94	4767.95	0.63	0.94
Test 4	3292.67	0.60	0.94	10118.23	0.58	0.93
Test 5	4416.22	0.64	0.94	8408.46	0.63	0.94
Test 6	4937.68	0.49	0.91	10265.39	0.48	0.93





A2 DATA FUSION OF HISTOLOGICAL AND IMMUNOHISTOCHEMICAL IMAGE DATA
FOR BREAST CANCER DIAGNOSTICS USING TRANSFER LEARNING

Submitted on: 14 September 2020

Accepted on: 13 November 2020

Published on: 10 February 2021

Data Fusion of Histological and Immunohistochemical Image Data for Breast Cancer Diagnostics using Transfer Learning

Pranita Pradhan^{1,2}, Katharina Köhler^{3,4}, Shuxia Guo^{1,2}, Olga Rosin^{3,4}, Jürgen Popp^{1,2}, Axel Niendorf^{3,4} and Thomas Wilhelm Bocklitz^{*,1,2}

¹*Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University, Helmholtzweg 4, Jena, 07743, Thüringen, Germany*

²*Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, Jena, 07745, Thüringen, Germany*

³*MVZ Prof. Dr. med. A. Niendorf Pathologie Hamburg-West GmbH, Lornsenstraße 4-6, Hamburg, 22767, Hamburg, Germany*

⁴*Institute for Histology, Cytology and Molecular Diagnostics, Lornsenstraße 4, Hamburg, 22767, Hamburg, Germany*


Keywords: Breast Cancer, Transfer Learning, Histology, Immunohistochemistry.


Abstract: A combination of histological and immunohistochemical tissue features can offer better breast cancer diagnosis as compared to histological tissue features alone. However, manual identification of histological and immunohistochemical tissue features for cancerous and healthy tissue requires an enormous human effort which delays the breast cancer diagnosis. In this paper, breast cancer detection using the fusion of histological (H&E) and immunohistochemical (PR, ER, Her2 and Ki-67) imaging data based on deep convolutional neural networks (DCNN) was performed. DCNNs, including the VGG network, the residual network and the inception network were comparatively studied. The three DCNNs were trained using two transfer learning strategies. In transfer learning strategy 1, a pre-trained DCNN was used to extract features from the images of five stain types. In transfer learning strategy 2, the images of the five stain types were used as inputs to a pre-trained multi-input DCNN, and the last layer of the multi-input DCNN was optimized. The results showed that data fusion of H&E and IHC imaging data could increase the mean sensitivity at least by 2% depending on the DCNN model and the transfer learning strategy. Specifically, the pre-trained inception and residual networks with transfer learning strategy 1 achieved the best breast cancer detection.


1 INTRODUCTION


Breast cancer is one of the most prevalent cancers among women. It is diagnosed by a routine procedure which is based on morphological tissue features in hematoxylin and eosin (H&E) stained tissue sections (figure 1a). The morphological tissue features include tumour size and type, which are regularly documented to assess the histological grade of breast cancer tissue (Webster et al., 2005). These morphological tissue features are also used to prevent recurrence risk of breast cancer and prescribe personalized therapies. Breast cancer is additionally verified by other staining technique called the immunohistochemical

(IHC) staining technique. The IHC staining technique uses antibodies to highlight specific antigens in the tissue region (Veta et al., 2014), and includes estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor-2 (Her2) (figure 1b-d). Studies have shown that the IHC examination with ER, PR, Her2 and Ki-67 can detect five molecular breast cancer sub-types to provide adequate personalized therapies (Perou et al., 2000; Sørliet et al., 2001; Cheang et al., 2009). However, none of the studies report a combination of histology (H&E) and IHC staining techniques (ER, PR, Her2 and Ki-67) for breast cancer diagnosis. Therefore, in this work, an integration of IHC imaging technique i.e. hormone receptors including ER, PR, Her2 and Ki-67 nuclear protein stained images with H&E stained images is proposed to gain new insights into breast cancer biology (Elledge et al., 2000; Damodaran and Olson, 2012). The combination of histology and IHC stain-

^a <https://orcid.org/0000-0002-0558-2914>

^b <https://orcid.org/0000-0001-8237-8936>

^c <https://orcid.org/0000-0003-4257-593X>

^d <https://orcid.org/0000-0003-2778-6624>

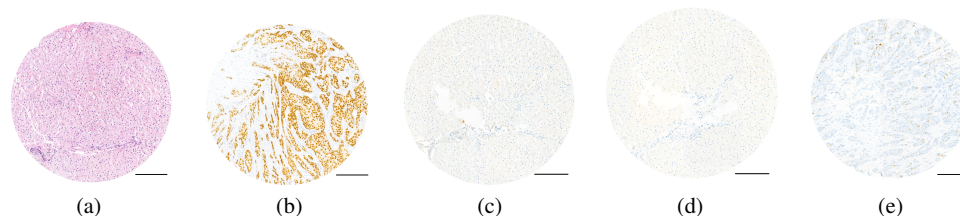


Figure 1: Five stain type images. (a) Hematoxylin and eosin (H&E), (b) Estrogen receptor (ER), (c) Progesterone receptor (PR), (d) Human epidermal growth factor-2 (Her2) and (e) Ki-67 protein are shown. Scale bar is 200 μm .

ing technique is referred to as ‘Data fusion’ approach.

Data fusion approach by combining the histological and IHC stained images can provide various tissue features associated with the disease stage and relapse of breast cancer. However, visual inspection of all five stained images is a tedious process which can prolong the diagnosis. Therefore, automation of breast cancer detection based on the combination of histological and IHC imaging data is needed. In this regard, researchers (Pham et al., 2007; Dobson et al., 2010) used computer-assisted image analysis techniques to automatically monitor changes in the tissue features of histological and IHC stained images separately. However, computer-assisted image analysis can be limited due to the need for specific software systems or the need for user-specific input to analyze the images. This slows down the process of analyzing images and providing personalized therapies to the patients. To increase the analysis speed and reduce human intervention, this work proposes machine learning (ML) instead of computer-assisted image analysis techniques.

Conventional ML methods can automatize breast cancer detection based on the fusion of histological and IHC imaging data in the following way. First, the features (e.g. color, shape and texture features) from the five stain type of imaging data (H&E, ER, PR, Her2 and Ki-67) can be extracted using image analysis methods. The feature extraction step in the conventional ML method is subjective and requires the effort of an image analyst. Based on the extracted features, a classification, or a regression model can be constructed. Subsequently, the classification or the regression model can be used to make ‘predictions’ (i.e. to predict a class like tumour or normal) on a new or unseen dataset. Thus, the extracted features affect the predictions made by the ML model. However, recently developed ML methods are capable of performing automatic feature extraction for classification or regression purpose. These self-learning methods are categorized into a broad family of ML called ‘Deep learning’ (DL). The DL models can have many types of network architectures. Widely used

DL model for images is the deep convolutional neural network (DCNN) and its numerous applications are reported in the field of digital pathology (Liu et al., 2017); for example, cell segmentation or detection (Chen and Chef’Hotel, 2014), tumour classification (Cireřan et al., 2013; Wang et al., 2016) and carcinoma localization (Janowczyk and Madabhushi, 2016; Coudray et al., 2018; Khosravi et al., 2018; Sheikhzadeh et al., 2018). Nevertheless, a bottleneck for DL models is the requirement of huge dataset during training, which is difficult to acquire, particularly in the medical imaging field. In such cases, ‘transfer learning’ methods for DCNNs can be applied for improving the model performance (Tajbakhsh et al., 2016).

