Toward Data Science in Biophotonics: Biomedical Investigations-based Study

Kumulative Dissertation

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

Vorgelegt dem Rat der Chemisch-Geowissenchaftlichen Fakultät der Friedrich-Schiller-Universität Jena

Von M.Sc. Nairveen Ali

geboren am 27.08.1987 in Damaskus, Syrien

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 Verteidigung: 16. Dezamber 2021

Table of Contents

Table of Contents--i

List of Figures---iv

List of Abbreviations--v

Chapter 1 Introduction---1

1.1 Overview of Biophotonics Technologies---1

1.2 Biophotonics-Based Applications--4

1.3 Data Science for Biophotonics--5

Chapter 2 State-Of-The-Art--9

2.1 Experimental Design--- 9

2.1.1 The Analysis of Multifactorial Experimental Designs--------------------------10

2.1.2 Sample Size Planning--11

2.2 Data Preprocessing--12

2.3 Machine Learning-Based Data Modeling--14

2.4 Model Validation--17

Chapter 3 Scientific Contribution--19

3.1 Experimental Design--22

3.1.1 Sample Size Planning for Multivariate Data-------------------------------------22

3.1.2 WE-ASCA for Analyzing the Unbalanced Multifactorial Designs------------24

3.2 Data-Driven Modeling and Validation---27

3.2.1 One Class Model-Based Antibiotics Susceptibility Prediction in Bacteria--27

3.2.2 Deep Learning-Based Bladder Tumor Classification--------------------------29

3.2.3 Model Validation Strategies-Based Automatic Detection of Breast Cancer-32

Additional Work---37

Summary---39

Zusammenfassung--45

Bibliography--53

List of Publications ---73

1 Sample-Size Planning for Multivariate Data: A Raman-Spectroscopy-Based Example--73

2 WE-ASCA: The Weighted-Effect ASCA for Analyzing 3 Unbalanced Multifactorial Designs – A Raman Spectra Based-Example---84

3 Predictive modeling of antibiotic susceptibility in *E. coli* strains based on the U-Net network and one-class classification--101

4 Deep learning-based classification of blue light cystoscopy imaging during transurethral resection of bladder tumors--119

5 Automatic label-free detection of breast cancer using nonlinear multimodal imaging and the convolutional neural network ResNet50---------------------------------------134

Curriculum Vitae--149

Conferences and Workshops--151

Acknowledgments---153

Erklärungen--155

List of Figures

Figure 1. A schematic diagram of data lifecycle in biophotonics----------------------------------6

Figure 2. Examples of biophotonic data before and after data preprocessing-------------------13

Figure 3. Data-driven pipeline for classical ML and DL---15

Figure 4. Overview of studies contributed to data lifecycle in biophotonics--------------------20

Figure 5. Visualization of sample size planning for multivariate data---------------------------23

Figure 6. The classification results of PCA-LDA and PLS-DA models based on leave-one-mouse-out cross-validation--26

Figure 7. The automatic pipeline for predicting antibiotic resistances---------------------------28

Figure 8. DL-based automatic classification of bladder cancer-----------------------------------31

Figure 9. Overview of the ML algorithms and the validation strategies utilized to evaluate the automatic detection of breast cancer---34

List of Abbreviations

OCT	Optical coherent tomography
IR	Infrared
CARS	Coherent anti-Stokes Raman scattering
TPEF	Two-photon excited fluorescence
SHG	Second-harmonic generation
IR	Infrared spectroscopy
OCT	Optical coherent tomography
ML	Machine learning
ANOVA	Analysis of variance
MANOVA	Multivariate analysis of variance
PCA	Principal component analysis
PC	Principal component
ASCA	ANOVA-simultaneous component analysis
GLM	General linear model
SSP	Sample size planning
DL	Deep learning
CNN	Convolutional neural network
CV	Cross-validation
IPL	Inverse power law
RMSE	Root mean square errors
LDA	Linear-discriminant analysis
WE	Weighted effect

PLS	Partial least square regression
AST	Antibiotic susceptibility testing
OCSVM	One-class support vector machine
ROC	Receiver operating characteristic OCSVM
PDD	Photodynamic diagnosis
BL	Blue light
CLAHE	Contrast limited adaptive histogram equalization
ROI	Region of interest
L10PO-CV	Leave-10-patients-out cross validation
H&E	Hematoxylin and Eosin
LOPO-CV	Leave-one-patient-out cross-validation

Chapter 1 Introduction

Biophotonics is an interdisciplinary field that aims to grasp and investigate the characteristics of biological samples based on their interaction with incident light (1,2). The light-matter interaction within biological samples is usually measured using optical tools, and the corresponding scientific field is termed biophotonics. Over the past few decades, numerous biophotonic technologies have been designed and innovated to extract various sorts of biological and chemical information from the studied samples. Such biological and chemical information is not directly acquired since all biophotonic techniques produce complex data in which the information is contained (3,4). Consequently, it is desirable to translate biophotonic-associated data to high-level information like disease biomarkers or sample characteristics. In this context, computer and data science advances using data learning approaches have inspired researchers to automatically analyze the acquired biophotonic data. In this chapter, an overview of biophotonic technologies in addition to their applications is introduced. Then, a brief outline of data science for biophotonic-associated data is presented.

1.1 Overview of Biophotonic Technologies

Biophotonics has been implemented in medicine and life science to understand and probe various characteristics of different biological systems (1,2). Since the last century, a broad spectrum of biophotonic technologies has been developed, allowing investigations of biological systems on several levels and using different properties (1,2,5). The first established biophotonic technology is the (bright-field) light microscopy. This microscopic technique exploits the light absorption in biological samples to characterize the contained structures (6,7). Later, light microscopy has been further developed into different techniques such as phase-contrast microscopy and differential interference contract microscopy (8,9). These developed microscopic techniques have been widely utilized for biological and medical investigations, although the obtained information is limited to the morphological features of the studied samples. Such morphological features usually describe only a single aspect in biological investigations, while these features are barely detected due to the low contrast of light absorption in most biological samples. To enhance this contrast, sample staining has been introduced, and it became the gold standard procedure in histopathological investigations and biological imaging. Despite the wide applications of sample staining in biomedicine, staining procedures are time consuming and cause sample perturbation resulting in application restrictions regarding living systems. Moreover, several diseases affect the biomedical composition of the biological systems, and therefore it was desired to utilize not only the morphology but also the chemical contrast of biological samples.

Another improvement of light microscopy is fluorescence microscopy. This enhanced microscopic technique can detect the chemical contrast of biomolecules based on the native weak single-photon excited auto-fluorescence of these molecules (10). The sensitivity of that molecular auto-fluorescence can be further improved if a fluorescence label is used (11,12). Nevertheless, staining process causes changes of the sample, while data acquisition is affected by fluorophore photobleaching. To overcome the previous constraints, other imaging technologies were developed (1,13). Prominent examples of such imaging systems include optical coherent tomography (OCT) (14), endoscopic methods and endomicroscopic techniques (15,16). These imaging technologies aim to improve the visualization of biological systems, and subsequently, improve the understanding of those systems. However, the aforementioned imaging technologies provide different sorts of information. For example, OCT utilizes low coherence light to capture two- to three-dimensional morphological images of tissues or body organs. This imaging technology offers label-free visualization of biological systems allowing for broad implementations in clinics. However, OCT doesn’t feature cellular resolution. In contrast to OCT, endoscopic and endomicroscopic techniques enable an meeasurment of organs within the body based on transmitting images and videos that can depict molecular or metabolic functions of the measured tissues in th organs. These techniques feature the drawback that the measurement is uncomfortable for patients and can cause complications in the measured organs.

The current developments of spectroscopic techniques have introduced non-invasive and label-free tools to biomedicine and biology, which can extract spatial and spectral information of biological samples (15–19). Prominent examples of spectroscopic techniques are infrared (IR) spectroscopy and Raman spectroscopy. These spectroscopies are able to capture a myriad of molcular information presented in the biological samples as unique spectral profiles of all biomolecules (18–22). IR spectroscopy can extract structural and chemical information based on light absorption within the infrared range of the electromagnetic spectrum. However, the molecular information and the spectral resolution provided by IR spectroscopy are constrained in aqueous environments due to the large background noise from water (21). Beside IR spectroscopy, Raman spectroscopy is a label-free and non-invasive biophotonic tool that has been widely applied to probe the molecular structures and composition of biological samples (17,23). This spectroscopic technology relies on the inelastic scattering of light measured in the studied samples serving a chemical fingerprint of biomolecules. The resulting Raman data can be utilized as a diagnostic marker, for instance, as a marker for abnormalities (17,22). In addition to the previous spectroscopic techniques, hyperspectral imaging systems that combine imaging and spectroscopic techniques also provide a non-invasive visualization of the spatial and spectral information in biological samples (24,25). Besides hyperspectral imaging systems, nonlinear multimodal imaging, combining coherent anti-Stokes Raman scattering (CARS), two-photon excited fluorescence (TPEF), and second-harmonic generation (SHG), was newly introduced as a fast-imaging technique that can detect the molecular contrast in the biological samples (26–30). This nonlinear multimodal imaging is usually characterized as a label-free and non-invasive imaging technique allowing non-destructive investigations of cells and tissues. Consequently, it might be an appropriate tool for in-vivo investigations as an optical biopsy when engaged in fiber-based measurements.

Each of the above-mentioned technologies has its advantages and limitations with respect to the level of sample alterations and the type of extracted sample information, although several technologies have already found their way to application fields, as shown in the next section.

1.2 Biophotonic Technology-Based Applications

The capability of biophotonic technologies to capture several biological and chemical information in biological systems enables multiple applications in biology and medicine (1,2,13). For instance, biophotonic technologies have shown a great potential to analyze the basic functionalities of biological systems in fundamental biomedical research (17,31,32). Such analysis is essential in monitoring the health condition and intends to understand disease genesis for early detection or even prevention of various diseases (1,2). Beside the implementation of biophotonics technologies in fundamental biomedical research, different tools were integrated into clinical procedures for early cancer identification and treatment, dentistry, cardiology, disease diagnosis, ophthalmology, and vascular medicine (31,33–37). In addition to the technology utilized in medical and biological investigations, biophotonics technologies have also been established well in the pharmaceutical industry and drug development (5,38–41). Common examples in this case cover flow-cytometry and fluorescence detection-based techniques. The aim here is to perform rapid investigations and assessments of biological matter reactions toward drugs (42–45). Nevertheless, several possible application fields of biophotonic tools were also demonstrated, including environmental monitoring, process control, food safety, and the point of care tests (1,5,13,17). The last application refers to evaluation procedures of healthcare, product, and clinician services provided for patients at the care time in clinics.

For any of the aforementioned disciplines, the utilized biophotonic technologies allow measuring different sorts of morphological and chemical information. This information is commonly contained in high-dimensional data like images or spectra. Furthermore, many of the biophotonic technologies are label-free, and subsequently the obtained data is untargeted making the interpretation of such data difficult (46). Therefore, data science is needed to use biophotonic data to the full extent.

1.3 Data Science for Biophotonics

Biophotonic technologies need to be coupled to data science methods to translate the biophotonic-associated data to information and knowledge, *e.g.*, disease biomarkers. This translation of biophotonic-associated data into interpretable information in the application context is challenging since biophotonic data show different levels of complexity (3,4,46). For instance, several biophotonic tools produce untargeted and high dimensional datasets that are difficult to be manually handled and subsequently difficult to extract any informative features from these datasets.

Recently, the revolution in data science has inspired advanced implementations of data learning approaches to analyze biophotonic-associated data. These approaches combine statistical learning techniques and machine learning algorithms in the so-called data lifecycle. Figure 1 depicts a systematic diagram of the data lifecycle when considering biophotonic-associated data. This lifecycle comprises experimental design, data acquisition, data cleaning and data preprocessing, data-driven modeling, and finally, model evaluation and deployment. In experimental design step, the aim of performing a certain study needs to be precisely determined. Therein, the experiment hypotheses and the required number of samples to test those formulated hypotheses are identified. Once an experiment is designed, data can be acquired from the planned sources according to the field of study, *i.e.*, survey-based data or experiment-based data. After data acquisition, the data preprocessing step is usually performed. This step revolves around techniques of data noise elimination, handling of missing data and data normalization. The obtained preprocessed data is utilized thereafter for data modeling and validation. While techniques for data-driven modeling combine statistical learning and mathematical algorithms to investigate data insights and then explore any potential pattern within the considered data, the goal of data validation is to evaluate the capability of a constructed model in predicting new datasets. In the last step of the data lifecycle, the evaluated models and the utilized data are stored to be deployed in future analyses.

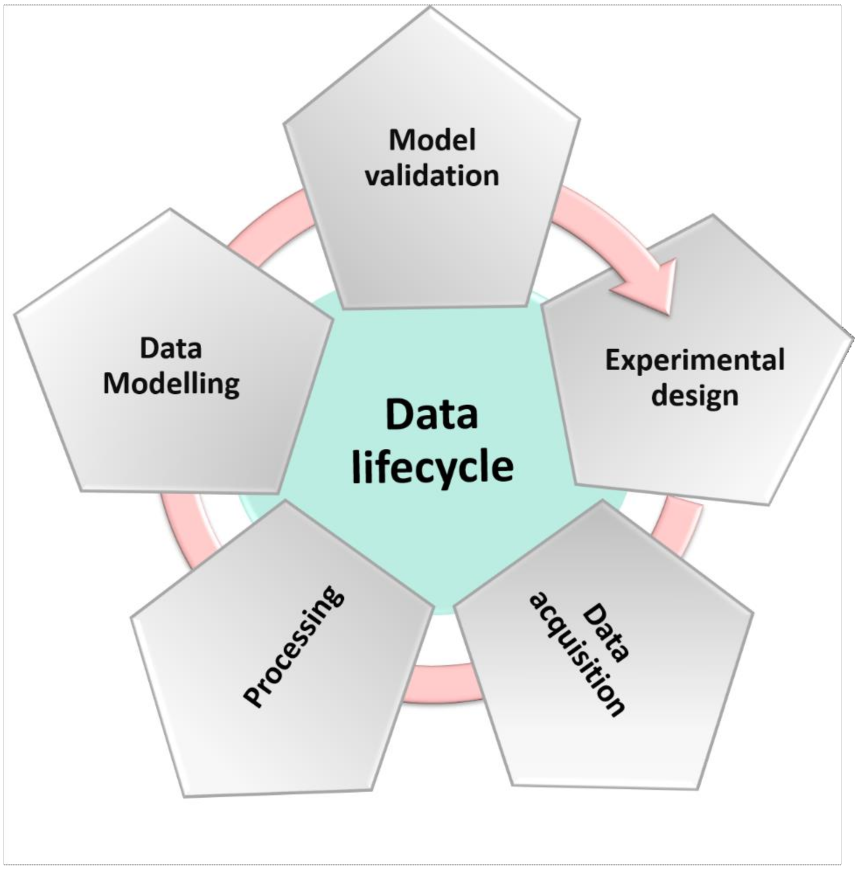


Figure 1. A schematic diagram of data lifecycle in biophotonics. This cycle describes a workflow that can be utilized to accomplish data-driven research. It starts by planning the study and deciding the number of samples needed to be collected. Thereafter, the acquired data are preprocessed and prepared to be used for constructing data models. These models can be validated and evaluated using several validation strategies in order to be utilized for further studies or applications.

The data lifecycle for biophotonic technologies, including statistical techniques and data learning approaches, is not fully researched and needs further developments. Chapter 2 introduces a systematic review of statistical techniques and data learning approaches for the analysis of biophotonics-associated data. The applications of many established statistical and machine learning techniques are still limited in biophotonics. Consequently, selected own scientific contributions to biophotonic data science are briefly discussed in Chapter 3. These contributions aimed to improve the planning and the design of biophotonic experiments then verify machine learning pipelines on several biophotonic imaging modalities.

Chapter 2 State-Of-The-Art

This chapter provides an overview of the statistical tools and machine learning (ML) techniques implemented in biophotonic data science. The presented approaches aim to improve the design of the experiments, suppress disturbing distortions of biophotonic data, and assess and validate ML techniques.

2.1 Experimental Design

The term "experimental design" refers to the protocols that formulate the statistical hypotheses needed to investigate the effect of specific treatments (variables) on a selected dataset (47,48). Three primary types of experimental design can be utilized for experimental research: pre-experimental design, true-experimental design, and quasi-experimental design (49–51). In pre-experimental design, the behavior of either one or multiple groups is observed to identify a potential effect of a studied treatment, which is characterized by "experimental factors". This exploratory, experimental approach aims to understand if a further investigation of the studied groups and treatments is required or not. Besides pre-experimental design, true-experimental design is often performed to checks how significant the experimental factors affect the considered dataset. Thereby, the response of samples selected from that dataset exposed to specific treatments is observed, and then this response is compared to other selected control samples, *i.e.*, the samples without any treatment. Regardless of the random sample assignment required in true-experimental designs, quasi-experimental designs can be performed similarly to true-experimental design (50). This quasi-experimental design is beneficial when the random assignment of control and treated groups is either irrelevant or not required.

To analyze any of the previous designs, the experimental factors in addition to the number of samples required to conduct such experimental studies need to be precisely defined. In the following, an overview of statistical techniques developed for the analysis of multifactorial experimental designs is presented. Thereafter, the established algorithms for determining the sample size required to achieve significant statistical results are briefly reviewed.

2.1.1 The Analysis of Multifactorial Experimental Designs

In order to investigate the effect of one factor or a number of factors on a conducted experiment, the analysis of factorial design can be utilized (52). This group of statistical techniques has been implemented in biomedical and biological research to explore hypothesized effects in a particular design. Herein, an experiment can be conducted using a specific dataset of different samples, then the effect of each experimental factor can be investigated according to its influence on the sampled data (53). The previous group of analyses has been established well when using one response variable to describe the selected samples, *i.e.*, univariate datasets, and it is known as analysis of variance (ANOVA) tests (54–56). In a classical ANOVA test, termed as a one-way ANOVA test, the influence of one factor on selected samples is evaluated based on studying the mean differences between factor levels. These factor levels usually indicate the possible values of the studied factor, *e.g.*, drug concentration. Besides the one-way ANOVA test, multi-way ANOVA tests search in a multifactorial design for any significant effect of the experimental factors and their interactions.

The above-mentioned tests were established for univariate data; however, only a few techniques for multifactorial designs were developed when multiple variables for the response data, *i.e.*, multivariate data, are utilized. Moreover, these analyses suffer from several limitations, which restrict their applications. For example, multivariate-ANOVA (MANOVA) tests analyze multivariate data in multifactorial experimental designs by performing an ANOVA test for each response variable (57,58). Although MANOVA tests allow determining the effect of one or more than one factor on these response variables, they are constrained to response datasets containing a much larger number of samples than the number of variables. Such datasets are rarely available for modern technologies which produce high dimensional measurements, like spectra or images. Therefore, combining principal component analysis (PCA) models with ANOVA tests offered an alternative to MANOVA tests for high dimensional multifactorial designs (59). In the PC-ANOVA test, the response matrix is fitted with a PCA model, and then the obtained principal components (PCs) are analyzed using ANOVA tests. Despite the wide applications of PC-ANOVA tests in small-sized datasets, several aspects related to factor contributions may be lost during the PCA projection. To overcome this drawback, ANOVA-simultaneous component analysis (ASCA) was introduced as a powerful tool to deal with multivariate data in multifactorial designs (60–62). Thereby, the response matrix can be decomposed into different effect matrices characterizing the contribution of each factor and each factor interaction in the designed model. These contributions are then measured based on the amount of variance explained by each possible effect, *i.e.*, each factor and each interaction. Finally, the ASCA test searches for significant effects in the designed model, and the dimensions of each effect matrix are reduced using a PCA model for better interpretation of the effect contributions (63).

The application of almost all previously described tests for multivariate data is constrained to balanced designs in which equal numbers of samples are needed for each factor level. As a result, the applications of these tests are limited. Alternatively, the ASCA+, an extension of ASCA, was introduced to analyze multivariate data in unbalanced designs (64). It utilizes a specific version of general linear models (GLMs) to decompose the response matrix into two main terms: The estimated response matrix and the estimation error (56,64). Even though the proposed ASCA+ provides a unique solution to estimate the contributions of experimental factors, it seems that this analysis underestimates these contributions in unbalanced designs (65). Therefore, an own scientific study was performed in [PII] as a new adjustment of the ASCA algorithm for unbalanced multifactorial designs.

2.1.2 Sample Size Planning

Sample size planning (SSP) represents strategies intended to determine a sufficient number of samples needed to perform robust and accurate statistical analysis (66,67). This SSP determination becomes more important in biomedical experiments due to the high costs and the ethical restrictions that may inhibit recruiting samples from patients and animals. Consequently, the aim of SSP in such experiments is to determine the minimal number of samples required to prove that group differences are significant. Initially, SSP techniques were established based on hypothesis testing (68). Thereby, the required sample size is estimated using a priori knowledge of the statistical distributions of these groups (69,70). However, these group distributions are mostly unavailable in the case of data acquired by modern biophotonic tools, and thus, modified SSP algorithms are needed.

So far, a few SSP algorithms for multivariate data were developed, while most of them were designed for classification tasks (71–73). In most of these algorithms, the learning curves were prominently implemented as an exploratory tool to describe the classification performance when increasing the training set sizes of that classification model. Thereafter, the sample size required to train this classifier can be predicted based on the inverse power law models (74). Nevertheless, the previously described learning curve-based SSP was established and checked for specific datasets like microarray datasets or resampling datasets. Moreover, the implementation of such algorithms was not clear for diagnostics tasks, *e.g.*, the estimate of the required number of patients. Therefore, a general SSP algorithm was presented in the scientific contribution [PI] to cover the mentioned challenges. Thereby, the sample sizes for a successful group differentiation can be determined for different levels within the data hierarchy, *i.e.*, spectra, biological replicates, patients, etc.

2.2 Data Preprocessing

Biophotonic tools usually deliver high-dimensional datasets containing various sorts of data variations. These variations are mainly categorized into informative variations and disturbing variations. While informative variations represent the differences between different states such as sample properties or disease states, disturbing variations may be assigned to systematic perturbations within experiments, *i.e.*, conducting an experiment using many devices or by several individuals (75,76). The later variation is very complicated and difficult to be controlled; hence, it might negatively influence the results of further statistical analyses.