Transfer learning is the transfer of knowledge learned on a source task using a source dataset to improve the performance on a target task using the target dataset (Torrey and Shavlik, 2010). Transfer learning using any DL model like DCNN can be performed by three strategies. First, a pre-trained DCNN can be used as a feature extractor. In this strategy, features for the target dataset are extracted using a DCNN trained on different or similar source dataset. The second strategy is fine-tuning the weights of the last layers of a pre-trained DCNN, and the third strategy is fine-tuning the weights of all layers of a pre-trained DCNN. In the second and third fine-tuning strategies, the weights of specific layers of a DCNN trained on a source dataset are further optimized based on the target dataset. The three transfer learning strategies like using a DCNN as a feature extractor or fine-tuning of a DCNN, requires adequate knowledge of the size and type of the source and the target dataset (Pan and Yang, 2010). Transfer learning, if used appropriately, can improve the initial and final performance of the DL model on the target dataset. It can also reduce the total training time of the DL model on the target dataset. Different transfer learning strategies acquire different results based on the source and target dataset which is evident in the next section.

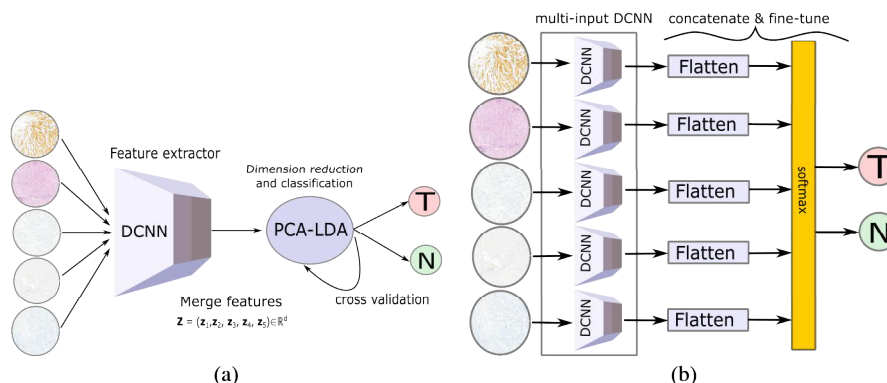


Figure 2: (a) visualizes transfer learning strategy 1 for data fusion approach where a pre-trained DCNN is used as a feature extractor. The features extracted from a pre-trained DCNN for all five stain type images are merged and classified into tumour and normal using the PCA-LDA model. (b) shows transfer learning strategy 2 for data fusion approach where fine-tuning of the last layer of a pre-trained multi-input DCNN is performed. The five DCNNs are pre-trained models like the VGG16, the Inceptionv3 or the ResNet50, each having a stain type as its input.

2 RELATED WORK

Transfer learning in medical imaging can be achieved by training a DCNN on a large medical or non-medical dataset, and transferring its knowledge to the target medical dataset (Bayramoglu and Heikkilä, 2016; Tajbakhsh et al., 2016). A recent study used a large non-medical dataset like the ImageNet dataset (Russakovsky et al., 2015) to pre-train a DCNN and transfer its off-the-shelf features to investigate two computer-aided detection (CADs) problems namely thoracoabdominal lymph node detection and interstitial lung disease detection (Shin et al., 2016). In their work, three different DCNNs including the CifarNet (Krizhevsky and Hinton, 2009), the AlexNet (Krizhevsky et al., 2012) and the GoogleNet (Szegedy et al., 2015) were evaluated with three transfer learning strategies. Similarly, a recent publication (Mormont et al., 2018) compared various transfer learning strategies based on pre-trained DCNNs using eight classification datasets in digital pathology. Their results showed that fine-tuning the ResNet (He et al., 2016) and the DenseNet (Huang et al., 2017) models outperformed the other tested models in the morphological classification task. Similar findings were observed in other references (Antony et al., 2016; Kieffer et al., 2017; Ravishankar et al., 2016).

In contrast to the previously mentioned applications where fine-tuning of a DCNN achieved the best performance, several other applications using a DCNN as feature extractor achieved significant performance on binary and multi-class classification tasks. These applications included prediction of mor-

phological changes in cells in microscopic images (Kensert et al., 2018), classification of colon polyps in endoscopic images (Ribeiro et al., 2016), identification of mammographic tumours (Huynh et al., 2016) and detection of pulmonary nodules in computed tomography scans (Van Ginneken et al., 2015). It is clear from the previous researches that transfer learning techniques are data-dependent, and a generalization of the above-mentioned results is not feasible, especially in the medical imaging field (Litjens et al., 2017). Therefore, no consensus of the proper application of transfer learning in the medical imaging field is established. Likewise, the application of transfer learning, especially for medical imaging data requires utmost care and further investigations.

In this contribution, data fusion of histological and immunohistochemical imaging data for classifying breast cancer is presented for the first time. Due to our small dataset size, the classification task is performed using two transfer learning strategies. From previous experience, the third transfer learning strategy i.e. the training of a DCNN from scratch is avoided, as it is computationally expensive and may lead to overfitting in the absence of large datasets. The performance of the two transfer learning strategies for the data fusion approach is compared with histological imaging data. Moreover, the two transfer learning strategies are performed using three pre-trained DCNN models like the VGG16 (He et al., 2016), the Inceptionv3 (Szegedy et al., 2016) and the ResNet50 network (Simonyan and Zisserman, 2014). The goal of this study was to verify whether the data fusion approach along with

transfer learning improves the breast cancer diagnosis based on the sensitivity and F1 score metric.

3 MATERIAL AND METHODS

3.1 Sample Preparation

A Tissue Microarray (TMA) with 97 cores representing 23 breast cancer cases (78 tumour cores, 18 non-cancerous tissue cores or the normal breast tissue and one control core of liver tissue) was produced using the Manual Tissue Arrayer MTA-1 by Estigen. The cases were randomly selected out of the daily routine of MVZ Prof. Dr. med. A. Niendorf Pathologie Hamburg-West GmbH and anonymized according to a statement of the ethics committee of the Hamburg Medical Chamber. Core tissue biopsies (1.0 mm in diameter) were taken from individual FFPE (formalin-fixed paraffin-embedded) blocks and arranged within a new recipient block. From the block, $2\ \mu\text{m}$ sections were cut, placed on glass microscope slides and H&E staining (figure 1a) following a standard protocol was performed. Digital images of histology (H&E) slides were obtained at $40\times$ magnification using the 3DHis-tech Panoramic 1000 Flash IV slide scanner with a spatial resolution of $0.24\ \mu\text{m}/\text{pixel}$ (.mrxs image file). Subsequently, immunohistochemistry staining (ER, PR, Her2 and Ki-67) (figure 1b-e) was performed on super frost charged glass slides.

3.2 Image Preprocessing

For the analysis, 96 TMAs or scans (78 tumour scans and 18 normal scans) from 23 patients were used, and each TMA had five stain types (H&E, PR, ER, Her2 and Ki-67). The pixel intensity I of each TMA was standardized using a min-max scaling $(I - I_{min}) / (I_{max} - I_{min})$, where I_{min} and I_{max} is the minimum and maximum intensity of a pixel in a TMA. The background pixels were cropped manually and non-overlapping patches of size 1024×1024 were extracted from a standardized TMA. This led to 9 patches per TMA (702 tumour and 162 normal patches). The four corner patches including a large number of background pixels were removed, leading to 390 tumour and 90 normal patches. Based on the 480 selected patches, three pre-trained models were used with two transfer learning strategies.

3.3 DCNN Architectures

To check the robustness of the data fusion approach, three DCNNs: the VGG network, the Inception net-

work and the residual network, with unique architectures were chosen. The VGG network is a DCNN that has acquired state-of-the-art performances for image classification tasks. However, the VGG network can exhibit the problem of vanishing gradients with an increasing number of layers (Hanin, 2018). Thus, the residual network which can solve the problem of vanishing gradients by adding the 'shortcut connections' was explored in this work. Furthermore, the inception network that provides width in addition to the depth to a conventional DCNN was utilized. A detailed explanation of the architecture of the three models is given further.

3.3.1 VGG Network

A VGG network is a DCNN with different configurations from 11 to 16 convolutional layers followed by three fully connected layers. The number of convolutional layers increases the depth of the VGG network. It is shown that an increase in the depth of the VGG network decreases the top-5 validation error (He et al., 2016). However, the decrease in the error for the VGG network from 16 to 19 convolutional layers is not significant. Thus, the VGG network with 16 convolutional layers referred to as VGG16 from Keras was used (Chollet et al., 2015). The input to the VGG16 network was an RGB image of size 224×224 , and each image was preprocessed by subtracting the mean RGB values computed over the training dataset.

3.3.2 Inception Network

Deep networks like VGG network require an appropriate selection of the number of convolution filters and filter sizes. For this reason, the inception network concatenates convolutional layers of different filter size, including the spatial dimension of 1×1 , 3×3 and 5×5 . This captures information at various scales while increasing the computational complexity. In order to reduce the computational cost, a convolutional layer of 1×1 filter size is applied before each convolutional layer of filter size 3×3 and 5×5 . These two salient features of the Inception network reduce the dimensionality in the feature space and thereby allows the network to be deeper and wider. Moreover, the inception network replaces the fully connected layer with global averaging layers which reduces the number of trainable weights, thus reducing over-fitting on the training dataset (Szegedy et al., 2016). The Inceptionv3 implementation from Keras, which has 95 layers and requires an RGB image as input with size 299×299 was used.

Table 1: This table shows confusion matrices, mean sensitivities and mean F1 scores for the VGG, the Inception and the residual networks using transfer learning strategy 1. Here, two feature sets extracted from pre-trained models are used; one feature set is extracted from H&E images only, while the other feature set is extracted from all the five stain types. All metrics are computed for 96 TMAs by taking majority voting of the predictions acquired for the patches using the PCA-LDA model. N: normal scans, T: tumour scans.

Data fusion (H&E+IHC imaging data)						Only histological imaging data					
DCNN		N	T	Sens (%)	F1 (%)	DCNN		N	T	Sens (%)	F1 (%)
VGG16	N	13	5	79.06	76.24	VGG16	N	14	4	80.56	76.61
	T	11	67				T	13	65		
Inceptionv3	N	16	2	89.32	85.47	Inceptionv3	N	15	3	88.46	86.97
	T	8	70				T	5	75		
ResNet50	N	14	4	86.97	87.80	ResNet50	N	14	4	85.68	84.96
	T	3	75				T	5	73		

Table 2: This table shows confusion matrices, mean sensitivities and mean F1 scores for the VGG, the Inception and the residual networks using transfer learning strategy 2. Data fusion approach used multi-input DCNN with the five stain type images as input, whereas a single-input DCNN was used only the H&E image as input. The last layers of both single-input and multi-input DCNNs were fine-tuned. The mean sensitivities are computed for 96 TMAs by taking majority voting of the predictions obtained for the patches. N: normal scans, T: tumour scans.

Data fusion (H&E+IHC imaging data)						Only histological imaging data					
DCNN		N	T	Sens (%)	F1 (%)	DCNN		N	T	Sens (%)	F1 (%)
VGG16	N	7	11	66.88	70.86	VGG16	N	3	15	55.13	57.57
	T	4	74				T	5	73		
Inceptionv3	N	0	18	50.00	44.83	Inceptionv3	N	9	9	72.44	75.66
	T	0	78				T	4	74		
ResNet50	N	0	18	50.00	44.83	ResNet50	N	12	6	81.41	83.78
	T	0	78				T	3	75		

Table 3: This table shows confusion matrices, mean sensitivities and mean F1 scores for the VGG, the Inception and the residual network using the two transfer learning strategies. All metrics are computed for 96 TMAs by taking majority voting of the predictions acquired by the models for patches.

Transfer learning strategy 1						Transfer learning strategy 2					
DCNN		N	T	Sens (%)	F1 (%)	DCNN		N	T	Sens (%)	F1 (%)
VGG16	N	13	5	79.06	76.24	VGG16	N	7	11	66.88	70.86
	T	11	67				T	4	74		
Inceptionv3	N	16	2	89.32	85.47	Inceptionv3	N	0	18	50.00	44.83
	T	8	70				T	0	78		
ResNet50	N	14	4	86.97	87.80	ResNet50	N	0	18	50.00	44.83
	T	3	75				T	0	78		

3.3.3 Residual Network

The configurations of the VGG network show that deep neural networks achieve good top-5 accuracy until a certain depth limit (He et al., 2016). An increase in the network depth causes a problem of vanishing or exploding gradients (Hanin, 2018) which affects the network convergence and degrades the performance (Simonyan and Zisserman, 2014). Therefore, the residual networks are built to solve this degradation problem by adding activations of the top layers into the deeper layers of the network. For instance, in a deep neural network the activation a of the $(l+2)^{th}$ layer with weight w and bias b is given as

$$a_{(l+2)} = f[(w_{(l+2)} \times a_{(l+1)}) + b_{(l+2)}], \quad (1)$$

where f is an activation function like linear rectified unit ($f = \max(a_{(l+2)}, 0)$). However, in a residual block the activation a of the l^{th} layer (or an identity mapping) is added via the ‘skip or shortcut connections’ (Bishop et al., 1995; Venables and Ripley, 2013) to the $(l+2)^{th}$ layer of the network. Therefore, the activation of the $(l+2)^{th}$ layer in a residual block can be given as

$$a_{(l+2)} = f[(w_{(l+2)} \times a_{(l+1)}) + b_{(l+2)} + a_{(l)}]. \quad (2)$$

This implies that in worse cases when the network

fails to learn representative features, i.e. $w_{(l+2)} = 0$ and $b_{(l+2)} = 0$, the output still remains an identity mapping of the input a_l . In residual networks, a series of residual blocks along with intermediate normalization layers was used; thus improving the learning of the deep neural networks. In this work, the ResNet50 implementation from Keras, which has 152 layers and requires an RGB image as an input with size 224×224 , was used.