Data preprocessing have shown a profound effect on reducing, or even eliminating, such disturbing variations. For biophotonic data, several preprocessing techniques have been established to eliminate these data variations according to the delivered data, *e.g.*, spectra or images. For instance, spectra collected from Raman spectroscopy are usually contaminated by cosmic spikes and fluorescence baseline in addition to several types of noise (see Figure 2-A). To deal with these corrupting effects in raw Raman spectra, a proper preprocessing pipeline was presented in (3,77–82). It starts by eliminating cosmic spikes within the acquired spectra and then calibrating the utilized spectrometer. Thereafter, the background effect is excluded

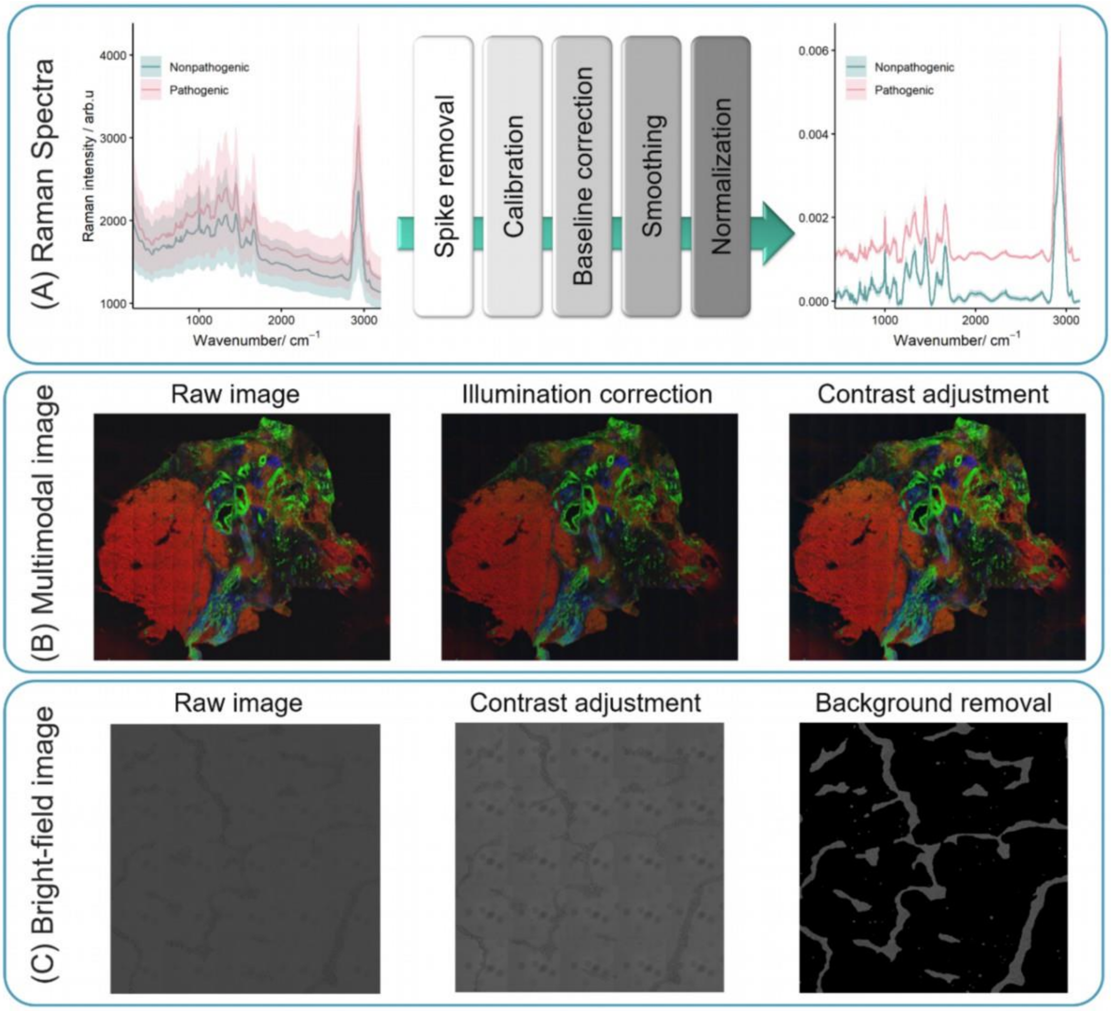


Figure 2. Examples of biophotonic data before and after data preprocessing. Most disturbing variations have been removed when proper preprocessing techniques are applied.

from the aligned spectra. Lastly, the corrected spectra are smoothed and normalized. This preprocessing workflow was successfully implemented in the own scientific contributions [PI] and [III] when considering Raman spectral datasets. Nevertheless, other preprocessing techniques have been developed to remove the corrupting effects produced when using biophotonic imaging tools. According to the acquired raw data, the image preprocessing techniques aim to improve the quality of the obtained image data. These preprocessing techniques include different algorithms for image background removal, image smoothing, image stitching, contrast adjustment, image registration, etc. (83–88).

In this thesis, an individual image processing pipeline was presented for each scientific contribution dealing with imaging data. Figure 2 presents three examples of spectral and imaging data before and after applying preprocessing techniques. It is obvious in this figure that the preprocessed data disposed of their disturbing variations if they are compared to the raw image and spectra. Consequently, the obtained preprocessed image and spectra seem to be more informative for further investigations.

2.3 Machine Learning-Based Data Modeling

Data-driven modeling has been commonly implemented to extract high-level information and informative features from the preprocessed data. Techniques for data-driven modeling combine statistical learning and mathematical algorithms to infer data insights and then explore any potential pattern presented within the collected datasets (89,90). Recently, ML algorithms have gained growing attention for data-driven modeling due to their remarkable capability in automating and assessing most phases of data-based learning (4,91). Concerning the goal of data-based learning, the utilized ML algorithms can be roughly categorized into unsupervised ML and supervised ML techniques (4,90,91). The later ML techniques are quite crucial, and they aim to map the acquired data via a predefined mathematical algorithm to predictor variables, which represent, for example, image labels or drug concentrations. Based on the supervised ML techniques, it is possible to build self-supervised models and systems that can automatically learn data patterns and retain the learned knowledge into model structures

(92,93). Moreover, several novel ML models have currently exposed a further capability to improve their prediction performance over time with minimal human intervention and without being explicitly designed for such tasks.

Until a few years ago, ML-based data modeling relied only on a manual feature design and extraction for exploring data patterns (94,95). This type of data modeling is known as classical ML algorithms. In Figure 3-A, a common pipeline of data-driven modeling based on classical ML algorithms is presented. It starts with manually designing and extracting data features that well represent the data (91,96). This new data representation is often described using too many features; hence, a feature extraction step is followed by a dimension reduction step. Therein, a small subset of uncorrelated features can be extracted, resulting in an adequate new representation of the original dataset (96–98). The subsequent features are finally mapped to the data predictors using a specific ML model regarding the task of interest, *e.g.*, regression or classification. The aforementioned group of ML techniques was successfully applied to evaluate several tasks in biophotonics. Prominent examples in this field include disease investigation, cancer detection, and bacteria identification (99–104).

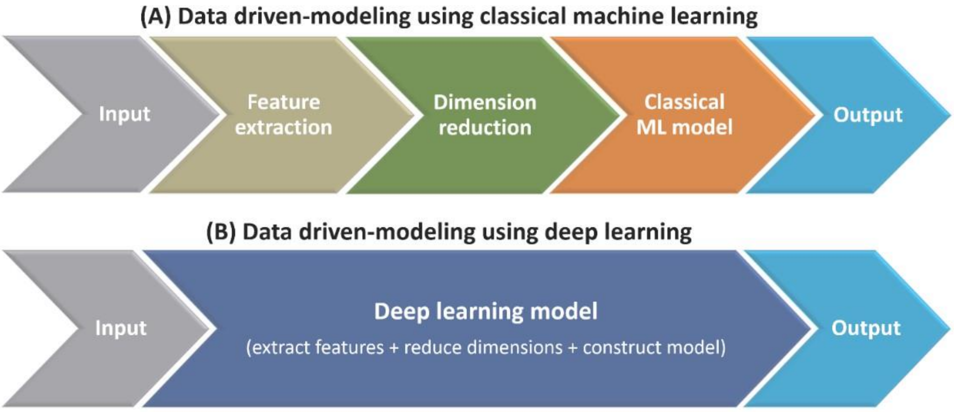

(A) Data driven-modeling using classical machine learning
Input
Feature extraction
Dimension reduction
Classical ML model
Output
(B) Data driven-modeling using deep learning
Input
Deep learning model
(extract features + reduce dimensions + construct model)
Output

Figure 3. Data-driven pipeline for classical ML and DL. While data-driven modeling based on classical ML models comprises feature extraction and dimension reduction in order to construct ML models, data-driven modeling based on DL models combine features extraction and model construction in one model.

In recent years, deep learning (DL), a subset of ML, has revolutionized the future of ML applications. With these DL techniques, it is possible to build ML models that can automatically learn data patterns using minimal human involvement (105,106). As depicted in Figure 3-B, data-driven modeling based on DL internally compresses all modeling stages of classical ML, starting from feature extraction and reaching decision-making. Nonetheless, one of the essential architectures of DL models was inspired by human brains, namely deep convolutional neural networks (CNNs) (106). These networks can be trained by passing a dataset of labeled images (or one-dimensional data) through multiple convolutional layers consisting of simple units called filters. These convolutional layers can detect a local combination of the data features from the previous layer, then they pass the resulted feature map into the next layer through a static nonlinearity, *e.g.*, replacing negative values with zeros. This layer is usually named the activation layer, and it is followed by some pooling layers that intend to reduce the number of image features. Later, CNNs process the input data as a sequence of visual representations in which each filter in a certain convolutional layer identifies a specific local region of the feature map obtained by the previous layer, while similar feature detectors exist across the locations in the feature map (106,107). The described training procedure of CNNs is commonly known as end-to-end CNN training (108). Based on this training procedure, the potential of DL models has been checked for several applications, including speech recognition, natural language processing, and healthcare. For the last application, DL models exhibited impressive performance, particularly in cell detection and cell counting (109–111), image segmentation (110,112–114), and tissue classification (115–118).

Unfortunately, the above-mentioned training procedure for DL models exhibits specific limitations in biophotonic-associated data due to the large sample size required to learn and optimize such models. Consequently, transfer learning of DL models was introduced as an alternative learning strategy to overcome the limitations of end-to-end CNN training. Thereby, the knowledge gained via training CNNs on a large-annotated dataset can be transferred to solve another task within a new small-sized dataset (46,119). In this context, DL-based transfer learning has shown a great performance in the diagnostic classifications of biomedical images using relatively small sample sizes (120,121). Therefore, transfer learning of several publicly available pre-trained CNNs was evaluated within three scientific contributions presented in Chapter 3, *i.e.*, [PIII], [PIV] and [PV]. The goal of two implementations was to automatically detect bladder and breast cancer using biophotonic imaging tools, namely: blue light cystoscopy and nonlinear multimodal imaging, respectively.

Apart from the training procedures of ML models, over-fitting has been addressed as another challenge in data-driven modeling. The term “Model over-fitting” describes ML models that are trained perfectly on specific training datasets, but they lack the prediction performance on new similar datasets. To avoid such over-fitting, the model performance needs to be evaluated on a new independent dataset named validation data. Thereafter, the optimized model is again checked on another dataset described as a test dataset. In the following section, common validation strategies that can be utilized to verify the performance of ML models are presented.

2.4 Model Validation

One of the main goals of data-driven modeling is to exploit the previously trained models in predicting new datasets. These models need to be carefully optimized and validated to rely on the new predictions of a trained model (122). In this context, the term “model validation” depicts data splitting strategies used to validate the performance of a model trained on a specific dataset for predicting new datasets. Model validation is usually based on two datasets: a dataset used for the model construction, *i.e.*, the training set, and a dataset not being used for the model construction, *i.e.*, the validation set. While the validation on the training set usually revolves around parameter optimization and tuning, model validation using the validation set checks the prediction performance of the trained model. Therefore, both training and validation datasets contribute significantly to test the quality and reproducibility of ML models (123,124).

Two classical validation strategies can be implemented to split data into training and validation sets: training-test validation and cross-validation (CV). In the training-test validation, the acquired labeled data is portioned randomly into two subsets: the training set and the test set. Subsequently, an ML model is trained on the first subset and checked on the later subset. In contrast to the training-test validation, the dataset is randomly split into k subsets when using CV. Subsequently, the considered ML model is trained on all data subsets except one subset, while the set-aside subset is utilized to test the prediction performance of the trained model. The previous procedure is iterated until all samples within all subsets are tested and predicted once.

In both validation strategies, the division of the considered dataset should not only be performed randomly but it is preferred to be also performed on the highest level of the data hierarchy, *i.e.*, patient level or replicate level (124). When using biophotonic tools, the later constrain becomes quite crucial for data modeling and validation. Thereby, multiple measurements can be acquired from the same patient (biological replicate), manifesting high internal correlation. Nevertheless, to avoid the previous correlation effect, the considered validation strategies for all ML models in all their own contributions were performed on the highest level of the data hierarchy, *i.e.*, patient level or replicate level. Therein, the model performance was evaluated using one of the following validation versions: training-test data validation on the patient level, leave-one-replicate-out CV, and leave-k-individuals-out CV; where k$\in \{1, 10\}$.

As a summary, this chapter presented an overview of several data science techniques in biophotonics to improve the experimental design and then automatically translate biophotonics-associated data to beneficial markers. These markers can be utilized to understand and investigate further many biological systems. In the next chapter, scientific studies of own research are presented to verify and adjust several statistical and ML algorithms on biophotonics-associated data.

Chapter 3 Scientific Contribution

As previously described, the implementation of statistical approaches combined with ML techniques can improve data investigations and help reducing human intervention. In the case of data produced using biophotonics technologies, such implementations require further study and adjustments. Therefore, the goal of the scientific contributions included in this chapter is to fill specific gaps related to the design and the evaluation of the biomedical experiments that use biophotonic technologies.

In Figure 4, the selected studies are allocated according to their contribution to the data lifecycle. On the side of experimental design, a general algorithm for estimating the required training set size for classification models was developed. Subsequently, the statistical analysis of experimental designs was improved for unbalanced multifactorial designs. Moving to data-driven modeling and validation, the performance of several ML techniques and validation strategies was evaluated for the automatic identification of three medical diagnostic studies. The scientific publications based on the above-mentioned studies are listed in the following with respect to their order in this chapter:

[PI] N. Ali, S. Girnus, P. Rösch, J. Popp, and T. Bocklitz
Sample-Size Planning for Multivariate Data: A Raman-Spectroscopy-Based Example
Analytical Chemistry, 2018, 90 (21), 12485-12492.

[PII] N. Ali, J. Jansen, A. Doel, G. H. Tinnevelt, and T. Bocklitz
WE-ASCA: The Weighted-Effect ASCA for Analyzing 3 Unbalanced Multifactorial Designs – A Raman Spectra Based-Example
Molecules, 2020, 26 (1), 66

[PIII] N. Ali, J. Kirchhoff, P. I. Onoja, A. Tannert, U. Neugebauer, J. Popp, T. Bocklitz
Predictive modeling of antibiotic susceptibility in *E. coli* strains based on the U-Net network and one-class classification
IEEE Access, 2020, 8, 167711-167720

[PIV] N. Ali, C. Bolenz, T. Todenhöfer, A. Stenzel, P Deetmar, M. Kriegmair, T. Knoll, S. Porubsky, A. Hartmann, J. Popp, M. C. Kriegmair, and T. Bocklitz
Deep learning-based classification of blue light cystoscopy imaging during transurethral resection of bladder tumors
Scientific reports, 2021, 11, 1169

[PV] N. Ali, E. Quansah, K. Köhler, T. Meyer, M. Schmitt, J. Popp, A. Niendorf, and T. Bocklitz
Automatic label-free detection of breast cancer using nonlinear multimodal imaging and the convolutional neural network ResNet50.
Translational Biophotonics, 2019, 1, e201900003

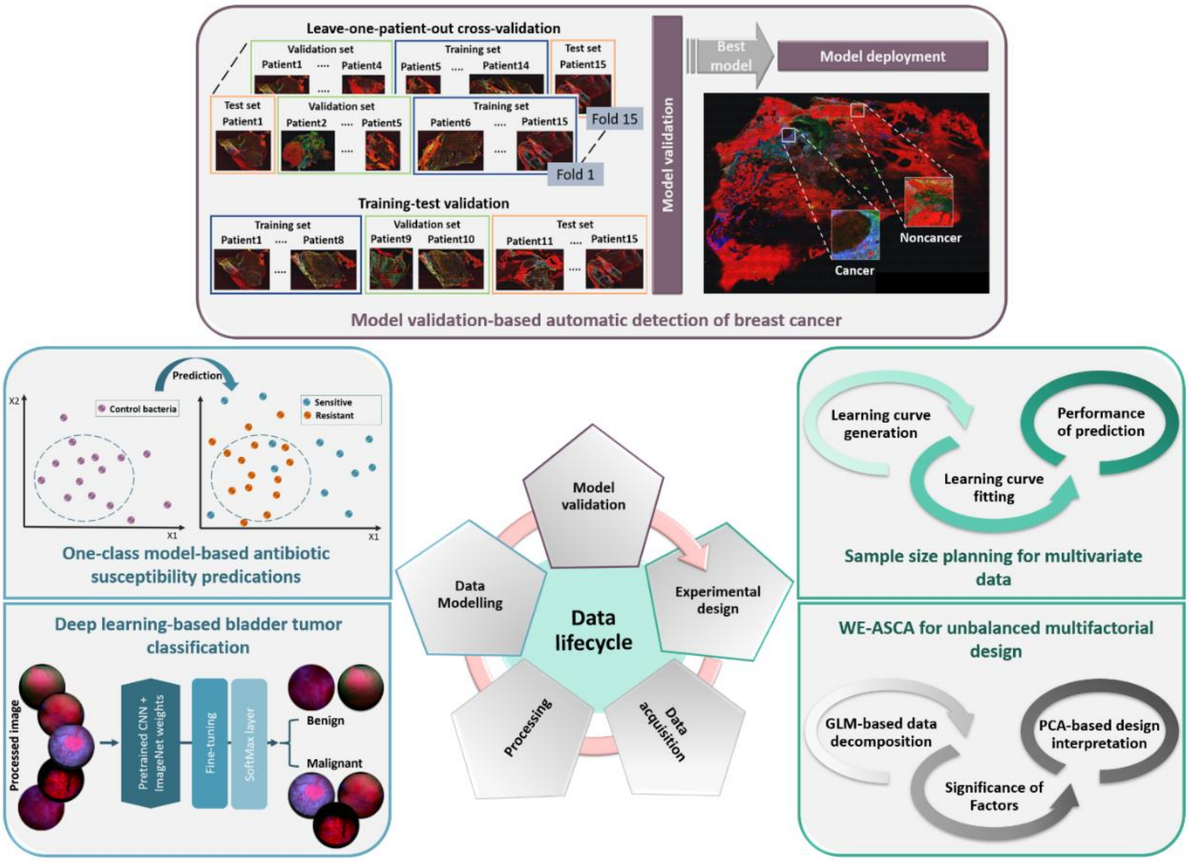


Figure 4. Overview of studies contributed to data lifecycle in biophotonics. The SSP and the WE-ASCA studies were performed to improve the experimental planning and the analysis of experimental designs, respectively. In contrast, the automated identification of antibiotic susceptibility in bacteria and the automatic classification of bladder cancer were demonstrated based on image-driven modeling. Finally, the performance of the presented validation strategies was evaluated for the automatic detection of breast cancer.

3.1 Experimental Design

The scientific contributions in this section were established to deal with specific issues related to the experimental design presented in section 2.1. These contributions focus on estimating the sample size required for group differentiation and on evaluating the influence of experimental factors on unbalanced multifactorial designs. Both methods were designed for multivariate data and were checked on biomedical Raman spectral datasets.

3.1.1 Sample Size Planning for Multivariate Data

The SSP aims to estimate the minimal number of measurements needed to achieve robust and significant statistical results. SSP has become more important for biophotonics-associated data because the generation of such data is time-consuming and features several limitations regarding the high costs and the ethical restrictions. Furthermore, the methods proposed for multivariate data-based SSP are still limited to specific applications (72,73,125). Therefore, the scientific contribution published in [PI] presented a developed SPP algorithm to estimate the training set size required for constructing a specific classification model in the case of multivariate data.

The suggested SSP algorithm was built based on learning curves and a specific version of inverse power law (IPL) models. Figure 5 shows a systematic pipeline of the proposed SSP algorithm. It starts by generating the learning curve of a specific classifier by quantifying the performance of this classifier when increasing the sizes of its training set. Thereafter, the generated learning curve is fitted using the nonlinear least-squares algorithm by the IPL (74,126). In [PI], the considered formula of the IPL model is:

$$\mathrm{IP}(n) = a \times n^{-b} + c,$$

where n denotes the training set size, whereas $\mathrm{IP}(n)$ estimates the quantified performance when training the classifier on n samples, and a, b, and c represent the parameters of the IPL model. Here, a refers to the learning rate, b is the decay rate, and c represents the final performance of the considered classifier if it is trained on an infinite number of samples. The later parameter is usually known as the Bayes error. It is utilized in the presented SSP algorithm to predict the training set size required to achieve 95% of Bayes error, named: $n_{95\%}$. After predicting the $n_{95\%}$, a new IPL model is fitted using only the training set sizes $\leq n_{95\%}$, while the performance of this fitted model is extrapolated for the training set sizes $> n_{95\%}$ named the extrapolated region. Finally, the performance of the fitted IPL model is evaluated based on the root mean square error (RMSE) of the IPL performance and the classification performance in the extrapolated region.

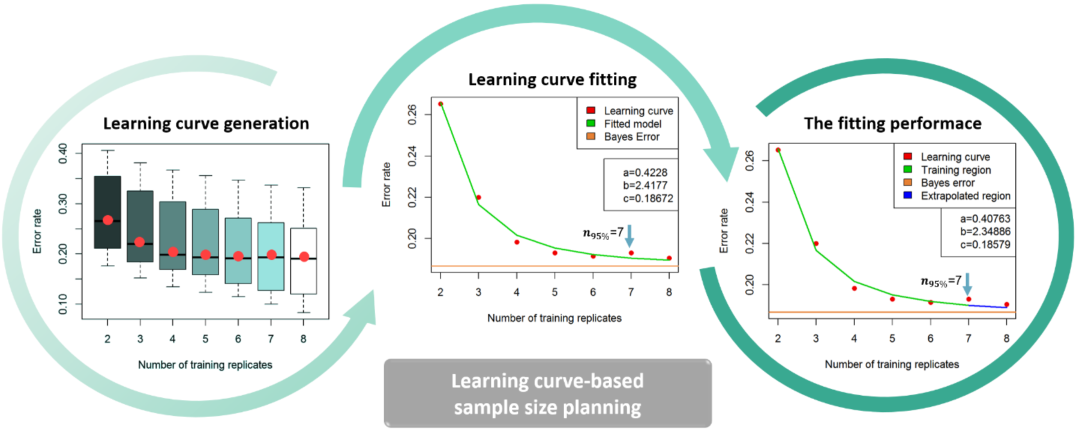


Figure 5. Visualization of sample size planning for multivariate data. The learning curve is generated using an increasing number of biological replicates. Thereafter, the obtained learning curve is fitted by the inverse power law, and the acquired fit is utilized to predict the training set size required to achieve 95% of Bayes error, *i.e.*, $n_{95\%}$. Finally, the predicted training set size is implemented to extrapolate the performance of inverse power law built upon training set sizes $\leq n_{95\%}$.

The established SSP algorithm was demonstrated on a Raman-spectral dataset consisting of six bacterial species cultivated in nine independent biological replicates. Thereby, the sample sizes needed to train a classification model that combines a PCA model with a linear-discriminant analysis (LDA) model were estimated for different data hierarchy levels, *i.e.*, spectral level and replicate level. The obtained results showed that 142 Raman spectra per bacterial species and seven biological replicates are required to achieve 95% of the final performance of the PCA-LDA model, *i.e.*, 95% of Bayes error.