The above explained three DCNN models were trained using two transfer learning strategies which are discussed in the next section.

3.4 Transfer Learning Strategies

The above-mentioned DCNNs were utilized for two transfer learning strategies. For the first strategy, a pre-trained DCNN model to extract off-the-shelf features followed by a linear classifier was used. In the second strategy, a multi-input pre-trained DCNN model followed by a softmax classifier was used. Both strategies were performed on a commercially available PC system intel® Core™ with NVIDIA GeForce GTX 1060, 6GB with python packages: Keras(Chollet et al., 2015), Tensorflow(Abadi et al., 2015), Scikit-learn (Pedregosa et al., 2011), Scipy (Jones et al., 2001) and Numpy (Oliphant, 2006).

3.4.1 DCNN as Feature Extractor

In the first strategy (figure 2a), features $\mathbf{z}_i \in \mathbb{R}^m, i = (1, 2, 3, 4, 5)$ were extracted for patches of each stain type i using the pre-trained VGG16, Inceptionv3 and ResNet50 networks. The patches were resized according to the model's input size requirement. For a patch of a single stain type, 25,088 features were extracted by the VGG16 (feature shape: 1, 7, 7, 512), 51,200 features were calculated by the Inceptionv3 (feature shape: 1, 5, 5, 2048) and 2048 features were obtained by the ResNet50 (feature shape: 1, 1, 1, 2048). For data fusion approach, the features from all five stain types were concatenated, $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4, \mathbf{z}_5) \in \mathbb{R}^d$ ($d \gg m$) resulting in ~ 0.12 million features by the VGG16 model, ~ 0.25 million features by the Inceptionv3 model and 10,240 features by the ResNet50 model per patch. For histological imaging data, i.e. without the data fusion approach, the features extracted only from the H&E images were used. In both cases, the large feature dimension of each patch was reduced by principal component analysis (PCA) model, and classified as normal or tumour using linear discriminant analysis (LDA) model (Hastie et al., 2009). The PCA-LDA model was evaluated using internal and external cross-validation scheme explained elsewhere (Guo

et al., 2017). Shortly, the internal cross-validation was used to optimize the number of PC's of the PCA-LDA model. The external cross-validation was used to predict an independent test dataset based on the PCA-LDA model. The external cross-validation used leave-one-patient-out cross-validation, such that the patches acquired from TMAs of 23 patients were used at least once as an independent test dataset. The internal cross-validation used 10 fold cross-validation. The predictions by the PCA-LDA model acquired for the patches from the external cross-validation step were voted to assign each TMA into a tumour or normal class. Based on the predicted TMA labels (obtained after majority voting of the patches) and true TMA labels, metrics like confusion matrix, mean sensitivity and mean F1 score were reported. The mean sensitivity and the mean F1 score were calculated using an average of the mean sensitivities and the mean F1 scores for the tumour and normal class, respectively. Lastly, the transfer learning strategy 1 was performed for all the three DCNNs and their classification performance based on TMAs was compared.

3.4.2 Fine-tuning of DCNN

In the second strategy (figure 2b), for histological imaging data, a single-input DCNN was used; whereas for the data fusion approach, a multi-input DCNN was used. The multi-input DCNN model \mathcal{N} was constructed using five pre-trained models of the same architecture; for instance, five pre-trained VGG networks each using a stain type image as an input. The input to the multi-input DCNN model was the five stained images (H&E, ER, Her2, Ki-67 and PR). The last layer of the multi-input DCNN models was concatenated and followed by a dense layer with two outputs (corresponding to the normal and tumour class) with a softmax activation layer. The softmax activation layer mapped the non-normalized output of the model \mathcal{N} to the distribution of K probabilities and is defined as

$$P(\mathbf{r})_i = \frac{\exp(r_i)}{\sum_{j=1}^K \exp(r_j)}, \quad (3)$$

where $\mathbf{r} = (r_1, \dots, r_K)$ and $K = 2$ for a binary classification task. During the training process, the last two layers were fine-tuned using Adam optimizer (Kingma and Ba, 2014) with a learning rate 0.001 and mini-batch size of 5 patches. To allocate higher class weight for the minority class (here, the normal class), the weighted binary cross-entropy loss function

$$\mathcal{L} = - \sum_i^K \alpha_i y_i \log(P(\mathbf{r})_i) \quad (4)$$

was used, where $\alpha_i = \frac{1}{\#K_i}$, y_i , $P(\mathbf{r})_i$ are the weight, ground truth and the probability from the softmax activation layer of the i^{th} class in K , respectively. The model was evaluated using the mean sensitivity and the mean F1 score similar to transfer learning strategy 1.

For the evaluation of the single and multi-input DCNN, the dataset was divided into three parts: training, validation and testing. In every iteration, patches of one patient were used as an independent test dataset and the patches of remaining patients were used as training and validation dataset. To avoid any training bias, the training and validation datasets were randomly split patient-wise such that patches from 30% patients were used as validation dataset and the rest as the training dataset. In other words, during each iteration, patches of one patient were used as the test dataset, patches of 16 patients formed the training dataset and patches of remaining 6 patients belonged to the validation dataset. The combination of 16 and 6 patients in training and validation datasets were chosen randomly. The iterations were repeated until all 23 patients were used as an independent test dataset. Further, every iteration was executed for ten epochs, and validation sensitivity was monitored for early stopping of the model training. The model with best validation sensitivity was used for predicting the independent test dataset in that iteration. In this way, the patches of all 23 patients were used individually as an independent test dataset, and majority voting of the patches similar to transfer learning strategy 1 was performed. The confusion matrices and average of the mean sensitivities for the normal and tumour classes were evaluated using the independent test dataset. Subsequently, transfer learning strategy 2 was performed for all the three pre-trained DCNN models with the same hyper-parameter setting.

3.4.3 ROC Curve Analysis for TMAs

The results of the two transfer learning strategies were obtained as ROC curves showing the true and the false positive rate for the tumour class. The ROC curves were evaluated for TMAs based on the majority voting of the selected patches. To achieve ROC curves for TMAs, the model output in the form of probabilities of each patch for the tumour class was thresholded using 100 different values in the range [0, 1]. This led to predictions for patches with different threshold values. Subsequently, the predictions for patches obtained for each threshold value were majority voted to obtain a prediction for a TMA. The predictions for TMAs were used to calculate the true positive rate, the false positive rate and the ROC curve, as shown in figure 3 and 4. The predictions

for the TMAs obtained with 0.5 threshold were used to obtain the confusion matrix, mean sensitivities and mean F1 scores as reported in table 1, 2 and 3.

4 RESULTS

The main aim of this work was to confirm that the data fusion approach can achieve better breast cancer diagnosis than histological imaging data based on performance metrics. This was confirmed by one of the two transfer learning strategies. The results are divided in three parts as shown in table 1, 2 and 3. Table 1 and 2 report performance metrics obtained for transfer learning strategy 1 and transfer learning strategy 2, with and without data fusion approach, respectively. Table 3 shows a comparison of the two transfer learning strategies using only the data fusion approach. In table 1, 2 and 3 report values for the VGG16, the Inceptionv3 and the ResNet50 models. These values were evaluated for 96 TMAs acquired by majority voting of the five patches extracted from each TMA.

The results in table 1 show that the pre-trained features acquired from the data fusion approach yield slightly higher mean sensitivities and mean F1 scores in comparison to the pre-trained features extracted from the histological imaging data. Higher mean sensitivities using the data fusion approach were seen for at least two of the three DCNNs. Higher mean F1 score using the data fusion approach was seen only for the ResNet50 model. Specifically, the pre-trained features obtained from the data fusion approach using the Inceptionv3 and the ResNet50 models showed mean sensitivities 89.32% and 86.97%, respectively. Similarly, the mean F1 scores for the two models were 85.47% and 87.80%, respectively. In comparison, the pre-trained features from the histological data using the same DCNN model showed mean sensitivities 88.46% and 85.68%, respectively. Thus, there was approximately 2% increase in the model performance by data fusion approach based on the mean sensitivity, which is significant from a clinical perspective. However, the VGG16 model showed higher mean sensitivity (80.56%) using histological imaging data compared to the mean sensitivity calculated for the data fusion approach (79.06%). Overall, it can be seen that transfer learning using pre-trained DCNN features and a linear classification model (PCA-LDA) based on data fusion approach show a slight improvement in breast cancer detection in some cases for a small dataset as in our study.

Contrarily, table 2 obtained by the transfer learning strategy 2 shows lower mean sensitivities for

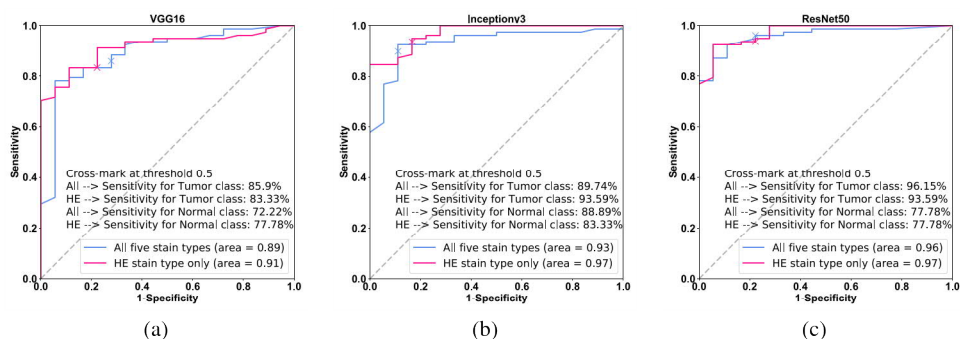


Figure 3: (a-c) show ROC curves for the VGG16, the Inceptionv3 and the ResNet50 networks using the transfer learning strategy 1 based on TMAs. The blue line shows ROC curve for the PCA-LDA model trained using the pre-trained DCNN features obtained from the histological and IHC imaging data, whereas the pink line shows ROC curve for the PCA-LDA model trained using pre-trained DCNN features extracted from the histological imaging data only. The cross-mark shows the true and the false positive rate on the ROC curve with 0.5 threshold.

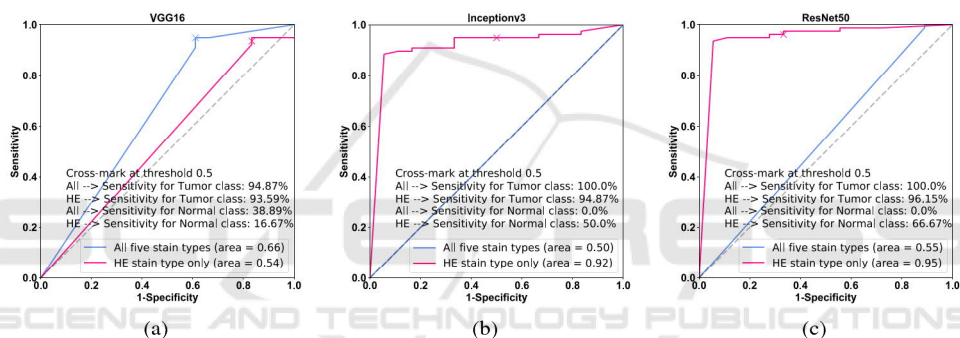


Figure 4: (a-c) show ROC curves for the VGG16, the Inceptionv3 and the ResNet50 networks using the transfer learning strategy 2 based on TMAs. The blue line shows ROC curve for the multi-input DCNN model fine-tuned using the histological and IHC imaging data, whereas the pink line shows the ROC curve for the single-input DCNN model fine-tuned using only the histological imaging data. The cross-mark shows the true and the false positive rate at 0.5 threshold.

the data fusion approach in comparison to the performance achieved by using histological imaging data alone. Except for the multi-input VGG16 network, the multi-input Inceptionv3 and the multi-input ResNet50 network trained with a combination of histological and IHC imaging data predicted all normal patches as tumour patches. Thus, the multi-input Inceptionv3 and the multi-input ResNet50 model achieved mean sensitivity of 50% and mean F1 score of 44.83%; whereas, the multi-input VGG16 network showed mean sensitivity of 66.88% and mean F1 score of 70.86% for the data fusion approach. The mean sensitivity of the single-input VGG16 network declined to 55.13% when only histological imaging data was used. On the other hand, the single-input Inceptionv3 and the single-input ResNet50 models using histological imaging data showed an opposite trend with comparatively higher mean sensitivities of

72.44% and 81.41%, and higher mean F1 scores of 75.66% and 83.78%, respectively. Overall, it was observed that transfer learning performed by fine-tuning the last layer of the pre-trained multi-input DCNNs result in lower mean sensitivities for the data fusion approach. This behaviour can be a consequence of the small sample size. It is clear from the results that fine-tuning the last layer of DCNNs is not the best approach for our small breast cancer dataset. Thus, it is suspected that the fine-tuning of all the layers of a DCNN will decrease the model performance further. However, fine-tuning of all layers for large breast cancer dataset should be investigated in the future.