3.1.2 WE-ASCA for Analyzing the Unbalanced Multifactorial Designs

The analysis of multifactorial designs is a group of statistical methods that are typically utilized to study the effect of multiple treatments on selected samples. Unlike univariate data, this group of exploratory tools is limited in unbalanced multifactorial designs when considering multivariate data (59,60). Consequently, the weighted-effect ASCA (WE-ASCA) presented in the scientific contribution [PII] was developed as an updated version of the classical ANOVA-simultaneous component analysis (ASCA) to deal with the unbalanced multifactorial designs of multivariate data.

The proposed WE-ASCA suggests characterizing the experimental design based on the general linear models (GLM) and the weighted-effect (WE) coding. Thereby, the response matrix can be decomposed into two terms: the estimated response and the error in this estimation. While this estimated response matrix usually consists of two matrices, *i.e.*, the design matrix and the parameter matrix, the error matrix is obtained by the difference between the response and its estimation. Nevertheless, the main improvement of the WE-ASCA is the implementation of the WE-coding, which forms a specific version of the dummy coding, to facilitate the inclusion of categorical variables in the GLM formula (127,128). This WE-coding offers a unique solution to solve the GLM equations in which the effect of each factor level represents the level deviation from the weighted mean. Therefore, the WE-coding was promoted to update the coding scheme of the design matrix in GLM equations when considering an unbalanced multifactorial experimental design. After that, the parameter matrix in GLM equations can be estimated easily based on the ordinary least square method (129,130). In the last step of WE-ASCA, the obtained estimated response based on the design matrix and the parameter matrix can be decomposed linearly as different effect matrices representing the experimental factors and their interactions. Besides, the significant effects in a particular design are determined using permutation tests, while the dimensions of the effect matrices are reduced using PCA models (63).

Using a Raman spectral dataset consisting of four colorectal tissue types collected from 47 mice in 387 scans, two applications of WE-ASCA were evaluated. This dataset was acquired with respect to four factors describing the experimental design: The individuals with 47 levels referring to the mice, the activity of the P53 gene, the mouse gender, and the location of samples (colon or rectum) (131). The first application intended to understand and analyze the design of that experiment and then determine which of the experimental factors contributed significantly to the considered design. The previous analysis was achieved by applying ASCA, ASCA+ and WE-ASCA and comparing their results based on the percentage of explained variances by all effects. It tuned out that the classical ASCA overestimated the effect contributions, while the ASCA+ underestimated these contributions. In contrast to ASCA and ASCA+, the presented WE-ASCA performed the best in estimating these effect contributions. Nevertheless, the three versions of ASCA showed that the individual factor has the largest significant effect in the considered design. Therefore, the influence of excluding such variations on the classification of colorectal tissues was checked in the second task using two classifiers, namely: the PCA-LDA and the combination of partial least square regression with LDA (PLS-DA). In this context, four different classification tasks were evaluated based on the leave-one-mouse-out cross-validation. Figure 6 visualizes the obtained results of the considered tasks. It is observed that excluding the contribution of the individual factor from the training set introduced more robust classification results, and it improved the mean sensitivity in most classification tasks. Moreover, training an LDA model on spectra, in which their individual effect was excluded, required a smaller number of principal components (or latent variables) and improved the reproducibility of CV results.

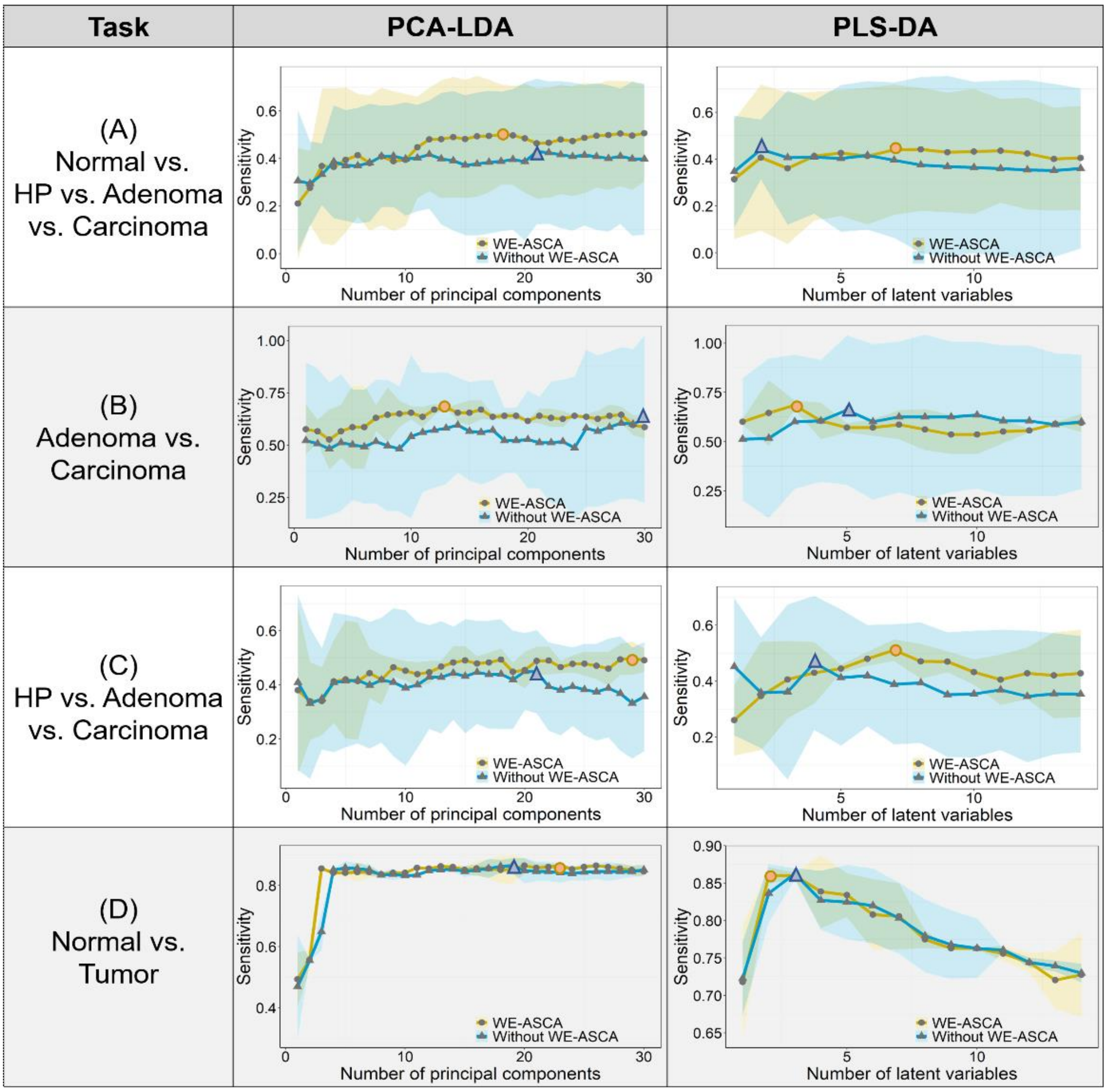


Figure 6. The classification results of PCA-LDA and PLS-DA models based on leave-one-mouse-out cross-validation. Each classifier was trained twice with and without applying WE-ASCA-based preprocessing. It turned out that removing the individual variations based on WE-ASCA improved the classification performance, and it significantly reduced the variance within the cross-validation results.

3.2 Data-Driven Modeling and Validation

This section briefly discusses the ML-based automatic detection and identification performed for three diagnostic tasks. In the first task, an improved ML pipeline to predict the antibiotic susceptibilities of *E. coli* bacteria was presented and evaluated based on images acquired by bright-field microscopy. Thereafter, transfer learning-based classification of bladder cancer was demonstrated in the second task using blue light cystoscopic images. Finally, different ML techniques and validation strategies were combined and checked in the third task. That task aimed to perform a label-free automatic detection of breast cancer based on a small-sized dataset of nonlinear multimodal images.

3.2.1 One-class Model-Based Antibiotics Susceptibility Prediction in Bacteria

The extensive and unwarranted application of antibiotics allowed many bacterial pathogens to developed new resistance mechanisms towards the existing drugs in the latest decades (45,132,133). As a result, the selection of an effective antibiotic to treat a specific bacterial species has become very complicated. Typically, the susceptibility determinations of bacterial pathogens are accomplished via antibiotic susceptibility testing (AST) (134). These ASTs need to be ideally rapid, accurate, and quantitative. Different technologies were recently developed to identify antibiotic susceptibility in bacteria; however, each of them features specific advantages and limitations in clinical application scenarios (135–137).

In the scientific contribution [PIII], an improved images-based automatic identification of bacterial susceptibilities toward antibiotics was presented using one-class classification models. Therein, a one-class support vector machine (OCSVM) was trained on images acquired from untreated controls of a specific bacterial strain, while the image labels of treated bacteria are predicted into control or non-control images. If a bacterial stain resists a specific antibiotic, it is expected that an image of these treated bacteria is predicted as control. In contrast, the bacterial strains sensitive to antibiotics show different morphology than the control untreated ones; and therefore, images collected after treating such bacteria with the antibiotics

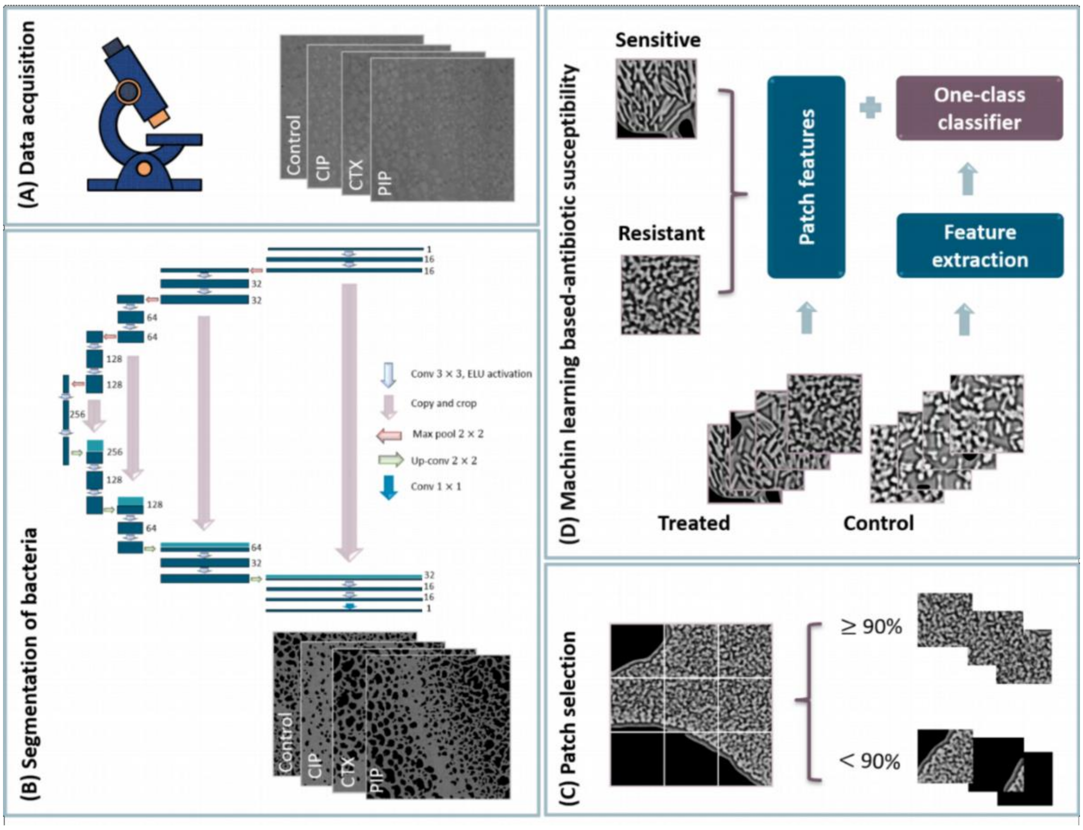


Figure 7. The automatic pipeline for predicting antibiotic resistances. (A) Images of cultivated bacteria, untreated control and treated bacteria with three antibiotics, are acquired using bright field microscopy. (B) The bacteria images are segmented using the U-Net network into high density bacteria regions and background. (C) The segmented bacteria images are sliced into patches of the size 265×256 pixels, and the patches that have 90% of their area covered by bacteria are considered for the statistical analysis. (D) The selected image patches of control bacteria are utilized to build one-class SVM (OCSVM) models. Lastly, the constructed OCSVM models are implemented to predict bacteria susceptibility towards the antibiotics using the extracted features from the selected patches of treated bacteria.

are expected to be identified as non-control. Under these assumptions, a complete pipeline for predicting the antibiotic susceptibility is presented as depicted in Figure 7. It starts by acquiring bright-field microscopic images of cultivated bacteria, untreated control, and treated bacteria with antibiotics. Subsequently, the collected bacteria images are segmented using the U-Net network into high-density bacteria regions and background (110). This U-Net network is a popular encoder-decoder CNN that has been constantly utilized for the semantic segmentation of biomedical tasks. After the image segmentation, the obtained images are sliced into patches, and patches with 90% of their area covered by bacteria are considered for further statistical modeling. In this case of study, the encoder part of the trained U-Net network is implemented as a feature extractor, then the extracted U-Net bottleneck features are utilized to predict the antibiotic susceptibility.

Using the proposed pipeline, the susceptibility detection of 12 *E. coli* strains towards three antibiotics, namely: ciprofloxacin, cefotaxime, and piperacillin, was performed based on the collected bright-field microscopic images. The results showed 83% area under the receiver operating characteristic (ROC) curve when OCSVM models were built on the U-Net bottleneck features of control bacteria images only. Moreover, the mean sensitivities of these one-class models are 91.67% and 86.61% for cefotaxime and piperacillin, respectively. In contrast, the classification means the sensitivity of ciprofloxacin is only 59.72% as the bacteria morphology was not fully detected based on the proposed method.

3.2.2 Deep Learning-Based Bladder Tumor Classification

Bladder cancer is one of the top 10 most frequently occurring cancers and one of the leading causes of death in Europe (138). To diagnose this cancer type, endoscopic techniques are commonly utilized (139,140). Recently, photodynamic diagnosis (PDD) based on blue light (BL) cystoscopy was introduced as a modern imaging technique for the detection of bladder cancer, especially for flat cancerous lesions (141–143). It offers characteristic information about tumor morphology based on the fluorescence properties of an extrinsic metabolic substrate, which metabolizes differently in cancerous tissues compared to healthy tissues (144). However, the main drawback of PDD is related to its low specificity in the differentiation between flat cancerous lesions and inflammable alterations after transurethral resection or instillation (141–143). Furthermore, due to the lack of experienced endoscopists, PDD-based image interpretation is quite subjective, leading to a high rate of false positives (145). Besides, PDD does not provide diagnostic information about cancer invasiveness or cancer grading. As a result, the applications of PDD are constrained to malignancy identification. Therefore, the aim of scientific work presented in [PIV] was to check the potential of deep learning models in automating the diagnosis of bladder cancer invasiveness and grading in addition to malignancy using the BL images only.

The outline of deep learning-based BL image classification is depicted in Figure 8. It starts by collecting images using BL cystoscopy, then preprocessing the acquired BL images according to a common image preprocessing pipeline. Therein, the illumination of the red and blue channels of all PDD images are corrected using the contrast limited adaptive histogram equalization (CLAHE) algorithm (85). Thereafter, the background area of each image is removed. Finally, the region of interest (ROI) of an image is acquired as an inscribed square region within the extracted image disk. The preprocessed images have the size of 384×384 pixels, and they refer to the image area containing the bladder tissue regions only. After image preprocessing, the obtained images are downsampled to the size of 224×224 pixels to fit the input size of the considered CNN architectures. In the presented analysis, four freely available pre-trained CNNs were fine-tuned by appending additional layers on top of each network, as shown in Figure 8-B. These CNNs are InceptionV3 network (146), MobileNetV2 network (147), ResNet50 network (148) and VGG16 network (149), and they were pre-trained on the ImageNet dataset (150). Nevertheless, the last additional layer of each fine-tuned CNN is the SoftMax layer, which offers label probabilities for each input image with respect to the considered classification task, *i.e.*, malignancy, cancer invasiveness and cancer grading.

The above-described outline was performed on a clinical dataset consisting of 216 BL images. These BL images were acquired prior to resection of the respective lesions from four urological departments retrospectively. Then, the collected biopsies were pathologically identified according to cancer malignancy, invasiveness, and grading. Meanwhile, two experienced urologists assessed the BL images only. The pre-trained fine-tuned CNNs were utilized subsequently to predict image labels using a leave-10-patients-out cross-validation (L10PO-CV). After predicting the BL image labels, the evaluation of each CNN performance was accomplished based on the classification mean sensitivity and mean specificity. Finally, the obtained metrics were compared to the urologist ratings.

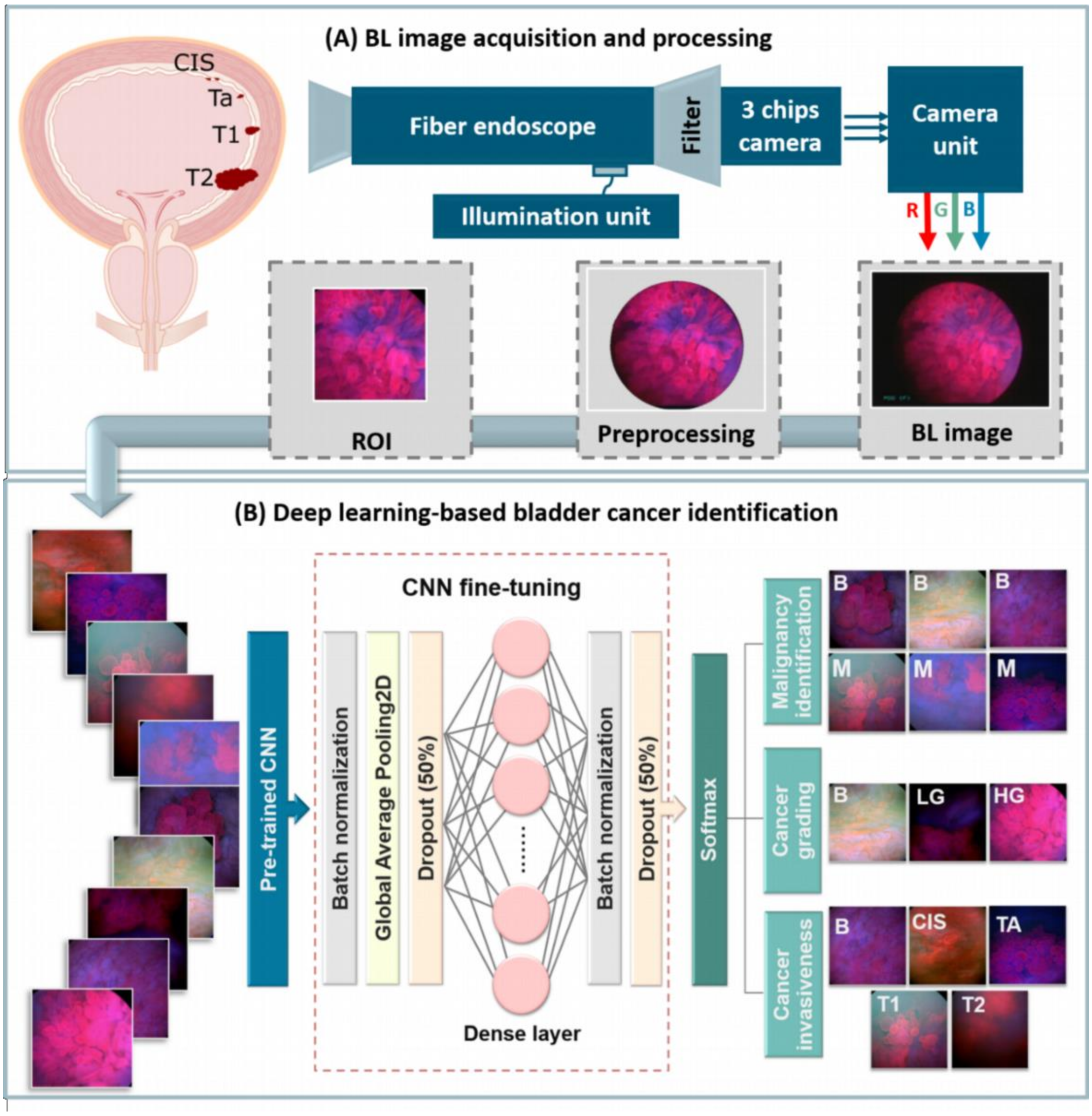


Figure 8. DL-based automatic classification of bladder cancer. (A) Cystoscopic images in the blue light mode are collected. These images are preprocessed, and then the background area of each image is excluded in order to get the image area, including bladder tissue only. (B) The obtained preprocessed images are resized and fed to the fine-tuned CNN. Here, the last additional layer of each CNNs, known as the SoftMax layer, provides label probabilities for each input image according to the considered classification task, namely: malignancy identification, the classification of cancer grading, and the identification of cancer invasiveness.

For the identification of malignant lesions, the fine-tuned MobileNetV2 showed the maximal sensitivity and specificity among the models. Thereby, the observed values of these statistics are 95.77% and 87.84%, respectively. Moving to the classification of tumor invasiveness, the fine-tuned MobileNetV2 also featured the best identification results. The mean sensitivity here is around 88%, and the mean specificity has a value of 96.56%. Furthermore, the detection of cancer stages Ta, T1 and T2 using the previous CNN has a class sensitivity of 93%, 100%, and 90.91%, respectively. For the identification of cancer grading, the maximum mean sensitivity is 92.07%, while the maximum mean specificity is 96.04%. These results were achieved when considering the fine-tuned ResNet50 network based on the L10PO-CV. In this case, the classification sensitivity of benign lesions is 95.95%, while the sensitivity of low-grade and high-grade cancer is 90.41% and 89.86%, respectively. Nevertheless, the identification ratings of both urologists were always much lower compared to classification performances of any of the previous CNNs.

3.2.3 Model Validation Strategies-Based Automatic Detection of Breast Cancer

Breast cancer is the most diagnosed cancer in women worldwide and the first cause of all female cancer deaths (151,152). According to the world health organization, breast cancer affects 2.1 million women yearly (153). The survival among patients of this cancer type largely depends on early detection, which is usually performed using imaging technologies in regular preventive checks (154). The challenge here is that breast cancer lacks early symptoms, while the current gold standard for definitive diagnosis is still the visual assessment of histopathological stained tissue sections after a biopsy of tissue material is taken. Therefore, new imaging technologies to enhance the low sensitivity of breast cancer screening and to supplement imaging technologies are highly appreciated. Ideally, these new tools permit a fast cancer diagnosis with a high potential for in-vivo investigations (155–157). Recently, the combination of coherent anti-Stokes Raman scattering (CARS), two-photon excited fluorescence (TPEF), and second harmonic generation (SHG) was introduced as a promising imaging tool. This combination provides a powerful label-free tool that can capture the biomolecular alterations of cancerous and noncancerous tissues.