Lastly, the performance of the two transfer learning strategies for the data fusion approach is summarized in table 3, where higher mean sensitivities are reported for strategy 1, i.e. using pre-trained features from the VGG16, the Inceptionv3 and the ResNet50

model. The training of the PCA-LDA model based on pre-trained features of the Inceptionv3 and the ResNet50 network yield promising results. The results from the VGG16 network are lower in comparison to the other two models for transfer learning strategy 1, but higher for transfer learning strategy 2.

The performance of the two transfer learning strategies based on TMAs is summarized in the form of ROC curves in figure 3 and 4. The ROC curve calculated for the data fusion approach and histological imaging data at various thresholds is depicted in blue and pink, respectively. The AUC values given in the figure legend show lower values for the data fusion approach in comparison to the AUC values calculated using histological imaging data. This trend is observed for both the transfer learning strategies. From figure 3 and 4, it can be inferred that the overall performance of DCNN models trained using an H&E image is better for both transfer learning strategies. However, the final performance of the models in terms of mean sensitivities evaluated at 0.50 threshold is better for the data fusion approach in some cases. The mean sensitivities cross-marked in each subplot of figure 3 and 4 are calculated at 0.50 threshold coincide with the values reported in table 1, 2 and 3. These values are evaluated for TMA's by performing majority voting of the five patches in each TMA. The ROC curves at threshold 0.50 which is mostly used to evaluate the model performance, show higher mean sensitivities for data fusion approach than using histological data, at least for the Inceptionv3 and the ResNet50 model in transfer learning strategy 1 (figure 3). Nevertheless, the AUC derived from the ROC curves for transfer learning strategy 2 (figure 4) show low mean sensitivities for all the DCNN networks. The inconsistency in the results of two transfer learning strategies can be due to various reasons discussed below.

5 DISCUSSION

Based on the results, three critical findings can be discussed.

5.1 Data Fusion vs. Histological Imaging

The results showed that the data fusion approach, i.e. combining histological and IHC imaging data, increases the model performance by $\sim 2\%$. However, the increase in model performance was achieved only for transfer learning strategy 1, where features were extracted from a pre-trained DCNN followed by binary classification using the PCA-LDA model. It is

important to mention that the analysis was performed on a limited number of TMAs and it is suspected that the results can improve with an increasing number of TMAs, at least for the transfer learning strategy 1. Furthermore, the data fusion approach can largely increase the feature dimension of the data, thus increasing computational complexity. Nevertheless, these limitations are the cost of performing reliable and early breast cancer diagnosis. In future studies, feature dimension can be reduced by extracting features from the last layers and a comparative study can be performed.

5.2 Strategy 1 vs. Strategy 2

From the results shown in table 3 it is clear that transfer learning strategy 1 outperforms the transfer learning strategy 2 for our breast cancer dataset. For transfer learning strategy 2, the misclassification of the under-represented normal class as tumour class is higher. This means that transfer learning strategy 2 performed by merging and fine-tuning the last layer of the pre-trained multi-input model causes 'negative transfer learning' showing lower binary classification performance. Although the past studies (Kensert et al., 2018; Mormont et al., 2018) have shown that transfer learning strategy 2 for medical imaging data can provide good classification performances, these studies used a single-input DCNN for fine-tuning; whereas, in this study a multi-input DCNN was used. Thus, training a large multi-input network on a small dataset can cause the model to overfit and degrade its performance. Degradation in model performance can also be a consequence of transferring features of top layers from two different domains (Yosinski et al., 2014). Specifically, the transferability of features can be negatively affected when the source task (e.g. classification of the ImageNet dataset) is different from the target task (e.g. breast cancer detection). Thus, transfer learning of features for different domains should be performed cautiously (Yosinski et al., 2014). Further, merging and fine-tuning only the last layer and initializing the weights of the whole network based on the ImageNet dataset transferred the specific features (learned in top layers) of the non-medical domain to the medical domain, thus decreasing the classification performance in the strategy 2. To improve the performance of a DCNN model by the transfer learning strategy 2, initializing and fine-tuning weights of the top and intermediate layers of the multi-input DCNN model should be investigated in future studies.

So far, limitations of the transfer learning strategy 2 were discussed, now it is important to discuss few

limitations of the transfer learning strategy 1. One of the limitations is the need for an aggressive down-sampling of the pathological images according to the input size of the pre-trained DCNN, ignoring the essential information. Although it is also possible to use a desired input image size by removing the fully connected layers of a pre-trained DCNN, downsampling our patches of size 1024×1024 to the model's input size facilitated the best classification performance. Extracting smaller size patches to increase the number of patches were also evaluated during the analysis. However, it was observed that small size patches increased the dataset size but decreased the biologically significant tissue features in each patch. Irrespective of our acceptable results using the pre-trained DCNNs as feature extractors, the interpretability of the transferred features is questionable. It is difficult to obtain an intuitive understanding of the transferability of non-medical features obtained from the ImageNet dataset to the medical domain. Thus, it is important to investigate transferring features from the medical domain to improve the breast cancer classification rate in future.

5.3 Effect of DCNN Architecture

It was clear from the results that acquiring a good classification rate using data fusion approach is dependent on the DCNN model. For transfer learning strategy 1, the Inceptionv3 and the ResNet50 network achieved better classification performances. While for transfer learning strategy 2, the multi-input VGG16 network achieved good classification performance. Furthermore, for transfer learning strategy 1, the Inceptionv3 and the VGG16 provided a large number of features (as they were combined from multiple modalities) in comparison to the ResNet50 network. Large feature dimension not only increased the dataset size but also increased the memory requirement. However, large feature dimension obtained by large DCNNs like the Inceptionv3 and the ResNet50 proved to be beneficial for training the PCA-LDA model in transfer learning strategy 1. While for transfer learning strategy 2, it was seen that large DCNN like the multi-input Inceptionv3 and the multi-input ResNet50 networks easily overfit and degrade model performance. It is suspected that large networks in multi-input fashion like the Inceptionv3 and the ResNet50 network generates a large number of trainable parameters which degrades model performance during fine-tuning. Furthermore, the time required to fine-tune the last layers of networks increases with network size.

6 CONCLUSION

The results show that combining histological imaging data along with IHC imaging data (estrogen receptor, progesterone receptor, human epidermal growth factor-2 and Ki-67) can improve breast cancer classification rate as compared to histological imaging data alone. The improvement in the classification performance was approximately 2% when deep convolutional neural networks (DCNN) were used as feature extractors (i.e. transfer learning strategy 1). However, the classification performance degraded when fine-tuning of the last layer of the multi-input DCNN (i.e. transfer learning strategy 2) was performed. Out of all three pre-trained networks, the pre-trained residual network and inception network as feature extractor outperformed the binary classification task (tumour vs normal), while the pre-trained VGG network as feature extractor obtained reasonable results. On the other hand, the VGG network showed better performances than the residual network and the inception network when fine-tuning of last layers was performed. The increase in performance by 2% for diagnosing breast cancer is explainable, because this task is normally performed using H&E, so the advancement is limited. Nevertheless, the data fusion approach can substantially improve differential diagnosis, which is important from a clinical perspective. Therefore, combining histology and IHC staining technique should be encouraged in future for more complicated tasks like a differential diagnosis or the prognosis of breast cancer patients. Overall, this comparative study showed that transfer learning could be utilized to diagnose breast cancer based on the combined histological and IHC imaging data with acceptable results. However, it is important to perform this study on a larger dataset in future. On large dataset, transfer learning strategy 3 i.e. training a DCNN from scratch can also be investigated. Furthermore, the data fusion approach can be performed to characterize stages of breast cancer in future.

ACKNOWLEDGEMENTS

Financial support of the German Science Foundation (BO 4700/1-1, PO 563/30-1 and STA 295/11-19) and funding by the BMBF for the project Uro-MDD (FKZ 03ZZ0444J) are highly acknowledged.

REFERENCES

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
- Antony, J., McGuinness, K., O'Connor, N. E., and Moran, K. (2016). Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1195–1200. IEEE.
- Bayramoglu, N. and Heikkilä, J. (2016). Transfer learning for cell nuclei classification in histopathology images. In *European Conference on Computer Vision*, pages 532–539. Springer.
- Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.
- Cheang, M. C., Chia, S. K., Voduc, D., Gao, D., Leung, S., Snider, J., Watson, M., Davies, S., Bernard, P. S., Parker, J. S., et al. (2009). Ki67 index, her2 status, and prognosis of patients with luminal b breast cancer. *JNCI: Journal of the National Cancer Institute*, 101(10):736–750.
- Chen, T. and Chefd'Hotel, C. (2014). Deep learning based automatic immune cell detection for immunohistochemistry images. In *International workshop on machine learning in medical imaging*, pages 17–24. Springer.
- Chollet, F. et al. (2015). Keras. <https://github.com/fchollet/keras>.
- Cireşan, D. C., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 411–418. Springer.
- Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567.
- Damodaran, S. and Olson, E. M. (2012). Targeting the human epidermal growth factor receptor 2 pathway in breast cancer. *Hospital Practice*, 40(4):7–15.
- Dobson, L., Conway, C., Hanley, A., Johnson, A., Costello, S., O'Grady, A., Connolly, Y., Magee, H., O'Shea, D., Jeffers, M., et al. (2010). Image analysis as an adjunct to manual her-2 immunohistochemical review: a diagnostic tool to standardize interpretation. *Histopathology*, 57(1):27–38.
- Elledge, R. M., Green, S., Pugh, R., Allred, D. C., Clark, G. M., Hill, J., Ravdin, P., Martino, S., and Osborne, C. K. (2000). Estrogen receptor (er) and progesterone receptor (pgr), by ligand-binding assay compared with er, pgr and ps2, by immuno-histochemistry in predicting response to tamoxifen in metastatic breast cancer: A southwest oncology group study. *International journal of cancer*, 89(2):111–117.
- Guo, S., Bocklitz, T., Neugebauer, U., and Popp, J. (2017). Common mistakes in cross-validating classification models. *Analytical Methods*, 9(30):4410–4417.
- Hanin, B. (2018). Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems*, pages 582–591.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, Germany.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.
- Huynh, B. Q., Li, H., and Giger, M. L. (2016). Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501.
- Janowczyk, A. and Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7.
- Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python.
- Kensert, A., Harrison, P. J., and Spjuth, O. (2018). Transfer learning with deep convolutional neural networks for classifying cellular morphological changes. *SLAS DISCOVERY: Advancing Life Sciences R&D*, page 2472555218818756.
- Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O., and Hajirasouliha, I. (2018). Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. *EBioMedicine*, 27:317–328.
- Kieffer, B., Babaie, M., Kalra, S., and Tizhoosh, H. R. (2017). Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, Citeseer.
- 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.
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.
- Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26.
- Mormont, R., Geurts, P., and Marée, R. (2018). Comparison of deep transfer learning strategies for digital pathology. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2262–2271.
- Oliphant, T. (2006). NumPy: A guide to NumPy. USA: Trelgol Publishing.
- Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000). Molecular portraits of human breast tumours. *nature*, 406(6797):747–752.
- Pham, N.-A., Morrison, A., Schwock, J., Aviel-Ronen, S., Iakovlev, V., Tsao, M.-S., Ho, J., and Hedley, D. W. (2007). Quantitative image analysis of immunohistochemical stains using a cmyk color model. *Diagnostic pathology*, 2(1):1–10.
- Ravishankar, H., Sudhakar, P., Venkataramani, R., Thiruvankadam, S., Annangi, P., Babu, N., and Vaidya, V. (2016). Understanding the mechanisms of deep transfer learning for medical images. In *Deep Learning and Data Labeling for Medical Applications*, pages 188–196. Springer.
- Ribeiro, E., Uhl, A., Wimmer, G., and Häfner, M. (2016). Exploring deep learning and transfer learning for colonic polyp classification. *Computational and mathematical methods in medicine*, 2016.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.
- Sheikhzadeh, F., Ward, R. K., van Niekerk, D., and Guillaud, M. (2018). Automatic labeling of molecular biomarkers of immunohistochemistry images using fully convolutional networks. *PLoS one*, 13(1):e0190783.
- Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.
- Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312.
- Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global.
- Van Ginneken, B., Setio, A. A., Jacobs, C., and Ciompi, F. (2015). Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 286–289. IEEE.
- Venables, W. N. and Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.
- Veta, M., Pluim, J. P., Van Diest, P. J., and Viergever, M. A. (2014). Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering*, 61(5):1400–1411.
- Wang, D., Khosla, A., Gargeya, R., Irshad, H., and Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.
- Webster, L., Bilous, A., Willis, L., Byth, K., Burgemeister, F., Salisbury, E., Clarke, C., and Balleine, R. (2005). Histopathologic indicators of breast cancer biology: insights from population mammographic screening. *British journal of cancer*, 92(8):1366–1371.
- Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