To exploit the above-mentioned biomolecular alterations, the obtained multimodal images need to be translated into high-level diagnostic information. In the scientific contribution presented in [PV], the potential of computer-aided diagnosis was implemented to extract breast cancer-related information based on multimodal nonlinear images. Therein, several image preprocessing techniques were combined with ML algorithms to automatically detect breast cancer regions of 15 multimodal images acquired from 15 patients. The analysis pipeline started by preprocessing the multimodal images as described in Figure 2-B, then comparing them to the annotated Haematoxylin and Eosin (H&E) stain images. Subsequently, two classification models were trained using the deep convolutional neural network ResNet50 [99]. Here, the ResNet50 was utilized either to identify the labels of the image patches directly or to extract image features that can be used afterwards by classical machine learning classifiers. For evaluating the performance of the utilized classifiers, two data validation strategies were additionally investigated. These strategies refer to the leave-one-patient-out cross-validation (LOPO-CV) and the training-test validation. In Figure 9, an overview of the utilized classification techniques and the validation strategies is presented. For all presented strategies, the statistical independence between the training, the validation, and the test sets was secured based on the following rule: The patient images utilized to train a classifier are totally different from images acquired from the patients considered to validate or test the learned classifier. The results of the presented classification and validation strategies were assessed based on the classification mean sensitivity for the binary cancer diagnostic model, *i.e.*, cancer and non-cancer, and for a three-class model, *i.e.*, carcinoma, fat and normal. This diagnostic model evaluates the model quality in identifying the cancerous tissues. It turned out that the best detection of cancerous tissues was achieved by the fine-tuned ResNet50 network and the LOPO-CV. Thereby, the mean sensitivity of LOPO-CV using the fine-tuned ResNet50 network is 86.23%, which decreased to 75.31% if the PCA-LDA model was implemented. For the training-test validation, the images split into a training set, a validation set, and a test set. Then, this split was iterated three different times to check the effect of the data division on the classification results (see Figure 4-D). The obtained mean sensitivity, in this case, varied between 43.80% and 69.21%. Hence, the classification was strongly influenced by the chosen data subsets, *i.e.*, training set, validation set, and test set.

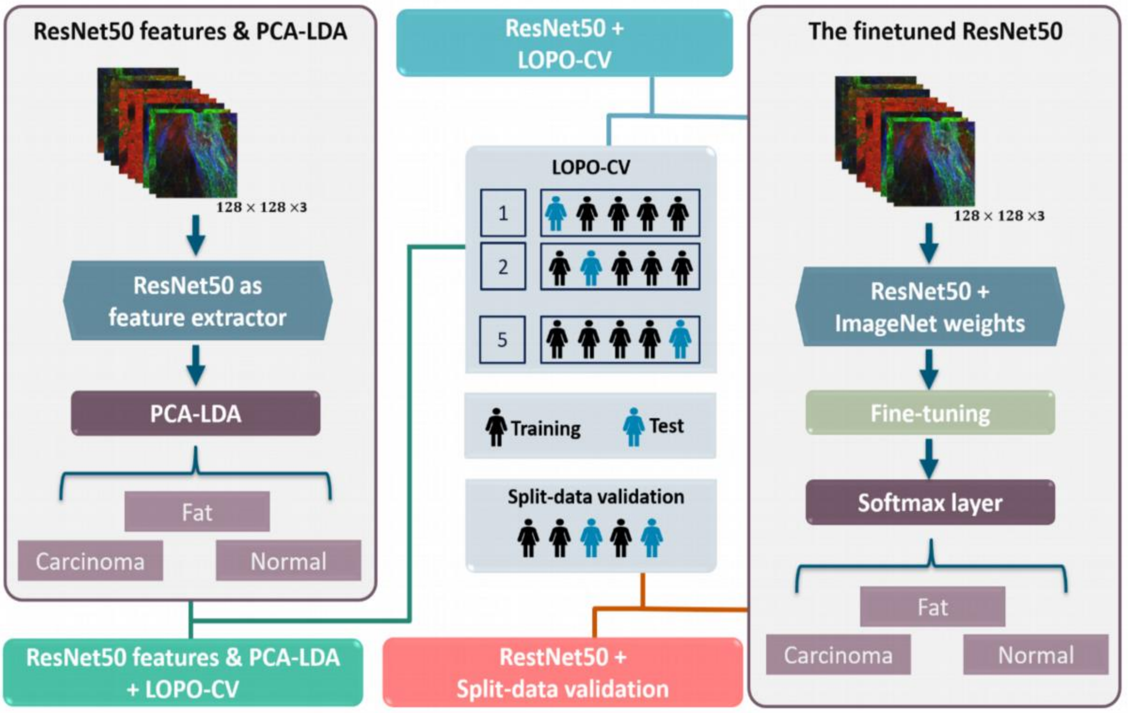


Figure 9. Overview of the ML algorithms and the validation strategies utilized to evaluate the automatic detection of breast cancer. The pre-trained ResNet50 network is either used as a feature extractor or fine-tuned to be used as a classification model. Besides, the leave-one-patient-out cross-validation and the training-test validation were implemented to evaluate the performance of the considered classifiers

All previous results were presented for the multimodal images that have corresponding annotated H&E images. However, introducing the presented nonlinear imaging technology into clinical routine needs to prove the diagnosis efficiency based on multimodal images of new patients without using the H&E annotation. To do so, the best performing ML models were deployed to predict the breast cancer patches of six multimodal images that missed their annotated H&E images. The fine-tuned ResNet50 network based on the LOPO-CV showed the best diagnostic performance among the presented ML models and validation strategies.

Therein, the best ResNet50 model was saved, and the patch labels of test patients were predicted for each iteration of the CV loop. It turned out that three models showed 100% prediction performance; therefore, these ResNet50 models were considered to identify breast cancer patches of six multimodal images. Despite the small-sized training set of ResNet50 and the overlapping of patch annotations, the results of patch prediction were very close to each other. Nevertheless, the obtained predictions still need to be verified, which was not available for the presented case of study.

Additional Work

Besides the scientific contributions being presented in Chapter 3, several publications from other studies conducted over the duration of the Ph.D. time are listed below:

[I] M.C. Kriegmair, A. Hartmann, T. Todenhöfer, N. Ali, G. Hipp, T. Knoll, P. Honeck, R. Oberneder, A. Stenzl, J. Popp, T. Bocklitz
Computer-assisted diagnosis during blue light cystoscopy using image analysis methods: Ahead of pathology?
European Urology Supplements, 2018, 17 (2), e1241

[II] S. Guo, O. Ryabchykov, N. Ali, R. Houhou, T. Bocklitz
Comprehensive Chemometrics. In Comprehensive Chemometrics: Chemical and Biochemical Data Analysis
Elsevier: Oxford, United Kingdom, 2020

[III] B. Lorenz, N. Ali, T. Bocklitz, P. Rösch, J. Popp
Discrimination between pathogenic and non-pathogenic E. coli strains by means of Raman microspectroscopy
Analytical and Bioanalytical Chemistry, 2020, 412, 8241–8247

[IV] J. Huang, N. Ali, E. Quansah, S. Guo, M. Noutsias, T. Meyer-Zedler, T. Bocklitz, J. Popp, U. Neugebauer, A. Ramoji
Vibrational Spectroscopic Investigation of Blood Plasma and Serum by Drop Coating Deposition for Clinical Application
International Journal of Molecular Sciences, 2021, 22 (4), 2191

Summary

The overall goal of data science in biophotonics is to improve the investigations of biological systems based on data collected using biophotonic technologies. In this context, several statistical tools can be combined with ML algorithems to enhance the experimental planning and then assess computer-aided identifications. Such statistical methods still need further investigations and adjustments in the case of life science and biomedicine-based studies. Therefore, the scientific contributions in Chapter 3 covered two main aspects within the data lifecycle for biophotonics: the design of statistical experiments and the implementation of data-driven modeling and validation.

Experimental design

The statistical techniques involved in designing the experiments deal with the planning and the analysis of controlled tests to evaluate the influence of experimental factors on selected data samples. In this thesis, the presented improvements related to experimental design were conducted to cover the sample size planning and the analysis of unbalanced multifactorial designs for multivariate data.

1- The designed algorithm of SSP aimed to estimate the number of samples required to train classification models. The presented algorithm started by generating LCs based on evaluating the classification performance as a function of an increasing set of training set sizes. Thereafter, the obtained LCs were fitted with the inverse power-law model, while its parameters were utilized to predict the training set size. Hither, the predicted size was calculated to describe 95% of the final classification performance. The last part of the SSP algorithm was designed to check the prediction performance based on comparing the behavior of the inverse power law model in the training region and the extrapolated region.

To evaluate the performance of the proposed SSP algorithm, a Raman-spectral dataset consisting of six bacterial species cultivated in nine independent biological replicates was considered. Thereby, the SSP for two levels of the data hierarchy was performed, while the focus was on the highest level of the data hierarchy, *i.e.*, biological replicates. Applying the proposed SSP algorithm showed that seven biological replicates are required to reach 95% of the final performance of the PCA-LDA model, which was introduced by the Bayes error rate. Moving to SSP for the required number of spectra, it turned out that 142 spectra per bacterial species are needed to achieve 95% of the final performance of the PCA-LDA model. The evaluation of both SSP tasks was carried out by calculating the RMSE of the extrapolated region and the training region. Nevertheless, the proposed SSP algorithm exhibited promising results for the prediction of the training set sizes required for both SSP tasks, *i.e.*, spectra and biological replicates. These predicted sizes were necessary to build a reliable and accurate PCA-LDA model. Although the SSP algorithm was performed on a Raman spectral dataset, the methodology can be utilized for any multivariate data, specifically in the case of biophotonic data. However, the estimations of sample size are strongly influenced by the experimental protocol, the considered data preprocessing techniques, and the utilized algorithm of statistical modeling. Subsequently, the estimated sample size is valid only for the same conditions, even though another analysis pipeline could require fewer or more measurements. The previous issue reflects the importance of the considered data-analysis pipeline for the sample size estimation. Therefore, the relation between these analytical pipelines and sample size estimation should be further investigated.

2- The weighted-effect ASCA (WE-ASCA) was presented as an extension of the classical (ASCA to analyze unbalanced multifactorial designs in the case of multivariate data. The main update of this ASCA version was to substitute the coding schemes of the design matrix in ASCA (or ASCA+) with the weighted-effect (WE) coding. This WE-coding is beneficial in such unbalanced designs as it uniquely estimates the effects of all factors considered in the designed model. Furthermore, it offers a zero value for the sum of all level effects in the design matrix, which is not the case when using other coding schemes.

Thus, the WE-coding was utilized instead of the deviation scheme implemented in the ASCA+ pipeline. In WE-ASCA, the response matrix is estimated based on a general linear model using the WE-coding-based design matrix and the calculated parameter matrix. Thereafter, the estimated response can be decomposed linearly as different effect matrices referring to the experimental factors and their interactions. Finally, the significant effects in the studied design are determined using a permutation test, while the dimensions of the effect matrices are reduced by applying PCA for each effect matrix. To infer the potential of the presented method, two possible applications were checked based on a Raman spectral dataset collected from colorectal tissues of 47 mice. The aims of the first application were to analyze the design of the studied experiment then to evaluate the performance of that analysis compared to the ASCA and the ASCA+. It tuned out that the classical ASCA overestimated the effect contributions, while the ASCA+ underestimated these contributions. In contrast to both, the proposed WE-ASCA showed the best performance with respect to the summation of the percentage of explained variances by effect contributions. Moving to the second application, the WE-ASCA was implemented as a preprocessing technique to exclude the disturbing variation presented in the Raman dataset. This was demonstrated for four different classification tasks using two classifiers and the leave-one-mouse-out cross-validation. The obtained results showed that excluding such variations from the training set introduced more robust classification results, and it improved the mean sensitivity in most classification tasks. In conclusion, the WE-ASCA was introduced as a powerful tool to analyze a complex unbalanced multifactorial design then to improve the classification performance and its reproducibility. Nonetheless, the WE-ASCA were checked only for Raman spectra for tissue classification tasks, but its applications are not limited. It can be expanded to cover the analysis of variance of any type of multivariate data and any statistical modeling task.

Data-driven modeling and validation

Recently, data-driven modeling based on ML algorithms has been implemented to extract high-level information by automating the extraction of data insights and inferring potential patterns within the acquired data. Therefore, the scientific contributions presented in this thesis were established to verify the capability of several ML models for three types of biophotonic data, *i.e.*, bright field microscopic images, fluorescence images, and nonlinear multimodal images.

1- Based on combining different image preprocessing techniques and ML algorithms, the automatic identification of antibiotics susceptibilities was presented for bacterial images collected from bright field microscopy. The proposed pipeline was designed to capture any morphological changes caused by applying antibiotics. It started by segmenting the bacterial regions using an autoencoder CNN, named the U-Net network. After that, the encoder part of the trained U-Net network was utilized as a feature extractor of the bacterial images. In the last step, a one-class classification model, specifically an OCSVM model, was implemented for the first time to detect the antibiotic effects on the bacterial strains. Thereby, the OCSVM was trained only on images acquired from control untreated bacteria, and then the trained model was utilized to predict the antibiotic resistance of treated bacteria cultivated within the same replicate. It turned out that the local OCSVM models introduced quite promising results in identifying the susceptibility of *E. coli*; hence, these models are self-correcting for the biological variations between different replicates or patients. Besides, such local-one-class classification is easy to apply for identifying any other antibiotic susceptibilities and for any image-based antibiotic susceptibility test (AST). Finally, this image-based method can be used as a fast-phenotypic AST as the morphological changes appear after short incubation times of antibiotics with bacteria. However, combining the image-based AST with other readout methods could improve the results of this detection.

2- The BL cystoscopy-based photodynamic diagnosis was introduced as a promising technology to improve the detection of bladder cancer. In the proposed scientific contribution, a BL image-based deep learning diagnostic platform was presented in order to predict the bladder cancer malignancy, invasiveness, and grading based on the BL images only. The potential of that platform in automating the classification of the endoscopic lesions and predicting histopathological results was checked using a small-sized dataset of BL images acquired from four different urological departments. Therein, the performance of four DL models was compared with the identification results provided by two experienced urologists for the three considered tasks. Despite the small sample size and the class imbalance in the BL dataset, the obtained results of these comparisons exhibited a high identification performance of DL-based transfer learning. For all tasks, the fine-tuned CNNs provided much better classification performance than both urologists. Moreover, the misclassification of BL images was expected due to the high variations between the images and other systematic errors. These errors were assigned mostly to specific fluorescence issues like the very low image fluorescence or the spotty fluorescence in other images. Besides, some images depicted flat lesions, while others were not close enough to capture the suspicious lesions; and consequently, such images were also misclassified. Overall, the presented study showed the promising potential of DL-based classification models for the diagnostics of bladder cancer when using the BL cystoscopic images only. However, further research needs to be performed in order to establish a fully automatic BL cystoscopic platform. The aim, in this case, should be to assist surgeons and aid the cancer diagnoses by offering a faster and lower-cost alternative of the classical biopsy-based pathological analysis.

3- The nonlinear multimodal imaging technologies provide a label-free tool that can offer a non-invasive characterization of the biomolecular alternations between cancerous and noncancerous tissues. The advantage of these technologies was used to detect breast cancer tissues based on the multimodal images only. To do so, 16 multimodal images of breast tissue acquired from 16 patients were considered. The challenge in that study was to translate the biomolecular information introduced by these images into an ML model that can be deployed in further identifications. Therefore, an image preprocessing pipeline was designed to enhance image quality, then three combinations of ML models and validation methods were checked. The best classification performance was achieved when using the pre-trained ResNet50 network as a classification model and the leave-one-patient-out cross-validation. Therefore, the best performing models within the CV loop were considered to detect cancerous and noncancerous tissues of not annotated multimodal images. In most cases, these classification models provided the same predictions of the multimodal image patches. Although these results were not validated and the training set was quite small, it was still possible to deploy ML models for the automatic diagnosis of breast cancer. Nevertheless, the non-invasive nature of the nonlinear imaging modalities allows for in-vivo examinations offering a low-risk diagnostic tool to supplement others.

Zusammenfassung

Das übergeordnete Ziel biophotonischer Datenwissenschaft ist die Verbesserung von Untersuchungen biologischer Systeme auf Grundlage von Daten, die mit biophotonischen Technologien gemessen wurden. In diesem Zusammenhang können verschiedene statistische Werkzeuge mit maschinellen Lern-Algorithmen (ML) kombiniert werden, um so die Versuchsplanung zu verbessern und anschließend eine computergestützte Identifikation durchzuführen. Diese statistischen Methoden bedürfen noch weiterer Erforschung und Anpassungen für biowissenschaftliche und biomedizinische Studien. Daher befassen sich die wissenschaftlichen Beiträge in Kapitel 3 mit zwei Hauptaspekten innerhalb des Datenlebenszyklus biophotonischer Daten: der statistischen Versuchsplanung und der Umsetzung datengetriebener Modellbildung sowie der Modellvalidierung.

Versuchsplanung

Die statistischen Techniken, die bei der Versuchsplanung zum Einsatz kommen, werden genutzt, um die Planung und Analyse von kontrollierten Versuchen zum Einfluss experimenteller Faktoren durchzuführen. In dieser Arbeit wurden Verbesserungen im Zusammenhang mit der Versuchsplanung erforscht, um die Planung des Stichprobenumfangs multivariater Studien durchzuführen und die Analyse von nicht balancierten multifaktoriellen Versuchsplänen für multivariate Daten zu erlauben.

1- Der entworfene Algorithmus zur Stichproben-Planung (SSP-Algorithmus) zielte darauf ab, den erforderlichen Stichprobeumfang für das Training von Klassifikationsmodellen zu schätzen. Der vorgestellte Algorithmus beginnt mit der Generierung der Lernkurve (LC), welche die Klassifizierungsleistung in Abhängigkeit von einer zunehmenden Anzahl von Trainingsdatensätzen quantifiziert. Anschließend werden die erhaltenen LCs mit dem inversen Potenzgesetzmodell gefittet, und die Fit-Parameter werden zur Vorhersage der

Trainingsdatensatzgröße verwendet. Dabei wurde die Trainingsdatensatzgröße so berechnet, dass sie zu 95 % der endgültigen Klassifizierungsleistung führt. Um die Vorhersageleistung zu überprüfen, wurde im letzten Teil des SSP-Algorithmus das Verhalten des inversen Potenzgesetzmodells in der Trainingsregion und in der extrapolierten Region verglichen. Um die Leistung des vorgeschlagenen SSP-Algorithmus zu bewerten, wurde ein Raman-Spektraldatensatz bestehende aus sechs Bakterienarten und neun unabhängigen biologischen Replikaten betrachtet. Dabei wurde die Fahlzahlplanung für zwei Ebenen der Datenhierarchie durchgeführt, wobei der Schwerpunkt auf der höchsten Ebene der Datenhierarchie, das heißt den biologischen Replikaten, lag. Die Anwendung des vorgeschlagenen SSP-Algorithmus zeigte, dass sieben biologische Replikate erforderlich sind, um 95 % der endgültigen Leistung des PCA-LDA-Modells (Hauptkomponenten-Analyse in Kombination mit einer Linearen Diskriminanz-Analyse) zu erreichen. Die finale Leistung des Modells kann durch die Bayes-Fehlerrate charakterisiert werden. Bei der Anwendung des Algorithmus zur Bestimmung der erforderlichen Spektren-Anzahl zeigte sich, dass 142 Spektren pro Bakterienart erforderlich sind, um 95 % der endgültigen Leistung des PCA-LDA-Modells zu erreichen. Die Bewertung der beiden SSP-Aufgaben erfolgte durch Berechnung des RMSE in der extrapolierten Region und der Trainingsregion. Es zeigte sich, dass der vorgeschlagene SSP-Algorithmus beide erforderlicher Trainingsmengen vorhersagen konnte. Diese vorhergesagten Fallzahlen waren notwendig, um ein zuverlässiges und genaues PCA-LDA-Modell zu erstellen. Obwohl der SSP-Algorithmus an einem Raman-Spektraldatensatz erstellt und getestet wurde, kann die Methodik für jeden multivariaten Datensatz verwendet werden, insbesondere im Fall von biophotonischen Daten. Die Schätzung des Stichprobenumfangs wird jedoch stark durch das Versuchsprotokoll, die verwendeten Datenvorverarbeitungstechniken und das verwendete statistische Modell beeinflusst. Folglich gilt der geschätzte Stichprobenumfang nur für dieselben Bedingungen, auch wenn eine andere Analysepipeline weniger oder mehr Messungen erfordern könnte. Das vorstehende Problem spiegelt die Bedeutung der betrachteten Datenanalyse-Pipeline für die Schätzung des Stichprobenumfangs wider. Daher sollte die

Beziehung zwischen der Analysepipeline und der Schätzung des Stichprobenumfangs weiter untersucht werden.

2- Die *ANOVA simultaneous component analysis* (ASCA) mit gewichteten Effekten (WE-ASCA) wurde als Erweiterung der klassischen ASCA entwickelt, um unausgewogene multifaktorielle Designs im Falle multivariater Daten analysieren zu können. Die wichtigste Neuerung dieser ASCA-Version besteht darin, das Kodierungsschema der Designmatrix durch die Kodierung mit gewichteten Effekten (WE) zu ersetzen. Diese WE-Kodierung ist in nicht-balancierten Designs vorteilhaft, da sie die Effekte aller im entworfenen Modell berücksichtigten Faktoren eindeutig schätzt. Darüber hinaus bietet sie einen Nullwert für die Summe aller Niveaueffekte in der Designmatrix, was bei der Verwendung anderer Kodierungsschemata nicht der Fall ist. Daher wurde die WE-Kodierung anstelle des in der ASCA+ Pipeline implementierten Abweichungsschemas verwendet. In WE-ASCA wird die Antwortmatrix auf der Grundlage eines allgemeinen linearen Modells (GLM) sowie unter Verwendung der WE-Kodierungsbasierenden Designmatrix und der berechneten Parametermatrix geschätzt. Danach kann die geschätzte Antwortmatrix linear in verschiedene Effektmatrizen zerlegt werden, die sich auf die experimentellen Faktoren und ihre Interaktionen beziehen. Schließlich werden die signifikanten Effekte im untersuchten Design mit Hilfe eines Permutationstests bestimmt, während die Dimension der Effektmatrizen durch Anwendung einer Hauptkomponenten-Analysis (PCA) für jede Effektmatrix reduziert wird. Um das Potenzial der vorgestellten Methode zu ermitteln, wurden zwei mögliche Anwendungen anhand eines Raman-Spektraldatensatzes von Mausdarmgewebe geprüft. Das Ziel der ersten Anwendung war es, das Design des untersuchten Experiments zu analysieren und dann die Leistung dieser Analyse im Vergleich zu ASCA und ASCA+ zu bewerten. Es stellte sich heraus, dass die klassische ASCA die Effektbeiträge überschätzte, während die ASCA+ diese Beiträge unterschätzte. Im Gegensatz zu beiden existierenden ASCA Versionen zeigte die vorgeschlagene Methode (WE-ASCA) die beste Leistung in Bezug auf die Summierung der erklärten Varianzen der Effektbeiträge. In der zweiten Anwendung wurde die WE-ASCA Methode als Vorverarbeitungstechnik implementiert, um störende Variationen im

Raman-Datensatz auszuschließen. Dies wurde für vier verschiedene Klassifikationsaufgaben unter Verwendung von zwei Klassifikatoren und einer Kreuzvalidierung (auf Individuen-Ebene) demonstriert. Die erzielten Ergebnisse zeigten, dass die Entfernung solcher Variationen aus dem Trainingssatz zu robusteren Klassifizierungsergebnissen führt, und die gemittelte Sensitivität der Modelle bei den meisten Klassifizierungsaufgaben etwas verbessert wurde. Zusammenfassend lässt sich sagen, dass die WE-ASCA Methode als leistungsfähiges Instrument zur Analyse eines komplexen unausgewogenen multifaktoriellen Designs generiert wurde, um die Klassifizierungsleistung und ihre Reproduzierbarkeit zu verbessern. Die WE-ASCA wurde nur anhand Raman-Spektren zur Gewebeklassifizierung getestet, aber ihre Anwendung ist nicht auf diese Aufgaben beschränkt. Sie kann auf jede Varianzanalyse multivariater Daten und jede statistische Modellierungsaufgabe erweitert werden.

Datengestützte Modellierung und Validierung

In den letzten Jahren wurde die Daten-basierende Modellierung auf Grundlage von ML-Algorithmen eingeführt, um höhere Informationen aus Daten zu extrahieren. Diese ML-Verfahren extrahieren Datenerkenntnissen durch die Bestimmung von Mustern in den erfassten Daten. Die in dieser Arbeit vorgestellten wissenschaftlichen Beiträge wurden daher erstellt, um die Fähigkeit verschiedener ML-Modelle für drei Arten von biophotonischen Daten zu überprüfen, das heißt für mikroskopische Hellfeldbilder, Fluoreszenzbilder und nichtlineare multimodale Bilder.

1- Basierend auf der Kombination verschiedener Bildvorverarbeitungstechniken und ML-Algorithmen wurde eine automatische Bestimmung von bakteriellen Antibiotika-Resistenzen mittels Hellfeldmikroskopie-Bildern vorgestellt. Die vorgeschlagene Datenanalyse-Pipeline wurde entwickelt, um morphologischen Veränderungen der Bakterien durch die Anwendung von Antibiotika zu erfassen. Die Pipeline beginnt mit der Segmentierung von Bildregionen, die von Bakterien dominiert werden, unter Verwendung eines auf *Convolutional Neural Networks* (CNNs) basieren Autoencoders, dem sogenannten U-Net-Modell. Danach wurde der Kodierungsteil des trainierten U-Netzes als Merkmalsextraktor für die Bilder verwendet. Im letzten Schritt wurde zum ersten Mal ein

Einklassen-Klassifikationsmodell, genauer gesagt ein OCSVM-Modell (*One-Class-Support-Vector-Machine*-Modell), implementiert, um die Wirkungen von Antibiotika auf Bakterien zu erkennen. Dabei wurde das OCSVM-Modell nur mit Bildern von unbehandelten Kontrollbakterien trainiert, und dann wurde das trainierte Modell zur Vorhersage der Antibiotikaresistenz von behandelten Bakterien verwendet, welche im selben Replikat wie die Kontroll-Bakterien kultiviert wurden. Es stellte sich heraus, dass die lokalen OCSVM-Modelle vielversprechende Ergebnisse bei der Bestimmung der Antibiotika-Sensitivitäten von *E. coli* lieferten, da diese Modelle automatisch für die biologischen Variationen zwischen verschiedenen Replikaten oder Patienten korrigieren. Außerdem ist eine solche lokale Einklassen-Klassifizierung leicht für die Detektion anderer Antibiotika-Sensitivitäten und für jeden bildbasierten Antibiotika-Empfindlichkeitstest (AST) anwendbar. Schließlich kann diese bildbasierte Methode als schneller phänotypischer AST verwendet werden, da die morphologischen Veränderungen der Bakterien nach kurzen Inkubationszeiten mit den Antibiotika auftreten. Die Kombination des bildbasierten AST mit anderen Auslesemethoden könnte die Ergebnisse dieses Nachweises noch verbessern.

2- Die photodynamische Diagnose mittels der Blau-Licht-Zystoskopie (BL-Zystoskopie) wurde als vielversprechende Technologie zur verbesserten Erkennung von Blasenkrebs eingeführt. Im vorgeschlagenen wissenschaftlichen Beitrag wurde eine auf tiefen Lernverfahren basierende Diagnoseplattform für BL-Bildern vorgestellt, um die Bösartigkeit, Invasivität und Graduierung von Blasenkrebs vorherzusagen. Das Potenzial dieser Plattform bei der automatischen Klassifizierung endoskopischer Läsionen und der Vorhersage histopathologischer Ergebnisse wurde anhand eines kleinen Datensatzes von BL-Bildern aus vier verschiedenen urologischen Abteilungen überprüft. Dabei wurde die Leistung von vier tiefen Lernmodellen mit den Erkennungsergebnissen von zwei erfahrenen Urologen für die oben genannten drei Aufgaben verglichen. Trotz der geringen Stichprobengröße und eines starken Klassenungleichgewichts im BL-Datensatz zeigten die Ergebnisse eine hohe Identifikationsleistung der tiefen Lernverfahren durch die Anwendung von Transferlernen. Bei allen Aufgaben lieferte das Fine-tunning der CNNs deutlich bessere Klassifikationsleistungen als die Vorhersagen der beiden Urologen. Darüber hinaus waren die Fehlklassifizierungen von BL-Bildern aufgrund der hohen Variationen zwischen den Bildern und anderer systematischer Fehler zu erwarten. Es traten Fluoreszenzproblemen auf, wie zum Beispiel eine sehr geringe Fluoreszenzintensität in manchen Bildern oder eine fragmentierte Fluoreszenz in anderen Bildern. Außerdem zeigten einige Bilder flache Läsionen, während andere nicht nah genug an die verdächtigen Läsionen heranreichten, so dass auch diese Bilder falsch klassifiziert wurden. Insgesamt zeigte die vorgestellte Studie das vielversprechende Potenzial tiefer Lernmodelle für die Diagnose von Blasenkrebs basierend BL-Zystoskopie-Bildern. Es besteht jedoch noch weiterer Forschungsbedarf, um eine vollautomatische BL-Zystoskopie-Plattform zu etablieren. Ziel sollte es sein, Chirurgen zu unterstützen und die Krebsdiagnose zu erleichtern, indem eine schnellere und kostengünstigere Alternative zur klassischen Biopsie-basierten pathologischen Analyse angeboten wird.

3- Die nichtlineare multimodale Bildgebung stellt ein markerfreies Werkzeug dar, welches eine nicht-invasive Charakterisierung von biomolekularen Veränderungen zwischen krebsartigem und nicht krebsartigem Gewebe ermöglicht. Diese Technologien wurde genutzt, um Brustkrebsgewebe allein auf der Grundlage der multimodalen Bilder zu erkennen. Zu diesem Zweck wurden 16 multimodale Bilder von Brustgewebe von 16 Patientinnen untersucht. Die Herausforderung in dieser Studie bestand darin, die biomolekularen Informationen, die diese Bilder liefern, in ein ML-Modell zu übersetzen, das für weitere Identifizierungen eingesetzt werden kann. Daher wurde eine Bildvorverarbeitungspipeline entwickelt, um die Bildqualität zu verbessern. Anschließend wurden drei Kombinationen von ML-Modellen und Validierungsmethoden geprüft. Die beste Klassifizierungsleistung wurde bei der Verwendung des vortrainierten ResNet50-Netzwerks als Klassifizierungsmodell und der Kreuzvalidierung (Leave-one-patient-out) erzielt. Die leistungsfähigsten Modelle wurden genutzt, um innerhalb der CV-Schleife nicht annotierte multimodale Bilder vorherzusagen. In den meisten Fällen lieferten diese Klassifikationsmodelle (in der Schleife) stabile Vorhersagen für die nicht-annotierten multimodalen Bilder. Obwohl diese Ergebnisse nicht validiert werden konnten und die

Trainingsmenge recht klein war, war es dennoch möglich, ML-Modelle für die automatische Diagnose von Brustkrebs zu erstellen. Die nicht-invasive Natur der nicht-linearen Bildgebungsmodalitäten ermöglicht es perspektivisch In-vivo-Untersuchungen durchzuführen, die ein risikoarmes Diagnoseinstrument zur Ergänzung anderer Diagnose-Methoden darstellen.

Bibliography

1. Keiser G. Biophotonics: Concepts to Applications. Singapore: Springer; 2016. (Graduate Texts in Physics).

2. Marion Jüurgens TM and Jürgen Popp. Introduction to Biophotonics. In: Jürgen Popp VVT Arthur Chiou, Stefan H Heinemann, editor. Handbook of Biophotonics [Internet]. Berlin: Wiley-VCH; 2012. p. 1–38. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527643981.bphot001

3. Bocklitz TW, Guo S, Ryabchykov O, Vogler N, Popp J. Raman Based Molecular Imaging and Analytics: A Magic Bullet for Biomedical Applications!? Anal Chem [Internet]. 2016 Jan 5;88(1):133–51. Available from: https://doi.org/10.1021/acs.analchem.5b04665

4. Zhang Y-Q, Rajapakse JC. Machine learning in bioinformatics. Vol. 4. Wiley Online Library; 2009.

5. Goda K. Biophotonics and beyond. APL Photonics [Internet]. 2019;4(5):050401. Available from: https://aip.scitation.org/doi/abs/10.1063/1.5100614

6. Pluta M, Maksymilian P. Advanced light microscopy. Vol. 1. Amsterdam: Elsevier; 1988.

7. Sanderson J. Introduction to Light Microscopy. J Microsc [Internet]. 1999 Jan 1 [cited 2020 Nov 17];193(1):90–1. Available from: https://doi.org/10.1046/j.1365-2818.1999.0429b.x

8. Zernike F. Phase contrast, a new method for the microscopic observation of transparent objects. Physica [Internet]. 1942 Jul 1;9(7):686–98. Available from: http://www.sciencedirect.com/science/article/pii/S003189144280035X

9. Allen RD, David GB, Nomarski G. The zeiss-Nomarski differential interference equipment for transmitted-light microscopy. Z Wiss Mikrosk. 1969/11/01 ed. 1969 Nov;69(4):193–221.

10. Coons AH, Kaplan MH. Localization of antigen in tissue cells; improvements in a method for the detection of antigen by means of fluorescent antibody. J Exp Med. 1950/01/01 ed. 1950 Jan 1;91(1):1–13.

11. Sahoo H. Fluorescent labeling techniques in biomolecules: a flashback. RSC Adv [Internet]. 2012;2(18):7017–29. Available from: http://dx.doi.org/10.1039/C2RA20389H

12. Gonçalves MST. Fluorescent Labeling of Biomolecules with Organic Probes. Chem Rev [Internet]. 2009 Jan 14;109(1):190–212. Available from: https://doi.org/10.1021/cr0783840

13. Laura M, Stephen AB, Mark RH, Jürgen P, Brian CW. Biophotonics: the big picture. J Biomed Opt [Internet]. January 12;23(2):1–7. Available from: https://doi.org/10.1117/1.JBO.23.2.021103

14. Low AF, Tearney GJ, Bouma BE, Jang I-K. Technology Insight: optical coherence tomography—current status and future development. Nat Clin Pract Cardiovasc Med [Internet]. 2006 Mar 1;3(3):154–62. Available from: https://doi.org/10.1038/ncpcardio0482

15. Goetz M, Malek NP, Kiesslich R. Microscopic imaging in endoscopy: endomicroscopy and endocytoscopy. Nat Rev Gastroenterol Hepatol [Internet]. 2014 Jan 1;11(1):11–8. Available from: https://doi.org/10.1038/nrgastro.2013.134

16. Rafii-Tari H, Payne CJ, Yang G-Z. Current and emerging robot-assisted endovascular catheterization technologies: a review. Ann Biomed Eng. 2014;42(4):697–715.

17. Popp J, Tuchin VV, Chiou A, Heinemann SH. Handbook of biophotonics: Vol. 2: Photonics for health care. Vol. 2. John Wiley & Sons; 2011.

18. Stuart BH. Infrared spectroscopy: Fundamental and applications. Chichester: John Wiley & Sons; 2004. (Analytical Techniques in the Sciences).

19. Workman Jr J, Springsteen A. Applied spectroscopy: a compact reference for practitioners. Academic Press; 1998.

20. McCreery RL. Raman spectroscopy for chemical analysis. Vol. 225. John Wiley & Sons; 2005.

21. Humecki HJ. Practical guide to infrared microspectroscopy. CRC Press; 1995.

22. Schmitt M, Popp J. Raman spectroscopy at the beginning of the twenty-first century. J Raman Spectrosc [Internet]. 2006 Jan 1 [cited 2020 Nov 17];37(1-3):20–8. Available from: https://doi.org/10.1002/jrs.1486

23. Raman CV, Krishnan KS. A new type of secondary radiation. Nature. 1928;121(3048):501–2.

24. David B, Donald DD, Evan RH, Sean JK, Marcus L, Wiendelt S, et al. Laser speckle contrast imaging: theoretical and practical limitations. J Biomed Opt [Internet]. January 6;18(6):1–10. Available from: https://doi.org/10.1117/1.JBO.18.6.066018

25. Guolan L, Baowei F. Medical hyperspectral imaging: a review. J Biomed Opt [Internet]. January 1;19(1):1–24. Available from: https://doi.org/10.1117/1.JBO.19.1.010901

26. Boyd RW. Nonlinear optics. Academic press; 2003.

27. Zhang C, Cheng J-X. Perspective: Coherent Raman scattering microscopy, the future is bright. APL Photonics [Internet]. 2018 Sep 1 [cited 2020 Nov 17];3(9):090901. Available from: https://doi.org/10.1063/1.5040101

28. Diaspro A, Chirico G, Collini M. Two-photon fluorescence excitation and related techniques in biological microscopy. Q Rev Biophys [Internet]. 2006/02/16 ed. 2005 May;38(2):97–166. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16478566

29. Wang BG, Konig K, Halbhuber KJ. Two-photon microscopy of deep intravital tissues and its merits in clinical research. J Microsc [Internet]. 2010/04/14 ed. 2010 Apr 1;238(1):1–20. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20384833

30. Mohler W, Millard AC, Campagnola PJ. Second harmonic generation imaging of endogenous structural proteins. Methods [Internet]. 2003/01/25 ed. 2003 Jan;29(1):97–109. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12543075

31. Krafft C, Dietzek B, Popp J, Schmitt M. Raman and coherent anti-Stokes Raman scattering microspectroscopy for biomedical applications. J Biomed Opt. 2012;17(4):040801.

32. Rodrigues AF, Newman L, Lozano N, Mukherjee SP, Fadeel B, Bussy C, et al. A blueprint for the synthesis and characterisation of thin graphene oxide with controlled lateral dimensions for biomedicine. 2D Mater. 2018;5(3):035020.

33. Shakibaie F, George R, Walsh L. Applications of laser induced fluorescence in dentistry. Int J Dent Clin. 2011;3(3):38–44.

34. Spyratou E, Makropoulou M, Mourelatou E, Demetzos C. Biophotonic techniques for manipulation and characterization of drug delivery nanosystems in cancer therapy. Cancer Lett. 2012;327(1–2):111–22.

35. Krafft C. Modern trends in biophotonics for clinical diagnosis and therapy to solve unmet clinical needs. J Biophotonics. 2016;9(11–12):1362–75.

36. Watson TF, Atmeh AR, Sajini S, Cook RJ, Festy F. Present and future of glass-ionomers and calcium-silicate cements as bioactive materials in dentistry: biophotonics-based interfacial analyses in health and disease. Dent Mater. 2014;30(1):50–61.

37. Hara AK, Leighton JA, Sharma VK, Fleischer DE. Small bowel: preliminary comparison of capsule endoscopy with barium study and CT. Radiology. 2004;230(1):260–5.

38. Vankeirsbilck T, Vercauteren A, Baeyens W, Van der Weken G, Verpoort F, Vergote G, et al. Applications of Raman spectroscopy in pharmaceutical analysis. TrAC Trends Anal Chem. 2002;21(12):869–77.

39. Esmonde-White KA, Cuellar M, Uerpmann C, Lenain B, Lewis IR. Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing. Anal Bioanal Chem. 2017;409(3):637–49.

40. Zloh M. NMR spectroscopy in drug discovery and development: Evaluation of physico-chemical properties. ADMET DMPK. 2019;7(4):242–51.

41. Wang W, Zhang H, Yuan Y, Guo Y, He S. Research progress of Raman spectroscopy in drug analysis. AAPS PharmSciTech. 2018;19(7):2921–8.

42. Patra B, Peng C-C, Liao W-H, Lee C-H, Tung Y-C. Drug testing and flow cytometry analysis on a large number of uniform sized tumor spheroids using a microfluidic device. Sci Rep. 2016;6:21061.

43. Wong FH-S, Cai Y, Leck H, Lim T-P, Teo JQ-M, Lee W, et al. Determining the Development of Persisters in Extensively Drug-Resistant Acinetobacter baumannii upon Exposure to Polymyxin B-Based Antibiotic Combinations Using Flow Cytometry. Antimicrob Agents Chemother. 2020;64(3).

44. Huang Y, Song C, Li H, Zhang R, Jiang R, Liu X, et al. Cationic conjugated polymer/hyaluronan-doxorubicin complex for sensitive fluorescence detection of hyaluronidase and tumor-targeting drug delivery and imaging. ACS Appl Mater Interfaces. 2015;7(38):21529–37.

45. Kirchhoff J, Glaser U, Bohnert JA, Pletz MW, Popp J, Neugebauer U. Simple Ciprofloxacin Resistance Test and Determination of Minimal Inhibitory Concentration within

2 h Using Raman Spectroscopy. Anal Chem [Internet]. 2018 Feb 6;90(3):1811–8. Available from: https://doi.org/10.1021/acs.analchem.7b03800

46. Pradhan P, Guo S, Ryabchykov O, Popp J, Bocklitz TW. Deep learning a boon for biophotonics? J Biophotonics [Internet]. :e201960186. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jbio.201960186

47. Cash P, Stanković T, Štorga M. An Introduction to Experimental Design Research. 2016;3–12.

48. Horvath I. Theory Building in Experimental Design Research. 2016;

49. Farooq MA, Nóvoa H, Araújo A, Tavares SMO. An innovative approach for planning and execution of pre-experimental runs for Design of Experiments. Eur Res Manag Bus Econ [Internet]. 2016;22(3):155–61. Available from: http://www.sciencedirect.com/science/article/pii/S1135252315000064

50. DePoy E, Gitlin LN. Chapter 10 - Experimental-Type Designs. In: DePoy E, Gitlin LN, editors. Introduction to Research (Fifth Edition) [Internet]. Mosby; 2016. p. 134–57. Available from: http://www.sciencedirect.com/science/article/pii/B9780323261715000100

51. Steinberg DM, Hunter WG. Experimental Design: Review and Comment. Technometrics [Internet]. 1984;26(2):71–97. Available from: https://www.tandfonline.com/doi/abs/10.1080/00401706.1984.10487928

52. Mead R, Curnow RN, Hasted AM. Statistical methods in agriculture and experimental biology: Third edition. Statistical Methods in Agriculture and Experimental Biology: Third Edition. 2017. 1–472 p.

53. Durakovic B. Design of experiments application, concepts, examples: State of the art. Vol. 5, Periodicals of Engineering and Natural Sciences. 2017. 421–439 p.

54. St»hle L, Wold S. Analysis of variance (ANOVA). Chemom Intell Lab Syst [Internet]. 1989 Nov 1;6(4):259–72. Available from: http://www.sciencedirect.com/science/article/pii/0169743989800954

55. R. Sampford M, G. Cochran W. Sampling Techniques. Vol. 34, Biometrics. 1978. 332 p.

56. Christensen R. Analysis of Variance and Generalized Linear Models. In: Smelser NJ, Baltes PB, editors. International Encyclopedia of the Social & Behavioral Sciences [Internet]. Oxford: Pergamon; 2001. p. 473–80. Available from: http://www.sciencedirect.com/science/article/pii/B0080430767004526

57. V. Mardia K, Kent J, Bibby J. Multivariate Analysis. Vol. 37, Probability and Mathematical Statistics, London: Academic Press, 1979. 1979.

58. Martens M. Multivariate Analysis of Quality. An Introduction. Vol. 12, Measurement Science & Technology - MEAS SCI TECHNOL. 2001. 1746–1746 p.

59. Bratchell N. Multivariate response surface modelling by principal components analysis. J Chemom [Internet]. 1989;3(4):579–88. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.1180030406

60. Jansen JJ, Hoefsloot HCJ, van der Greef J, Timmerman ME, Westerhuis JA, Smilde AK. ASCA: analysis of multivariate data obtained from an experimental design. J Chemom [Internet]. 2005;19(9):469–81. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.952

61. Smilde AK, Jansen JJ, Hoefsloot HCJ, Lamers R-JAN, van der Greef J, Timmerman ME. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. Bioinformatics [Internet]. 2005 [cited 2019 May 15];21(13):3043–8. Available from: https://doi.org/10.1093/bioinformatics/bti476

62. Smilde AK, Hoefsloot HuubCJ, Westerhuis JohanA. The geometry of ASCA. J Chemom [Internet]. 2008;22(8):464–71. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.1175

63. Anderson M, ter Braak C. Permutation tests for multi-factorial analysis of variance. J Stat Comput Simul. 10;73:85–113.

64. Thiel M, Féraud B, Govaerts B. ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. J Chemom [Internet]. 2017;31(6):e2895. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.2895

65. Madsen H, Thyregod P. Introduction to General and Generalized LInear Models. Journal of Applied Statistics - J APPL STAT. 2010.

66. Kadam P, Bhalerao S. Sample size calculation. Int J Ayurveda Res. 2010/06/10 ed. 2010 Jan;1(1):55–7.

67. Suresh K, Chandrashekara S. Sample size estimation and power analysis for clinical research studies. J Hum Reprod Sci. 2012/08/08 ed. 2012 Jan;5(1):7–13.

68. Cochran WG. Sampling Techniques, 3rd Edition. John Wiley; 1977.

69. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inf. 2014/03/04 ed. 2014 Apr;48:193–204.

70. Maxwell SE, Kelley K, Rausch JR. Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. Annu Rev Psychol [Internet]. 2008;59(1):537–63. Available from: https://doi.org/10.1146/annurev.psych.59.103006.093735

71. Cho J, Lee K, Shin E, Choy G, Do S. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy. ArXiv Learn. 2015;

72. Mukherjee S, Tamayo P, Rogers S, Rifkin R, Engle A, Campbell C, et al. Estimating Dataset Size Requirements for Classifying DNA Microarray Data. J Comput Biol [Internet]. 2003;10(2):119–42. Available from: https://doi.org/10.1089/106652703321825928

73. Balki I, Amirabadi A, Levman J, Martel AL, Emersic Z, Meden B, et al. Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review. Can Assoc Radiol J [Internet]. 2019 Nov 1 [cited 2020 Nov 11];70(4):344–53. Available from: https://doi.org/10.1016/j.carj.2019.06.002

74. Clauset A, Shalizi CR, Newman ME. Power-law distributions in empirical data. SIAM Rev. 2009;51(4):661–703.

75. Wood RH, Muhlbauer WC, Thompson PT. Systematic errors in free energy perturbation calculations due to a finite sample of configuration space: sample-size hysteresis. J Phys Chem. 1991;95(17):6670–5.

76. Hamann BC, Hartwig JF. Systematic variation of bidentate ligands used in aryl halide amination. Unexpected effects of steric, electronic, and geometric perturbations. J Am Chem Soc. 1998;120(15):3694–703.