PEER-REVIEWED PUBLICATIONS

1. P. Pradhan, T. Meyer, M. Vieth, A. Stallmach, M. Waldner, M. Schmitt, J. Popp, T. Bocklitz, Semantic Segmentation of Non-linear Multimodal Images for Disease Grading of Inflammatory Bowel Disease: A SegNet-based Application, *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, 2019, Vol. 1: 396–405, SciTePress, Prague.
2. M. Yarbakht, P. Pradhan, N. Köse-Vogel, H. Bae, S. Stengel, T. Meyer, M. Schmitt, A. Stallmach, J. Popp, T. W. Bocklitz and T. Bruns, Nonlinear Multimodal Imaging Characteristics of Early Septic Liver Injury in a Mouse Model of Peritonitis, *Analytical chemistry*, 2019, Vol. 91(17): 11116–11121, ACS Publications, Washington.
3. P. Pradhan, S. Guo, O. Ryabchykov, J. Popp, T. W. Bocklitz, Deep learning a boon for Biophotonics?, 2020, *Journal of Biophotonics*, Vol. 13(6): e201960186, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.
4. T. Kirchberger-Tolstik, P. Pradhan, M. Vieth, P. Grunert, J. Popp, T. W. Bocklitz, A. Stallmach, Towards an interpretable classifier for characterization of endoscopic Mayo scores in ulcerative colitis using Raman Spectroscopy, 2020, *Analytical Chemistry*, ACS Publications, Washington
5. P. Pradhan, K. Köhler, S. Guo, O. Rosin, J. Popp, A. Niendorf and T. Bocklitz, Data fusion of histological and immunohistochemical image data for breast cancer diagnostics using Transfer learning. *Accepted as conference proceeding for 10th International Conference on Pattern Recognition Applications and Methods*, 2021.
6. P. Pradhan, T. Meyer, M. Vieth, A. Stallmach, M. Waldner, M. Schmitt, J. Popp, T. Bocklitz, Computational tissue staining of non-linear multimodal imaging using Generative Adversarial Networks, *Submitted to Journal Optica*.

CONFERENCES AND SUMMER SCHOOL

- Talks

1. P. Pradhan, T. Kirschberger-Tolstik, J. Popp and T. Bocklitz, Characterization of Inflammatory Bowel Disease based on Raman Spectroscopy, DokDok 2018, Friedrichroda, Germany, 17th-21th September, 2018.
2. P. Pradhan, M. Yarbakht, T. Meyer, M. Schmitt, J. Popp, T. Bruns and T. Bocklitz, Non-linear multimodal imaging of early septic liver injury, 4th International Symposium "Image-based Systems Biology" (IbSB), Jena, Germany, 6th-7th September, 2018.
3. P. Pradhan, T. Meyer, M. Vieth, A. Stallmach, M. Waldner, M. Schmitt, J. Popp and T. Bocklitz, Semantic segmentation of Non-Linear Multimodal images for disease grading of Inflammatory Bowel Disease – A SegNet-based application 8th International Conference on Pattern Recognition 2019, Prague, Czech Republic, 19th-21th February, 2019.

- Posters

1. P. Pradhan, J. Popp and T. Bocklitz, Automated classification of mucosal and crypt regions using non-linear multimodal imaging, DokDok 2017, Suhl, Germany, 18-22, September, 2017.
2. P. Pradhan, T. Kirschberger-Tolstik, A. Stallmach, J. Popp and T. Bocklitz, Pattern recognition analysis for prognosis of inflammatory bowel disease, Medical Imaging summer school (MISS 2018), 29th July- 4th August, 2018.
3. P. Pradhan, M. Yarbakht, N. Koese, H. Bae, S. Stengel, T. Meyer, M. Schmitt, J. Popp, T. Bruns and T. Bocklitz, Non-linear Multimodal Imaging for potential use in early septic liver injury investigation, 14th European Molecular Imaging Meeting (EMIM 2019), Glasgow, Scotland, 19-22, March, 2019.

WORKSHOP

1. Ex-vivo and in-vivo optical molecular technology: Potential and Trends, Institute for Physical chemistry, University Jena, 2017.
2. Academic writing skills, Graduate Academy, University Jena, 2018.
3. Scientific presentations, Graduate Academy, University Jena, 2018.
4. Scientific image processing and analysis, Graduate Academy, University Jena, 2018.
5. R as flexible tool for statistical analysis, Graduate Academy, University Jena, 2018.
6. Scientific writing and publishing for natural scientist, Graduate Academy, University Jena, 2018.

TEACHING AND SUPERVISION ACTIVITY

1. JSMC course on “Image analysis and Machine learning with R and Python”, Leibniz Institute of Photonic Technology and Institute of Physical Chemistry, Jena, 22nd -23rd November, 2018.
2. Practical course for M.Sc. Medical Photonics on “Quantum chemical investigations on the inhibition process of cysteine proteases”, Institute of Physical Chemistry, Jena, SS 2019.
3. Seminar project on “Stone, Paper, scissors game using Raspberry Pi and deep learning”, Carl-Zeiss-Gymnasium, Jena, WS 2019-2020.

Acknowledgments

I want to thank Prof. Dr. Juergen Popp for allowing me to conduct my PhD studies in his multidisciplinary working group. I very much admire his enthusiasm for topics like artificial intelligence in biophotonics, which has always helped me to improve my work and my thesis.

I want to express my sincerest gratitude to my supervisor PD Dr. Thomas Bocklitz, for his continued support and guidance throughout my PhD years. Starting from the beginning of my PhD studies until today, his endless patience and motivation, invaluable suggestions, constructive criticism, and acknowledgement of my ideas have been an integral part in the completion of my PhD. His support for my passion for working with “deep learning” for biophotonic data will be forever appreciated.

My special gratitude goes to Prof. Dr. Knut Moeller, whose lectures on “Statistical model building” during my master’s in Villingen-Schwenningen created an immense liking for this topic and led me to finally embrace it as my PhD work.

I want to thank Prof. Dr. Andreas Stallmach and Prof. Dr. med. Tony Bruns for sharing their vast biological knowledge and providing the data required to analyze my PhD work. My co-operations in Jena with Prof. Dr. Michael Schmitt, Dr. Tobias Meyer, Dr. Tatiana Kirchberger-Tolstik, and Dr. Melina Yarbakht have primarily contributed to my work and improved my manuscripts. Likewise, my co-operations with Prof. Dr. med. Michael Vieth and Prof. Dr. med. Maximilian Waldner (Erlangen), Prof. Dr. med. Tony Bruns (Aachen), Prof. Dr. Axel Niendorf, and Dr. Katharina Koehler (Hamburg) are highly acknowledged. Furthermore, a special thanks to my colleagues – Oleg, Shuxia, Nairveen, Rola, Darina, and Mehul for their advice and scientific conversations.

I want to thank my friends from Germany and India, who have always been very caring and understanding. Finally, I would thank my family and especially my father for his constant support during my studies. I want to thank my husband, Tanmay Raje, for being very supportive and patient throughout my PhD tenure. I also thank him for proofreading my thesis.

For financial support, I thank the “Deutsche Forschungsgemeinschaft” for granting me funding for the three years of my PhD work.

ERKLÄRUNGEN

SELBSTÄNDIGKEITSERKLÄRUNG

Ich erkläre, dass ich die vorliegende Arbeit selbständig und unter Verwendung der angegebenen Hilfsmittel, persönlichen Mitteilungen und Quellen angefertigt habe.

Pranita Rajan Pradhan

Datum

Ort

Unterschrift