77. Bocklitz T, Walter A, Hartmann K, Rösch P, Popp J. How to pre-process Raman spectra for reliable and stable models? Anal Chim Acta [Internet]. 2011 Oct 17;704(1):47–56. Available from: http://www.sciencedirect.com/science/article/pii/S0003267011008749

78. Afseth NK, Segtnan VH, Wold JP. Raman spectra of biological samples: A study of preprocessing methods. Appl Spectrosc. 2007/01/16 ed. 2006 Dec;60(12):1358–67.

79. Bocklitz TW, Dörfer T, Heinke R, Schmitt M, Popp J. Spectrometer calibration protocol for Raman spectra recorded with different excitation wavelengths. Spectrochim Acta A Mol Biomol Spectrosc [Internet]. 2015;149:544–9. Available from: http://europepmc.org/abstract/MED/25978023 https://doi.org/10.1016/j.saa.2015.04.079

80. Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal Chem [Internet]. 1964 Jul 1;36(8):1627–39. Available from: https://doi.org/10.1021/ac60214a047

81. Kuzmin AN, Pliss A, Prasad PN. Ramanomics: New Omics Disciplines Using Micro Raman Spectrometry with Biomolecular Component Analysis for Molecular Profiling of Biological Structures. Biosensors [Internet]. 2017;7(4):52. Available from: https://pubmed.ncbi.nlm.nih.gov/29140259
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5746775/

82. Ryabchykov O, Bocklitz T, Ramoji A, Neugebauer U, Foerster M, Kroegel C, et al. Automatization of spike correction in Raman spectra of biological samples. Chemom Intell Lab Syst [Internet]. 2016 Jul 15;155:1–6. Available from: http://www.sciencedirect.com/science/article/pii/S0169743916300600

83. Legesse FB, Chernavskaia O, Heuke S, Bocklitz T, Meyer T, Popp J, et al. Seamless stitching of tile scan microscope images. J Microsc. 2015/03/20 ed. 2015 Jun;258(3):223–32.

84. Acharya T, Ray AK. Image Processing - Principles and Applications. Wiley-Interscience; 2005.

85. Reza A. Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement. VLSI Signal Process. January 8;38:35–44.

86. Reeves SJ. Chapter 6 - Image Restoration: Fundamentals of Image Restoration. In: Trussell J, Srivastava A, Roy-Chowdhury AK, Srivastava A, Naylor PA, Chellappa R, et al., editors. Academic Press Library in Signal Processing [Internet]. Elsevier; 2014. p. 165–92. Available from: http://www.sciencedirect.com/science/article/pii/B9780123965011000066

87. Maini R, Aggarwal H. A Comprehensive Review of Image Enhancement Techniques. J Comput. 3;2.

88. Martin L, Marco E, Guido MS, Aggelos KK. Framework for efficient optimal multilevel image thresholding. J Electron Imaging [Internet]. January 1;18(1):1–10. Available from: https://doi.org/10.1117/1.3073891

89. Montáns FJ, Chinesta F, Gómez-Bombarelli R, Kutz JN. Data-driven modeling and learning in science and engineering. Comptes Rendus Mécanique [Internet]. 2019 Nov 1;347(11):845–55. Available from: http://www.sciencedirect.com/science/article/pii/S1631072119301809

90. Siddesh G. Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications. Springer Nature; 2020.

91. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York, USA: Springer Science & Business Media; 2009. (Springer Series in Statistics).

92. Kording K, Benjamin AS, Farhoodi R, Glaser J. The roles of machine learning in biomedical science. Bridge. 12;47:23–30.

93. Foster KR, Koprowski R, Skufca JD. Machine learning, medical diagnosis, and biomedical engineering research - commentary. Biomed Eng OnLine [Internet]. 2014 Jul 5;13(1):94. Available from: https://doi.org/10.1186/1475-925X-13-94

94. Gjoreski H, Bizjak J, Gjoreski M, Gams M. Comparing deep and classical machine learning methods for human activity recognition using wrist accelerometer. In 2016.

95. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science [Internet]. 2015;349(6245):255. Available from: http://science.sciencemag.org/content/349/6245/255.abstract

96. Abe S. Feature selection and extraction. In: Support vector machines for pattern classification. Springer; 2010. p. 331–41.

97. Guo S, Ryabchykov O, Ali N, Houhou R, Bocklitz T. Comprehensive Chemometrics. Oxford, United Kingdom: Elsevier; 2020. (Brown S Tauler, R, Walczak, B, editor. In Comprehensive Chemometrics: Chemical and Biochemical Data Analysis).

98. Agarwal B, Mittal N. Prominent feature extraction for review analysis: an empirical study. J Exp Theor Artif Intell. 2016;28(3):485–98.

99. Lorenz B, Ali N, Bocklitz T, Rösch P, Popp J. Discrimination between pathogenic and non-pathogenic E. coli strains by means of Raman microspectroscopy. Anal Bioanal Chem [Internet]. 2020 Oct 8; Available from: https://doi.org/10.1007/s00216-020-02957-2

100. Chernavskaia O, Heuke S, Vieth M, Friedrich O, Schürmann S, Atreya R, et al. Beyond endoscopic assessment in inflammatory bowel disease: real-time histology of disease activity by non-linear multimodal imaging. Sci Rep [Internet]. 2016 online;6:29239. Available from: https://doi.org/10.1038/srep29239

101. Heuke S, Chernavskaia O, Bocklitz T, Legesse FB, Meyer T, Akimov D, et al. Multimodal nonlinear microscopy of head and neck carcinoma - toward surgery assisting frozen section analysis. Head Neck [Internet]. 2016/04/22 ed. 2016 Oct;38(10):1545–52. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27098552

102. Kong K, Rowlands CJ, Varma S, Perkins W, Leach IH, Koloydenko AA, et al. Diagnosis of tumors during tissue-conserving surgery with integrated autofluorescence and Raman scattering microscopy. Proc Natl Acad Sci [Internet]. 2013;110(38):15189. Available from: http://www.pnas.org/content/110/38/15189.abstract

103. Wang S, Summers RM. Machine learning and radiology. Med Image Anal [Internet]. 2012 Jul 1;16(5):933–51. Available from: http://www.sciencedirect.com/science/article/pii/S1361841512000333

104. Goodacre R. Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules. Vib Spectrosc [Internet]. 2003 Aug 5;32(1):33–45. Available from: http://www.sciencedirect.com/science/article/pii/S0924203103000456

105. Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw [Internet]. 2015 Jan 1;61:85–117. Available from: http://www.sciencedirect.com/science/article/pii/S0893608014002135

106. Goodfellow I and B Yoshua and Courville, Aaron. Deep Learning. The MIT Press; 2015. (Adaptive computation and machine learning).

107. Kietzmann TC, McClure P, Kriegeskorte N. Deep Neural Networks in Computational Neuroscience. 2019 Jan 25; Available from: https://oxfordre.com/neuroscience/view/10.1093/acrefore/9780190264086.001.0001/acrefore-9780190264086-e-46

108. Silver D, Hasselt H, Hessel M, Schaul T, Guez A, Harley T, et al. The Predictron: End-To-End Learning and Planning. In: Doina P, Yee Whye T, editors. Proceedings of Machine Learning Research: PMLR; 2017. p. 3191--3199. Available from: http://proceedings.mlr.press

109. Chen CL, Mahjoubfar A, Tai L-C, Blaby IK, Huang A, Niazi KR, et al. Deep Learning in Label-free Cell Classification. Sci Rep [Internet]. 2016 online;6:21471. Available from: https://doi.org/10.1038/srep21471

110. Falk T, Mai D, Bensch R, Cicek O, Abdulkadir A, Marrakchi Y, et al. U-Net: deep learning for cell counting, detection, and morphometry. Nat Methods. 2018/12/19 ed. 2019 Jan;16(1):67–70.

111. Habibzadeh M, Jannesari M, Rezaei Z, Baharvand H, Totonchi M. Automatic white blood cell classification using pre-trained deep learning models: ResNet and Inception [Internet]. SPIE; 2018. (Tenth International Conference on Machine Vision; vol. 10696). Available from: https://doi.org/10.1117/12.2311282

112. Pradhan P, Meyer T, Vieth M, Stallmach A, Waldner M, Schmitt M, et al. Semantic Segmentation of Non-linear Multimodal Images for Disease Grading of Inflammatory Bowel Disease: A SegNet-based Application. 2019. 396–405 p.

113. Rodner E, Bocklitz T, von Eggeling F, Ernst G, Chernavskaia O, Popp J, et al. Fully convolutional networks in multimodal nonlinear microscopy images for automated detection of head and neck carcinoma: Pilot study. Head Neck [Internet]. 2019;41(1):116–21. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/hed.25489

114. Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans Pattern Anal Mach Intell [Internet]. 2017 Dec;39(12):2481–95. Available from: http://europepmc.org/abstract/MED/28060704 https://doi.org/10.1109/TPAMI.2016.2644615

115. Kumar A, Kim J, Lyndon D, Fulham M, Feng D. An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification. IEEE J Biomed Health Inform. 2017;21(1):31–40.

116. Aubreville M, Knipfer C, Oetter N, Jaremenko C, Rodner E, Denzler J, et al. Automatic Classification of Cancerous Tissue in Laserendomicroscopy Images of the Oral Cavity using Deep Learning. Sci Rep [Internet]. 2017 Sep 20;7(1):11979. Available from: https://doi.org/10.1038/s41598-017-12320-8

117. Srinidhi C, Ciga O, Martel A. Deep neural network models for computational histopathology: A survey. ArXiv. 2019;abs/1912.12378.

118. Shkolyar E, Jia X, Chang T, Trivedi D, Mach K, Meng M, et al. Augmented Bladder Tumor Detection Using Deep Learning. Eur Urol. January 9;76.

119. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Trans Med Imaging. 2016/02/18 ed. 2016 May;35(5):1285–98.

120. Wu E, Hadjiiski LM, Samala RK, Chan HP, Cha KH, Richter C, et al. Deep Learning Approach for Assessment of Bladder Cancer Treatment Response. Tomography. 2019/03/12 ed. 2019 Mar;5(1):201–8.

121. Xu H, Park S, Lee SH, Hwang TH. Using transfer learning on whole slide images to predict tumor mutational burden in bladder cancer patients. bioRxiv [Internet]. 2019;554527. Available from: https://www.biorxiv.org/content/biorxiv/early/2019/02/19/554527.full.pdf

122. Breck E, Polyzotis N, Roy S, Whang SE, Zinkevich M. Data Validation for Machine Learning. In 2019.

123. Huber L. Validation and Qualification in Analytical Laboratories. Boca Raton: CRC Press; 2007.

124. Guo S, Bocklitz T, Neugebauer U, Popp J. Common mistakes in cross-validating classification models. Anal Methods [Internet]. 2017;9(30):4410–7. Available from: http://dx.doi.org/10.1039/C7AY01363A

125. Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. Sample size planning for classification models. Anal Chim Acta [Internet]. 2013;760:25–33. Available from: http://www.sciencedirect.com/science/article/pii/S0003267012016479

126. Kelley CT. Iterative methods for optimization. SIAM; 1999.

127. Nieuwenhuis R, Grotenhuis M te, Pelzer B. Weighted Effect Coding for Observational Data with wec. R J [Internet]. 2017;9(1):477–85. Available from: https://doi.org/10.32614/RJ-2017-017

128. te Grotenhuis M, Pelzer B, Eisinga R, Nieuwenhuis R, Schmidt-Catran A, Konig R. A novel method for modelling interaction between categorical variables. Int J Public Health [Internet]. 2017 Apr;62(3):427–31. Available from: https://doi.org/10.1007/s00038-016-0902-0

129. Hutcheson GD. Ordinary least-squares regression. Moutinho GD Hutcheson SAGE Dict Quant Manag Res. 2011;224–8.

130. de Souza SV, Junqueira RG. A procedure to assess linearity by ordinary least squares method. Anal Chim Acta. 2005;552(1–2):25–35.

131. Vogler N, Bocklitz T, Subhi Salah F, Schmidt C, Bräuer R, Cui T, et al. Systematic evaluation of the biological variance within the Raman based colorectal tissue diagnostics. J Biophotonics [Internet]. 2016;9(5):533–41. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jbio.201500237

132. Antibiotic resistance, World Health Organization (WOH) [Internet]. https://www.who.int/topics/antimicrobial_resistance/en/. 2020. Available from: https://www.who.int/topics/antimicrobial_resistance/en/

133. Ventola CL. The antibiotic resistance crisis: part 1: causes and threats. P T Peer-Rev J Formul Manag [Internet]. 2015;40(4):277–83. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25859123 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4378521/

134. Giuliano C, Patel CR, Kale-Pradhan PB. A Guide to Bacterial Culture Identification And Results Interpretation. P T Peer-Rev J Formul Manag [Internet]. 2019;44(4):192–200. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30930604 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6428495/

135. Pulido MR, Garcia-Quintanilla M, Martin-Pena R, Cisneros JM, McConnell MJ. Progress on the development of rapid methods for antimicrobial susceptibility testing. J Antimicrob Chemother. 2013/07/03 ed. 2013 Dec;68(12):2710–7.

136. van Belkum A, Welker M, Pincus D, Charrier JP, Girard V. Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry in Clinical Microbiology: What Are the Current Issues? Ann Lab Med [Internet]. 2017;37(6):475–83. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28840984 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5587819/

137. Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, et al. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report

from the EUCAST Subcommittee. Clin Microbiol Infect [Internet]. 2017 Jan 1;23(1):2–22. Available from: http://www.sciencedirect.com/science/article/pii/S1198743X16305687

138. Antoni S, Ferlay J, Soerjomataram I, Znaor A, Jemal A, Bray F. Bladder Cancer Incidence and Mortality: A Global Overview and Recent Trends. Eur Urol. 2016/07/03 ed. 2017 Jan;71(1):96–108.

139. Burger M, Catto JW, Dalbagni G, Grossman HB, Herr H, Karakiewicz P, et al. Epidemiology and risk factors of urothelial bladder cancer. Eur Urol. 2012/08/11 ed. 2013 Feb;63(2):234–41.

140. Cina SJ, Epstein JI, Endrizzi JM, Harmon WJ, Seay TM, Schoenberg MP. Correlation of cystoscopic impression with histologic diagnosis of biopsy specimens of the bladder. Hum Pathol. 2001/06/30 ed. 2001 Jun;32(6):630–7.

141. Burger M, Grossman HB, Droller M, Schmidbauer J, Hermann G, Dragoescu O, et al. Photodynamic diagnosis of non-muscle-invasive bladder cancer with hexaminolevulinate cystoscopy: a meta-analysis of detection and recurrence based on raw data. Eur Urol. 2013/04/23 ed. 2013 Nov;64(5):846–54.

142. Rink M, Babjuk M, Catto JW, Jichlinski P, Shariat SF, Stenzl A, et al. Hexyl aminolevulinate-guided fluorescence cystoscopy in the diagnosis and follow-up of patients with non-muscle-invasive bladder cancer: a critical review of the current literature. Eur Urol. 2013/08/03 ed. 2013 Oct;64(4):624–38.

143. Daneshmand S, Bazargani ST, Bivalacqua TJ, Holzbeierlein JM, Willard B, Taylor JM, et al. Blue light cystoscopy for the diagnosis of bladder cancer: Results from the US prospective multicenter registry. Urol Oncol. 2018/06/04 ed. 2018 Aug;36(8):361.e1-361.e6.

144. Mari A, Abufaraj M, Gust KM, Shariat SF. Novel endoscopic visualization techniques for bladder cancer detection: a review of the contemporary literature. Curr Opin Urol. 2017/10/19 ed. 2018 Mar;28(2):214–8.

145. Gravas S, Efstathiou K, Zachos I, Melekos MD, Tzortzis V. Is there a learning curve for photodynamic diagnosis of bladder cancer with hexaminolevulinate hydrochloride? Can J Urol. 2012/06/19 ed. 2012 Jun;19(3):6269–73.

146. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. In 2016. p. 2818–26.

147. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In 2018. p. 4510–20.

148. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In 2016. p. 770–8.

149. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv 14091556. April 9;

150. Deng J, Dong W, Socher R, Li L, Kai L, Li F-F. ImageNet: A large-scale hierarchical image database. In 2009. p. 248–55.

151. Ataollahi MR, Sharifi J, Paknahad MR, Paknahad A. Breast cancer and associated factors: a review. J Med Life [Internet]. 2015/01/01 ed. 2015;8(Spec Iss 4):6–11. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28316699

152. Milosevic M, Jankovic D, Milenkovic A, Stojanov D. Early diagnosis and detection of breast cancer. Technol Health Care. 2018/08/21 ed. 2018;26(4):729–59.

153. World Health Organization (WHO). Breast Cancer [Internet]. [cited 2019 Jul 30]. Available from: https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

154. Migowski A. [Early detection of breast cancer and the interpretation of results of survival studies]. Cien Saude Colet [Internet]. 2015/04/30 ed. 2015 Apr;20(4):1309. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25923642

155. Hellquist BN, Czene K, Hjalm A, Nystrom L, Jonsson H. Effectiveness of population-based service screening with mammography for women ages 40 to 49 years with a high or low risk of breast cancer: socioeconomic status, parity, and age at birth of first child. Cancer [Internet]. 2014/09/23 ed. 2015 Jan 15;121(2):251–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25242087

156. Lee JM, Halpern EF, Rafferty EA, Gazelle GS. Evaluating the correlation between film mammography and MRI for screening women with increased breast cancer risk. Acad Radiol [Internet]. 2009/07/28 ed. 2009 Nov;16(11):1323–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19632865

157. Onega T, Goldman LE, Walker RL, Miglioretti DL, Buist DS, Taplin S, et al. Facility Mammography Volume in Relation to Breast Cancer Screening Outcomes. J Med Screen [Internet]. 2015/08/13 ed. 2016 Mar;23(1):31–7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26265482

List of Publications

I. Sample-Size Planning for Multivariate Data: A Raman-Spectroscopy-Based Example. Analytical chemistry

Reprinted with permission from [N. Ali, S. Girnus, P. Rösch, J. Popp, and T. Bocklitz: Sample-Size Planning for Multivariate Data: A Raman-Spectroscopy-Based Example. Analytical chemistry, 2018, 90 (21), 12485-12492] Copyright © 2018, American Chemical Society

Erklärungen zu den Eigenanteilen des Promovenden sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation.

N. Ali, S. Girnus, P. Rösch, J. Popp, and T. Bocklitz, *Sample-Size Planning for Multivariate Data: A Raman-Spectroscopy-Based Example*, Analytical Chemistry, 2018, 90 (21), 12485-12492.					
Beteiligt an (Zutreffendes ankreuzen)					
	1	2	3	4	5
Konzeption des Forschungsansatzes	×			×	×
Planung der Untersuchungen	×		×	×	×
Datenerhebung		×	×		
Datenanalyse und -interpretation	×				×
Schreiben des Manuskripts	×		×	×	×
Vorschlag Anrechnung Publikationsäquivalente	1.0				

analytical chemistry
Article
Cite This: Anal. Chem. 2018, 90, 12485−12492
pubs.acs.org/ac

Sample-Size Planning for Multivariate Data: A Raman-Spectroscopy-Based Example

Nairveen Ali,[†,‡] Sophie Girnus,[†] Petra Rösch,[†] Jürgen Popp,[†,‡,§,⊥] and Thomas Bocklitz*,[†,‡]

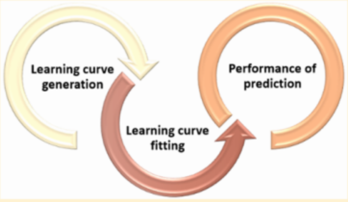
[†]Institute of Physical Chemistry and Abbe Center of Photonics (IPC), Friedrich-Schiller-University, Helmholtzweg 4, D-07743 Jena, Germany

[‡]Leibniz Institute of Photonic Technology (IPHT), Albert-Einstein-Straße 9, D-07745 Jena, Germany

[§]Center for Sepsis Control and Care (CSCC), Jena University Hospital, Erlanger Allee 101, D-07747 Jena, Germany

[⊥]InfectoGnostics, Forschungscampus Jena, Philosophenweg 7, D-07743 Jena, Germany

Supporting Information



ABSTRACT: The goal of sample-size planning (SSP) is to determine the number of measurements needed for statistical analysis. This SSP is necessary to achieve robust and significant results with a minimal number of measurements that need to be collected. SSP is a common procedure for univariate measurements, whereas for multivariate measurements, like spectra or time traces, no general sample-size-planning method exists. Sample-size planning becomes more important for biospectroscopic data because the data generation is time-consuming and costly. Additionally, ethical reasons do not allow the use of unnecessary samples and the measurement of unnecessary spectra. In this paper, a general sample-size-planning algorithm is presented that is based on learning curves. The learning curve quantifies the improvement of a classifier for an increasing training-set size. These curves are fitted by the inverse-power law, and the parameters of this fit can be utilized to predict the necessary training-set size. Sample-size planning is demonstrated for a biospectroscopic task of differentiating six different bacterial species, including *Escherichia coli*, *Klebsiella terrigena*, *Pseudomonas stutzeri*, *Listeria innocua*, *Staphylococcus warneri*, and *Staphylococcus cohnii*, on the basis of their Raman spectra. Thereby, we estimate the required number of Raman spectra and biological replicates to train a classification model, which consists of principal-component analysis (PCA) combined with linear-discriminant analysis (LDA). The presented algorithm revealed that 142 Raman spectra per species and seven biological replicates are needed for the above-mentioned biospectroscopic-classification task. Even though it was not demonstrated, the learning-curve-based sample-size-planning algorithm can be applied to any multivariate data and in particular to biospectroscopic-classification tasks.

In almost all disciplines of science, statistics is utilized to determine if a significant difference between different groups exists. This determination is achieved by collecting measurements from each group, and afterward statistical methods are applied. To reach reliable and robust results, a suitable number of measurements needs to be collected. The number of necessary measurements for a certain significance level is usually known as a sample size, and this sample size is determined on the basis of already existing knowledge about the samples or the experiment.[1,2] This information might be deduced from pre-experiments or existing literature values. The sample-size estimation becomes a critical issue for biomedical studies, because high costs and ethical reasons restrict the collection of unnecessary data. An example for ethical reasons is that in a biomedical study involving patient material, only a defined sample volume can be used for spectral measurements. In this case the number of spectral measurements is restricted by the approval of the ethics committee. Therefore, sample-size planning (SSP) is necessary to decide how many measurements are needed minimally to differentiate between various groups in a significant manner. In the last few decades, classical statistics was applied for sample-size estimation for univariate measurements.[3,4]

Sample-size planning for univariate data uses hypothesis testing to predict a suitable sample size.[4−7] This procedure is usually called sample-size planning or power analysis. The basic idea is to test whether the groups are significantly different on the basis of the respective group means, and the minimal number of measurements necessary to prove this difference of means is calculated. A drawback of power analysis is that previous knowledge of the statistical distribution of the measurement values is required. The most prominent example is the test of mean difference of two groups that are Gaussian distributed. Maxwell et al.[6] introduced a formula for sample-size planning based on the standardized difference of means (or effect size). The effect size is a robust measure for quantifying the significance of the mean difference. In this case,

Received: May 15, 2018
Accepted: October 1, 2018
Published: October 1, 2018

Downloaded via THURINGER UNIV LANDESBIBLIOTHEK on October 7, 2020 at 07:51:23 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ACS Publications © 2018 American Chemical Society

DOI: 10.1021/acs.analchem.8b02167
Anal. Chem. 2018, 90, 12485−12492

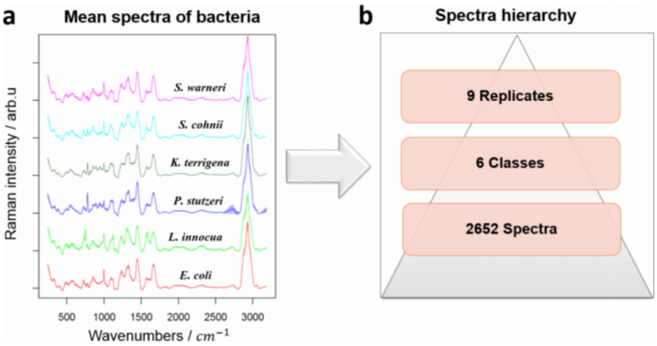


Figure 1. Raman-spectral data set. (a) Mean Raman spectra of bacteria. The preprocessed mean Raman spectra of six bacterial species were calculated on the basis of the normalized and preprocessed Raman spectra of the bacterial species. (b) Hierarchy of the Raman-spectral data set. The data set consists of six bacterial species: *E. coli*, *K. terrigena*, *P. stutzeri*, *L. innocua*, *S. warneri*, and *S. cohnii*. These species were cultivated in nine independent biological replicates. In total, 2652 Raman spectra were measured. The biological replicates represent the independent measurements in the Raman-spectral data set.

the sample size ($n_{(\alpha,\beta)}$) is defined for the significance level, α, and the power of the test, $1 - \beta$, as the following:

$$n_{(\alpha,\beta)} \geq \frac{2(Z_{1-\beta} + Z_{1-\alpha/2})^2}{d^2}$$

where $d \neq 0$ represents the effect size, and $Z_{1-\beta}$ and $Z_{1-\alpha/2}$ are the quantiles of the standard normal distribution.[3] This formula provides a good approximation of the number of measurements required to differentiate two groups in the case of univariate data. Nevertheless, for multivariate data, this formula is not applicable, as it is not clear how a univariate standardized difference can be calculated. This mentioned issue is in stark contrast with the intensive use of spectroscopic, spectrometric, and other multivariate measurements for the characterization and measurement of biomedical samples, which have been commonly performed in the last few decades. To mention a few, Raman spectroscopy has been used to identify microorganisms,[8−11] predict cell types,[12−15] and detect cancer areas.[16−18] Besides Raman spectroscopy, other multivariate measurement techniques have been used to characterize biomedical samples. Among the techniques that are commonly applied, mass spectrometry,[19] nuclear-magnetic-resonance (NMR) spectroscopy,[20] and IR spectroscopy[21] are the most utilized measurement techniques. Therefore, a different SSP algorithm is needed that can deal with multivariate data sets. There are only a few publications that work on SSP algorithms for multivariate data.[22−24] In all of them, the learning curve (LC) is one of the cornerstones. The learning curve quantifies the classification performance with respect to the training-set size and characterizes the learning behavior of a classifier. Mukherjee et al.[22] utilized learning curves to estimate the data-set size required to classify microarray data sets. In their paper, the inverse-power law was implemented to fit the learning curves. Their procedure was demonstrated and tested for several DNA-microarray data sets and binary-classification tasks. A simulation-based study regarding SSP was performed by Beleites et al.[24] The necessary number of Raman spectra was estimated for biomedical (multiclass)-classification tasks on the basis of resampling.

In this paper, we introduce a general SSP algorithm for multivariate data and in particular for biospectroscopic data. This algorithm is constructed using mathematical methods that are combined together to predict the minimum sample size needed for successful group differentiation. Thereafter, the classifier performance is extrapolated on the basis of this predicted sample size. Although the methodology of the presented algorithm is the same for all multivariate data, different factors could influence the final sample-size prediction. Examples of these influencing parameters are the utilized experimental protocol or the selected parameters for the data preprocessing. In biomedical studies, the data sets additionally feature a hierarchical structure, which originates if multiple measurements on samples from the same biological replicate (e.g., samples from the same patient or cultivation batch) are performed. This data structure leads to the fact that spectral measurements within these replicates show a strong connection that violates the statistical independence between the measurements. Our algorithm incorporates this fact by performing the statistical analysis on the highest level of the data hierarchy. In this case, the statistical model is constructed on the basis of a number of training replicates that are different than the test replicates. This additionally ensures statistical independence between the training and validation sets in the applied cross-validation (CV).[25−27] Therefore, we implemented a version of CV called leave-one-replicate-out cross-validation (LORO-CV) for all classification tasks. This method of CV represents a robust alternative to classical CV, where the validation fold (or validation measurement) in classical CV is replaced with all the measurements of a replicate. Nevertheless, our SSP algorithm was designed on the basis of learning curves to estimate the required number of measurements for different hierarchical levels of the biological data (number of biological replicates and number of single spectral measurements), whereas the statistical independence was ensured by applying LORO-CV. The test of this algorithm was carried out on the basis of a biological Raman-spectral data set consisting of six classes and nine biological replicates. This data set includes classification problems of different difficulties because of the similar and diverse bacterial species.

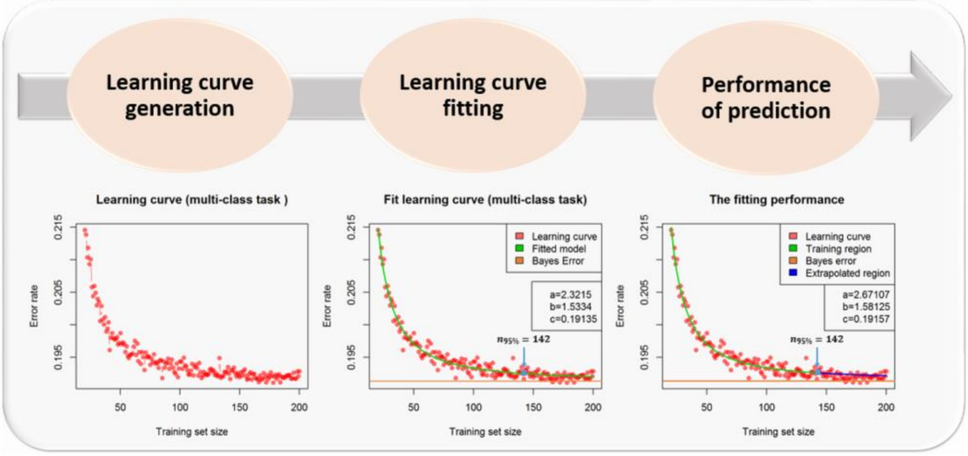


Figure 2. Visualization of the SSP algorithm for the prediction of the required size of the training set of spectra. First, the LC is generated by constructing a classifier on the basis of a training set of size n, where the maximum training-set size is $N = 200$. See the text for details on LC generation. The obtained error rates are represented by red points, and the LC is fitted by the inverse-power law. This fit includes calculating and optimizing the inverse-power-law parameters. These parameters are the learning rate, a; the decay rate, b; and the Bayes error, c. The last parameter represents the final performance of a classifier, which is trained with an infinite training-set size. The model fitted by the inverse-power law is plotted as a green line. Finally, the required training set is calculated on the basis of 95% of the Bayes error ($n_{95\%}$). Then, we fitted an inverse-power law until $n_{95\%}$ and checked the extrapolation performance by the root-mean-square error (RMSE) of the predicted area (blue line).

■ MATERIAL AND METHODS

Raman-Spectra Data Set. In order to examine and evaluate our sample-size-planning algorithm, Raman spectra of bacterial species were utilized. This Raman-spectral data set consisted of six bacterial species, *Escherichia coli* DSM 423, *Klebsiella terrigena* DSM 2687, *Pseudomonas stutzeri* DSM 5190, *Listeria innocua* DSM 20649, *Staphylococcus warneri* DSM 20316, and *Staphylococcus cohnii* DSM 20261, from Deutsche Sammlung von Mikroorganismen and Zellkulturen GmbH (DSMZ). All species were cultivated in nine independent biological replicates. The species were cultivated using Nutrition-Bouillon at 37 °C for 24 h. After cultivation, the species were diluted with distilled water, washed three times by centrifugation for 1 min at 10 844*g* (Hettich Rotina 380 R), and finally suspended in distilled water. The samples were immediately prepared for the Raman measurements by placing 1 μL of the suspension on nickel foil and air drying at room temperature. For the measurements, a Raman microscope (Bio Particle Explorer, rap.ID Particle Systems GmbH) using a 532 nm solid-state frequency-doubled Nd:YAG laser (LCM-s-111-NNP25, Laser-export Company Ltd.) was used. The laser beam was focused through a 100× objective (MPLFLN-BD, Olympus) on the nickel foil with a lateral resolution below 1 μm, and the maximum laser power after the objective was approximately 3 mW. A single-stage monochromator (HE532, Horiba Jobin Yvon, grating of 920 lines) dispersed the backscattered Raman light before its detection with a thermoelectrically cooled CCD (DV 401_BV, Andor Technology) resulting in a spectral resolution of approximately 8 cm^{-1}. In Figure 1a, the preprocessed mean spectra of the bacterial species are shown, and the data hierarchy can be found in Figure 1b. Within the considered Raman-spectral data set, different classification difficulties were present, such as the differentiation between *S. warneri* and *S. cohnii* species, which belong to same genus (*Staphylococcus*), and the classification of *E. coli* and *L. innocua*, which have similar genera. These different difficulties together with the methods of data preprocessing have an influence on final outcome of the classification model (e.g., the classification accuracy).

Data Preprocessing. An essential part of the Raman-spectral-data analysis is the spectral preprocessing. It aims to reduce unwanted variations between different Raman spectra and enhance the differences between different species. For the Raman-spectral data set of bacteria, a common preprocessing workflow was implemented. The bacterial spectra were uploaded into the statistical programming language R.[28] A median filter was applied to remove cosmic spikes within the data. Then, a wavenumber calibration was utilized to correct the peak position.[29] All Raman spectra were aligned between 240−3190 cm^{-1}. To exclude the background effects from the raw spectra, we applied the iterative restricted least-squares (IRLS) algorithm for baseline correction.[30] The last two steps of spectral preprocessing were spectral smoothing and normalization. A Savitzky−Golay filter was utilized to smooth the spectra.[31] The mean spectra after preprocessing can be seen in Figure 1a. We can note that the differences between the Raman spectra of the bacterial species are very tiny. In particular, closely related bacteria show almost identical mean Raman spectra. Therefore, the Raman spectra need to be analyzed using advanced statistical methods. Here, we implemented a simple classification model which suited the high dimensionality of Raman spectra and did not have many parameters to be optimized. This classification model is described in the following sections.

Classification Model and Computations. In this paper, a combination of principal-component analysis (PCA) and linear-discriminant analysis (LDA) was used for classification.[29] PCA is an unsupervised method aiming to reduce the high dimensionality of the multivariate data by projecting the data into a lower-dimensional subspace, whereas LDA is a parametric classification method that can be easily applied. The combination of PCA and LDA (PCA−LDA model) allows choosing one free parameter (e.g., the number of principal components, PCs). In our paper, we did the sample-size planning for multiclass-classification tasks. Therefore, we utilized the mean sensitivity to choose the optimal number of PCs and predict the training-sample size. For validation, leave-one-replicate-out cross-validation (LORO-CV) was applied.[25] This validation method guarantees the independence of training and validation data within the evaluation of the classification model. It can be explained easily by excluding one replicate as a validation set, and the classification model is constructed on the basis of the remaining replicates. Then, this procedure is repeated for all replicates as a validation set. All computational steps were carried out on the basis of in-house written functions in RStudio version 3.4.2, and the R packages baseline,[32] mdatool,[33] oce,[34] caret,[35] and MASS[36,37] were utilized. These functions are available upon request.

■ RESULTS AND DISCUSSION

In this section, we introduce our SSP algorithm and discuss the obtained results for the Raman-spectral data set of bacteria. Within biospectral data sets the number of independent measurements is typically very small and represents the highest level of the data hierarchy (Figure 1b). In our case the independent measurements are the biological replicates or cultivated batches of bacteria. In all the previous publications, including our own,[22−24] SSP was done to estimate the required number of single measurements, such as spectra or genes. In this publication we present an SSP algorithm to estimate the required number of independent measurements (biological replicates) and single measurements (spectra or genes). The proposed algorithm utilizes the learning curve to predict and evaluate the performance of a classifier (Figure 2). The SSP algorithm for the estimation of training-set size for both cases (biological replicates and Raman spectra) works in three steps: a learning curve (LC) is generated, the inverse-power law is fitted to this curve, and the sample size is calculated on the basis of this fit. These steps represent the mathematical pipeline of our algorithm, which can be applied independently of data source to predict the required training-set size for all multivariate data. Our SSP-algorithm workflow is shown in detail in Figure 2 and described in the following section in detail.

Sample-Size-Planning (SSP) Algorithm. The main part of our SSP algorithm is the generation of the learning curve (LC). Thereby, the LC describes the classification performance by quantifying how a classifier learns when the training-set size increases. Here, the LC is implemented to represent the empirical classification-error rate for different training-set sizes, which can be accomplished by calculating the errors for increasing the sequences of training-set sizes by progressive sampling. In progressive sampling, one determines a maximum size of the training set, N, and then the classifier is built for all possible training sets whose sizes are smaller than the upper limit, N. The obtained classifiers based on each training set are tested on an independent data set. The independence between the test set and training set is very important for retrieving reliable and accurate classification results. In this paper we implemented LORO-CV to ensure the independence between the test and training sets. The classification quality is quantified by the classification mean sensitivity, μ_{sen}, in the multiclass problems, whereas the sensitivity for a specific class represents the true positive rate for that class.

In detail, the algorithm works as follows: First the number of PCs is determined, which can be performed in multiple ways.[29,38,39] Nevertheless, internal cross-validation is advisable (Figure S1a). To generate a reliable LC, we applied the following algorithm, which ensures the independence of the training and testing data sets. We started by selecting a specific replicate as a test set. Then, we sampled n spectra (or replicates) from each of the six classes from the other replicates and trained the classification model on the basis of that sampling. We tested the model on the test data and calculated the classification error on the basis of the mean sensitivity. This sampling procedure was repeated 500 times, and the mean LC for a given replicate as the test set was constructed. In Figure S1c, the mean LC for a given replicate as a test set can be seen. The classification error for a training sample of size n is defined as

$$E(n) = 1 - \mu_{\text{sen}}(n)$$

The final estimate of the error rate (ER) is defined by the median of the classification errors, $E(n)$. By doing so, an average LC is generated by calculating the median of all $E(n)$ for a given n. After generating the LC, we fit the LC by the inverse-power law,[22] which is defined as

$$\text{IP}(n) = a \times n^{-b} + c$$

In this formula, a, b, and c are the inverse-power-law parameters. Parameter a refers to the learning rate, whereas b is the decay rate. Parameter c represents the final performance of a classifier, if it is trained with an infinite training set. Parameter c is also known as Bayes error. The fit of the above-defined model is done by the nonlinear least-squares algorithm. The model is approximating the inverse relation between the classification-error rate, ER(n), and the training-set size, n. This relation passes through three phases: First, a small difference in training size improves the classification performance strongly. Second, larger training sizes decrease the classification-error rate, but the classification improvement is smaller compared with that of the first phase. In the last phase, the increase in training-set size is not significant anymore, and the LC reaches its asymptotic behavior. The fitted IP model and its parameters can be used to investigate the performance of a model, so we can check the Bayes error and learning-related parameters (e.g., learning rate and decay rate). Both of these parameters are influenced by the standardization of the experiment via a standard operating procedure (SOP), the data preprocessing, the chosen model, and the difficulty of the investigated classification task itself. In parameters a and b, different effects are mixed, so we focus on the interpretation and utilization of the Bayes error, c. In the framework of SSP, we need to predict the required training-set size for a given performance and check the extrapolation quality of the inverse-power-law model. To do so, we estimate the training-set size of $n_{95\%}$, which gives 95% of Bayes error.[40] This 95% of Bayes error shows a 5% error range of the final performance of the PCA−LDA model, which is given by parameter c. After predicting training-set size of $n_{95\%}$, we fit an LC with this maximal training-set size of $n_{95\%}$. The model obtained by the

DOI: 10.1021/acs.analchem.8b02167
Anal. Chem. 2018, 90, 12485−12492

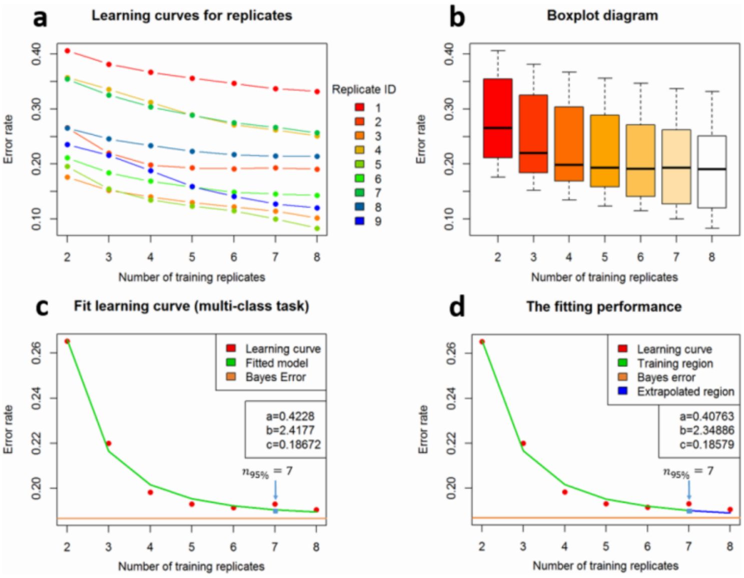


Figure 3. SSP results for biological replicates. (a) LCs for different test replicates. All the test replicates show inverse relations between the classification-error rate and the number of training replicates. (b) Boxplot for the different numbers of replicates as training sets. Increasing the number of training replicates improves the classification results. This improvement decreases for training-set sizes larger than four biological replicates and becomes nonsignificant for training-set sizes larger than six replicates. (c) LC fit. The fitted model suits the LCs quite well and converges fast to the final performance of PCA−LDA, which is represented by parameter *c*. (d) Model extrapolated on the basis of the prediction of replicate-training-set size. The predicted number of training replicates is $n_{95\%} = 7$. This number represents the required number of training replicates to reach 95% of Bayes error. The fitted model based on seven replicates is very good, and the extrapolated region represents the LC points nicely.

inverse-power law is implemented again to validate the extrapolation performance. The evaluation of the fit within the extrapolation region (e.g., sample size larger than $n_{95\%}$) is achieved by comparing the root-mean-square errors (RMSEs) in the extrapolation region and the training region (e.g., sample size smaller than $n_{95\%}$). This RMSE is given by

$$\mathrm{RMSE} = \sqrt{\frac{\sum[\mathrm{ER}(n) - \mathrm{IP}(n)]^2}{m}}$$

where m is the number of points in the region of interest, and $\mathrm{ER}(n)$ is the observed classification-error rate in the respective region. In Figure 2, the red points represent the LC, whereas the green and blue lines represent the model fitted by the inverse-power law and the extrapolated inverse-power law on the basis of the estimated training-set size of $n_{95\%}$, respectively.

Sample-Size Planning for Biological Replicates. In order to study the influence of the utilized PCs on the training-set-size prediction, we checked the SSP algorithm for three ranges of PCs. Figure S1 illustrates the results of our algorithm for 10, 20, and 30 PCs as maximal included PCs. Every column of the plot represents 10, 20, and 30 PCs, respectively. Figure S1a shows the classification results based on LORO-CV. Here, we constructed and evaluated a classification model utilizing a defined number of PCs. As an evaluation merit we utilized the mean sensitivity over all species. On the basis of the maximal sensitivity, the PC number was determined, which was utilized in further modeling. This method of PC determination represents a fast and simple technique to select the number of PCs, which is the only free parameter of our classification model. Importantly this method secures the independence between training and validation sets by utilizing LORO-CV. Figure S1b shows the mean sensitivity for each replicate as test set, and if these learning curves are averaged, Figure S1a results. In Figure S1c, the generated LC by the PCA−LDA model, was plotted together with the standard deviation. Figure S1d shows the fit of the respective learning curves with the inverse-power law. The prediction of training-set size can be found in the last row of Figure S1. For 10 PCs, the maximum mean sensitivity that can be reached by the PCA−LDA model is 75.53%. In this case, the LC does not show asymptotic behavior within the given number of replicates. This can be observed in the fitted coefficients of the inverse-power law. The coefficient *b*, which represents the learning, is small. Therefore, the required number of replicates is very large. Also, the Bayes error is small for this low-dimensional model. Both of these facts indicate that the dimensionality of this model is too low. Increasing the utilized number of PCs up to 20, the highest mean sensitivity of PCA−LDA for the six bacterial species is 81.4%. Here, the LC reaches its asymptotic behavior, and the fitted inverse-power law reflects that. The predicted number of replicates is seven biological replicates which represents the training-size $n_{95\%}$. If the number of PCs is increased to 30 PCs, the maximum mean sensitivity is 81.89%,

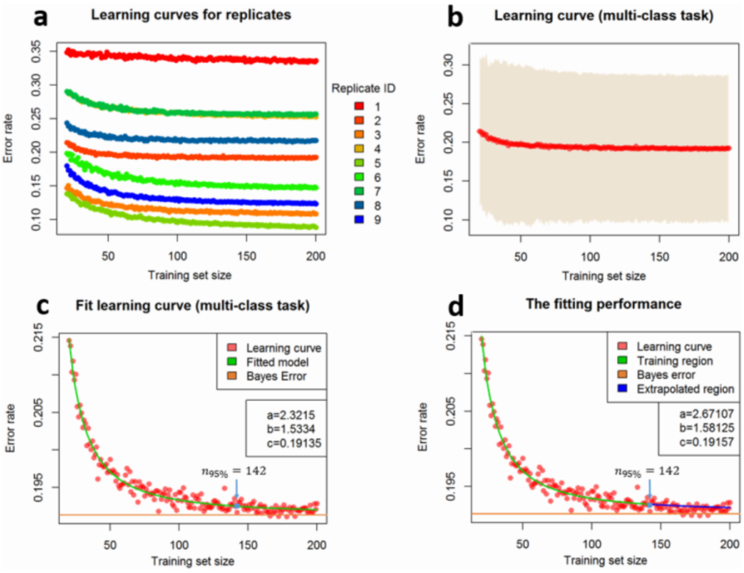


Figure 4. SSP algorithm for spectra. (a) LCs for different replicates as a test set. Here, $N = 200$ spectra per class were set as the maximal spectra number. All LCs achieve their asymptotic behavior in the region smaller than N. (b) Average LC for the Raman-spectral data set of bacteria. This LC reaches its stable behavior in the presented region. (c) Fitted inverse-power-law model. The fitted model suits the LC quite well, and it converges fast to the final performance of classification (orange line). (d) Training and extrapolated regions based on the predicted training-set size. Our algorithm implies that at least $n_{95\%} = 142$ spectra per class are required to build a reliable PCA–LDA model. This size was used to extrapolate the inverse-power law. The evaluation of the training region and the extrapolated region leads to the assumption that the fitted model based on $n_{95\%} = 142$ suits the LC very well.

which is given for 30 PCs. On the basis of 30 PCs, the generated LC is fitted well by the inverse-power law. It turns out that 24 replicates are necessary for the PCA–LDA model to reach 95% of its final performance. The error rate by using 24 replicates as a training set is around 12.61%.

To summarize the previous results, we compare the SSP results for different ranges of PCs. If only 10 PCs are utilized, the construction of a reliable model is not possible, because the dimension is too small. That is why the learning is not finished with 10 PCs. If the individual LCs for 20 PCs are compared, some replicates with stationary behavior can be seen, whereas a decrease in classification performance for some replicates can be already seen if up to 30 PCs are incorporated. This might originate from overfitting for these replicates. Also, the improvement in classification results is not significant if we increase the number of PCs up to 30. Therefore, the use of 20 PCs was considered to construct the PCA–LDA model for both SSP tasks (biological replicates and spectra). However, in the following, the proposed SSP algorithm is implemented with some adjustment to fit SSP for biological replicates. This adjustment involves two parameters: The first parameter is the maximum training size, N, which is considered as $N = 8$ replicates. The reason behind that is that the bacterial data set consists of nine biological replicates, and one of these replicates should be the test set. The second adjusted parameter is the maximum number of iterations. This number is defined as the maximum number of combinations between the training replicates, whereas the other parameters in the SSP algorithm are fixed as in the SSP description. The results of SSP for the replicates are presented in Figure 3. The generated LCs for different test replicates can be found in Figure 3a. In most LCs, increasing the training-set size up to six replicates improves the classification results, whereas this improvement becomes smaller for seven and eight training replicates. In Figure 3b, the boxplot of the previous LCs is shown. It is clear that increasing the training size to more than six replicates does not improve the classification results. In Figure 3c, the generated LC and the inverse-power-law model can be seen. In this figure, the averaged LC is represented by red points, and the fitted inverse-power law is represented by the green line.

It is observed that the fitted model suits the LC quite well, and it converges very fast to Bayes error rate, c. As mentioned earlier, Bayes error represents the final classification performance for an infinite replicate-training-set size, and in our algorithm, Bayes error is employed for the prediction of training-set size. This implementation leads to the assumption that seven replicates are sufficient to construct a PCA–LDA model. The error rate by using seven replicates is almost 19%, which is 95% of the Bayes error rate. In Figure 3d, the predicted training-set size is utilized to extrapolate the performance of the PCA–LDA model. Here, the training size $n_{95\%} = 7$ replicates was considered as a maximum size to fit the LC with the inverse-power law. The fitted models based on $n_{95\%}$ and the extrapolated inverse-power law are represented by green and blue lines, respectively. These two regions were utilized to evaluate the extrapolation performance. This evaluation was done on the basis of the RMSE for the training and extrapolation regions. For the training region, the observed RMSE is 0.2432%, which represents the RMSE of the fitted model based on the training size $n_{95\%} = 7$ replicates. For the extrapolation region, the observed RMSE is 0.2453%. Therefore, using seven replicates as a training set is enough to fit the LC and to extrapolate the performance of the inverse-power law well. Moreover, this number of replicates ($n_{95\%} = 7$) is the required training size to achieve 95% of the final performance of a PCA−LDA model.

Sample-Size Planning for Raman Spectra. In the following section, we present the results of our SSP algorithm for the prediction of required spectrum-training-set size. In this case, we used the same parameters of SSP for the replicates. However, the PCA−LDA model was built on the basis of 20 PCs, and it was trained by training sets of the size $n \leq 200$ spectra per class and afterward tested by an independent test replicate. Finally, this procedure was iterated 500 times and repeated for all replicates as a test set. After that, the calculated error rate for each training-set size of $n \leq 200$ were plotted in Figure 4a. This figure shows the generated LC for each test replicate. All LCs achieve their asymptotic behavior, which can be observed as a decrease in the classification improvement for training-set sizes of $n > 100$. Moving to Figure 4b, the averaged LC is generated by the median of the individual LCs, whereas the standard deviation of this LC is represented by the beige area. In Figure 4c, the inverse-power law is implemented to fit the averaged LC (Figure 4b). Here, the obtained fitted model suits the LC very well, and it converges fast to Bayes error rate. This Bayes error rate is used to predict the training-set size of $n_{95\%}$. In our case, 142 spectra per class are needed to train and construct the PCA−LDA model, and the corresponding error rate is 19.25%. Coming up to Figure 4d, the predicted training-set size of $n_{95\%}$ is used as the maximum training size (N) to fit the LC with the inverse-power law. On the basis of $n_{95\%} = 142$, the training region of the fitted model fits the LC quite well and extrapolates the performance of the inverse-power law perfectly. However, to evaluate this performance, we compared the RMSEs in the training and extrapolated regions. For the training region, the RMSE is 0.0877%, which decreases to 0.0529% for the extrapolated region. Finally, the previous results can be summarized as the following: the predicted size $n_{95\%} = 142$ spectra per class is required as a minimal training-set size to construct a PCA−LDA model, and the classification-error rate based on this training size represents 95% of the final performance of the PCA−LDA model

■ SUMMARY AND CONCLUSIONS

In our paper we provided a sample-size-planning (SSP) algorithm for multivariate data and tested it for a Raman-biospectroscopic data set. The core of this algorithm was established on the basis of learning curves (LCs), which describe the classification-error rate as a function of training-set size. The proposed algorithm started by LC generation, and then these LCs were fitted with the inverse-power law. The parameters of this model were used afterward to predict the training-set size of $n_{95\%}$. The final part of our algorithm was the evaluation of training-set-size prediction, which was done on the basis of the extrapolation of the inverse-power-law model, and then we compared the observed RMSE in the training region and the extrapolated region.

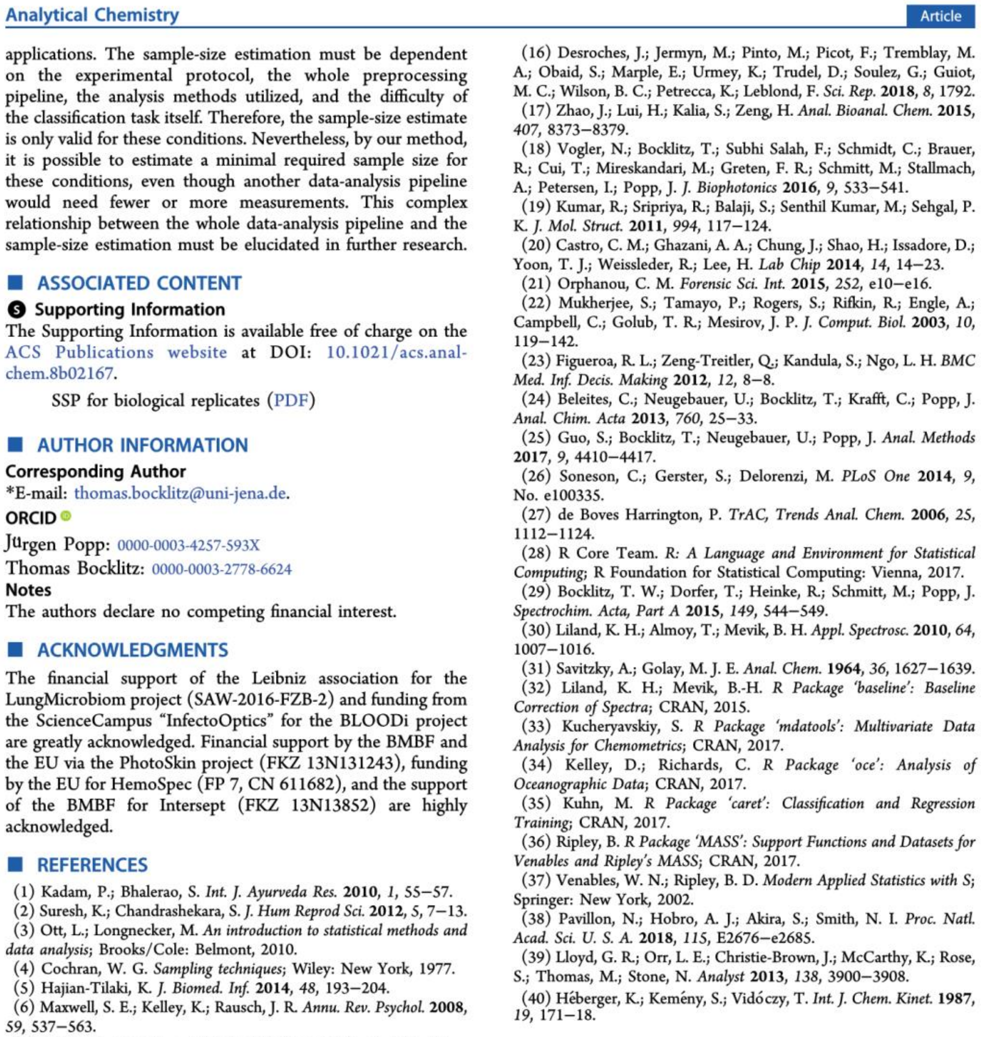
As mentioned above, the implementation of this algorithm was tested on a standard Raman-spectral data set. This data set consists of six bacterial species, which were cultivated in nine independent biological replicates. Then, the collected Raman spectra were uploaded into the statistical programming language R and preprocessed in order to reduce unwanted variations. After the preprocessing, we defined two SSP tasks (SSP for biological replicates and SSP for spectra), and we focused on presenting the SSP algorithm for biological replicates. The main reason is that all previous studies performed SSP for single measurements (spectra or genes), and no SSP method for the prediction of training-set size for biological replicates or patients exists. However, our algorithm was designed to predict the training-set sizes for both tasks (i.e., spectra and replicates).

The statistical part of the SSP algorithm is the following. A combination of PCA−LDA and leave-one-replicate-out cross-validation (LORO-CV) was implemented as the classification model and evaluation method. This combination of PCA−LDA and LORO-CV was used first to study the influence of the number of principle components (PCs) on the prediction of training-set size. Therefore, the SSP algorithm was checked for three ranges of PCs ({1, 2, ..., 10}, {1, 2, ..., 20}, and {1, 2, ..., 30}); then, the selection of the number of PCs was made according to the improvement in classification results for each single test replicate and the mean sensitivity. We decided to use 20 PCs to construct the PCA−LDA model for both SSP tasks.

Applying the proposed algorithm determined that seven biological replicates were required to achieve 95% of the final performance of the PCA−LDA model. The final performance, which was given by the Bayes error rate, was 18.67%, and the classification-error rate by seven replicates was 19.01%. Finally, we evaluated the performance of our prediction by fitting the inverse-power law with $n_{95\%} = 7$ replicates and then extrapolating the performance of this fit. The obtained extrapolation quality was very good, which is illustrated by the RMSE values for the training and extrapolated regions.

After predicting the required number of biological replicates, we implemented our SSP algorithm to estimate the required number of spectra. Here, the parameters in the SSP description were utilized. We generated the averaged LC and fitted this LC with the inverse-power law. The obtained fitted model approximated the LC quite well, and the final performance of the PCA−LDA model was 19.13%. On the basis of this performance, the predicted training-set size was 142 spectra per class, and the classification mean sensitivity was 81.75%. Again, the evaluation was carried out by comparing the RMSEs of the extrapolated region and the training region. This evaluation showed that the inverse-power-law model based on $n_{95\%} = 142$ was perfect for approximating the LC.

To conclude this work, the proposed SSP algorithm showed very good performance in predicting the required training-set sizes for both SSP tasks (spectra and biological replicates). By our algorithm the sample size was predicted, which was necessary for building a reliable and accurate PCA−LDA model. The test of our SSP algorithm was performed for a Raman-biospectroscopic data set, but the methodology can be applied for any multivariate data set. The SSP method is especially suitable for biomedical studies if the required number of biological replicates and spectral measurements need to be calculated. It should be noted here that the methodology of our SSP algorithm is the same for both



applications. The sample-size estimation must be dependent on the experimental protocol, the whole preprocessing pipeline, the analysis methods utilized, and the difficulty of the classification task itself. Therefore, the sample-size estimate is only valid for these conditions. Nevertheless, by our method, it is possible to estimate a minimal required sample size for these conditions, even though another data-analysis pipeline would need fewer or more measurements. This complex relationship between the whole data-analysis pipeline and the sample-size estimation must be elucidated in further research.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.8b02167.

SSP for biological replicates (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: thomas.bocklitz@uni-jena.de.

ORCID

Jürgen Popp: 0000-0003-4257-593X

Thomas Bocklitz: 0000-0003-2778-6624

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The financial support of the Leibniz association for the LungMicrobiom project (SAW-2016-FZB-2) and funding from the ScienceCampus "InfectoOptics" for the BLOODi project are greatly acknowledged. Financial support by the BMBF and the EU via the PhotoSkin project (FKZ 13N131243), funding by the EU for HemoSpec (FP 7, CN 611682), and the support of the BMBF for Intersept (FKZ 13N13852) are highly acknowledged.

■ REFERENCES

(1) Kadam, P.; Bhalerao, S. *Int. J. Ayurveda Res.* **2010**, *1*, 55−57.

(2) Suresh, K.; Chandrashekara, S. *J. Hum Reprod Sci.* **2012**, *5*, 7−13.

(3) Ott, L.; Longnecker, M. *An introduction to statistical methods and data analysis*; Brooks/Cole: Belmont, 2010.

(4) Cochran, W. G. *Sampling techniques*; Wiley: New York, 1977.

(5) Hajian-Tilaki, K. *J. Biomed. Inf.* **2014**, *48*, 193−204.

(6) Maxwell, S. E.; Kelley, K.; Rausch, J. R. *Annu. Rev. Psychol.* **2008**, *59*, 537−563.

(7) Kelley, K.; Rausch, J. R. *Psychol Methods* **2006**, *11*, 363−385.

(8) Stöckel, S.; Kirchhoff, J.; Neugebauer, U.; Rösch, P.; Popp, J. *J. Raman Spectrosc.* **2016**, *47*, 89−109.

(9) Read, D. S.; Whiteley, A. S. *J. Microbiol. Methods* **2015**, *109*, 79−83.

(10) Walter, A.; Schumacher, W.; Bocklitz, T.; Reinicke, M.; Rosch, P.; Kothe, E.; Popp, J. *Appl. Spectrosc.* **2011**, *65*, 1116−1125.

(11) Meisel, S.; Stockel, S.; Rosch, P.; Popp, J. *Food Microbiol.* **2014**, *38*, 36−43.

(12) Bocklitz, T. W.; Guo, S.; Ryabchykov, O.; Vogler, N.; Popp, J. *Anal. Chem.* **2016**, *88*, 133−151.

(13) Hobro, A. J.; Kumagai, Y.; Akira, S.; Smith, N. I. *Analyst* **2016**, *141*, 3756−3764.

(14) Neugebauer, U.; Bocklitz, T.; Clement, J. H.; Krafft, C.; Popp, J. *Analyst* **2010**, *135*, 3178−3182.

(15) Chen, M.; McReynolds, N.; Campbell, E. C.; Mazilu, M.; Barbosa, J.; Dholakia, K.; Powis, S. J. *PLoS One* **2015**, *10*, No. e0125158.

(16) Desroches, J.; Jermyn, M.; Pinto, M.; Picot, F.; Tremblay, M. A.; Obaid, S.; Marple, E.; Urmey, K.; Trudel, D.; Soulez, G.; Guiot, M. C.; Wilson, B. C.; Petrecca, K.; Leblond, F. *Sci. Rep.* **2018**, *8*, 1792.

(17) Zhao, J.; Lui, H.; Kalia, S.; Zeng, H. *Anal. Bioanal. Chem.* **2015**, *407*, 8373−8379.

(18) Vogler, N.; Bocklitz, T.; Subhi Salah, F.; Schmidt, C.; Brauer, R.; Cui, T.; Mireskandari, M.; Greten, F. R.; Schmitt, M.; Stallmach, A.; Petersen, I.; Popp, J. *J. Biophotonics* **2016**, *9*, 533−541.

(19) Kumar, R.; Sripriya, R.; Balaji, S.; Senthil Kumar, M.; Sehgal, P. K. *J. Mol. Struct.* **2011**, *994*, 117−124.

(20) Castro, C. M.; Ghazani, A. A.; Chung, J.; Shao, H.; Issadore, D.; Yoon, T. J.; Weissleder, R.; Lee, H. *Lab Chip* **2014**, *14*, 14−23.

(21) Orphanou, C. M. *Forensic Sci. Int.* **2015**, *252*, e10−e16.

(22) Mukherjee, S.; Tamayo, P.; Rogers, S.; Rifkin, R.; Engle, A.; Campbell, C.; Golub, T. R.; Mesirov, J. P. *J. Comput. Biol.* **2003**, *10*, 119−142.

(23) Figueroa, R. L.; Zeng-Treitler, Q.; Kandula, S.; Ngo, L. H. *BMC Med. Inf. Decis. Making* **2012**, *12*, 8−8.

(24) Beleites, C.; Neugebauer, U.; Bocklitz, T.; Krafft, C.; Popp, J. *Anal. Chim. Acta* **2013**, *760*, 25−33.

(25) Guo, S.; Bocklitz, T.; Neugebauer, U.; Popp, J. *Anal. Methods* **2017**, *9*, 4410−4417.

(26) Soneson, C.; Gerster, S.; Delorenzi, M. *PLoS One* **2014**, *9*, No. e100335.

(27) de Boves Harrington, P. *TrAC, Trends Anal. Chem.* **2006**, *25*, 1112−1124.

(28) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, 2017.

(29) Bocklitz, T. W.; Dorfer, T.; Heinke, R.; Schmitt, M.; Popp, J. *Spectrochim. Acta, Part A* **2015**, *149*, 544−549.

(30) Liland, K. H.; Almoy, T.; Mevik, B. H. *Appl. Spectrosc.* **2010**, *64*, 1007−1016.

(31) Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627−1639.

(32) Liland, K. H.; Mevik, B.-H. *R Package 'baseline': Baseline Correction of Spectra*; CRAN, 2015.

(33) Kucheryavskiy, S. *R Package 'mdatools': Multivariate Data Analysis for Chemometrics*; CRAN, 2017.

(34) Kelley, D.; Richards, C. *R Package 'oce': Analysis of Oceanographic Data*; CRAN, 2017.

(35) Kuhn, M. *R Package 'caret': Classification and Regression Training*; CRAN, 2017.

(36) Ripley, B. *R Package 'MASS': Support Functions and Datasets for Venables and Ripley's MASS*; CRAN, 2017.

(37) Venables, W. N.; Ripley, B. D. *Modern Applied Statistics with S*; Springer: New York, 2002.

(38) Pavillon, N.; Hobro, A. J.; Akira, S.; Smith, N. I. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, E2676−e2685.

(39) Lloyd, G. R.; Orr, L. E.; Christie-Brown, J.; McCarthy, K.; Rose, S.; Thomas, M.; Stone, N. *Analyst* **2013**, *138*, 3900−3908.

(40) Héberger, K.; Kemény, S.; Vidóczy, T. *Int. J. Chem. Kinet.* **1987**, *19*, 171−18.

Supporting Information for:

Sample size planning for multivariate data: a Raman spectroscopy based example

Nairveen Ali[§,¥], Sophie Girnus[§], Petra Rösch[§], Jürgen Popp[§,¥,ǂ,£], and Thomas Bocklitz[§,¥,*]

[§] Institute of Physical Chemistry and Abbe Center of Photonics (IPC), Friedrich-Schiller-University, Helmholtzweg 4, D-07743 Jena, Germany

[¥] Leibniz Institute of Photonic Technology (IPHT), Albert-Einstein-Straße 9, D-07745 Jena, Germany

[ǂ] Center for Sepsis Control and Care (CSCC), Jena University Hospital, Erlanger Allee 101, D-07747 Jena, Germany

[£] InfectoGnostics, Forschungscampus Jena, Philosophenweg 7, D-07743 Jena, Germany

[*] Corresponding Author: thomas.bocklitz@uni-jena.de

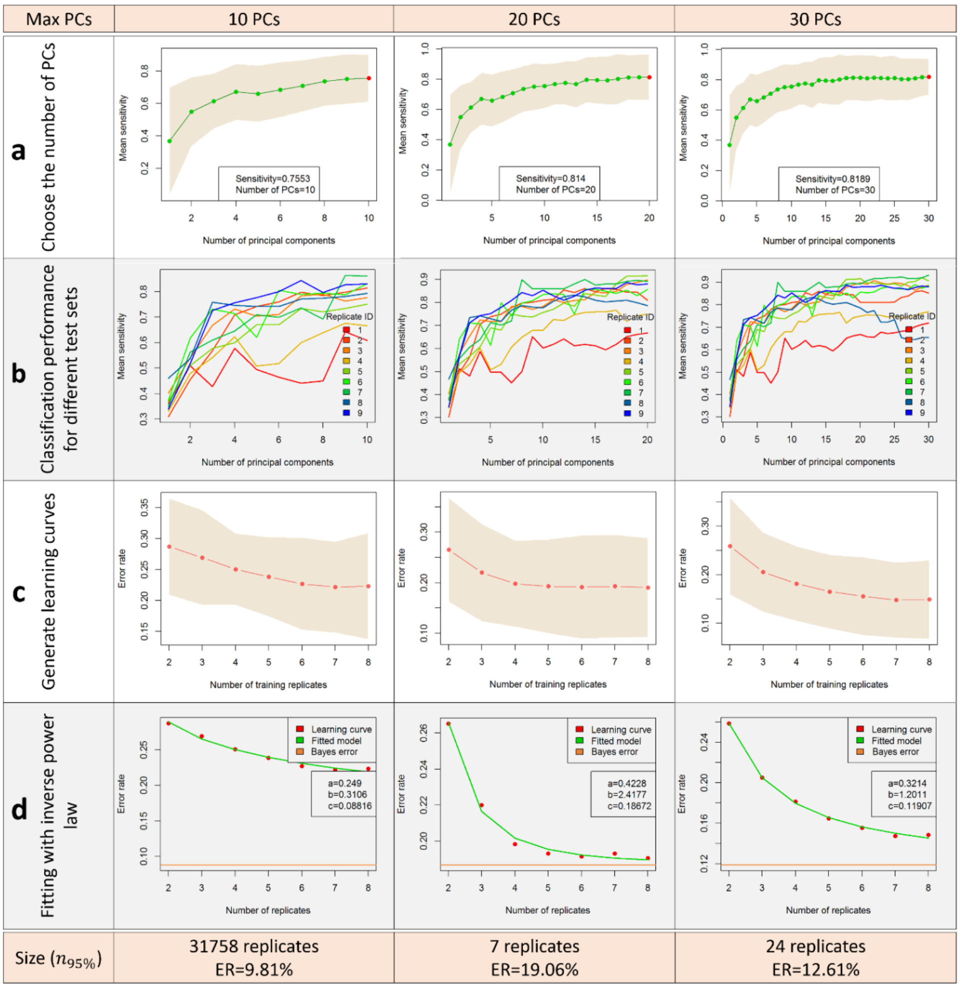


Figure SI. SSP for biological replicates and determination of PCs. a) The SSP algorithm was checked for different ranges of PCs ({1,2, … ,10}, {1,2, … ,20} and {1,2, … ,30}). The choice of the number of PCs was done due to the highest classification mean sensitivity, which was the maximal number of PC in the respective range. This number of PCs was utilized to build a PCA-LDA model with different numbers of training replicates. b) Here, the LCs for different test replicates are shown. All biological replicates show improvement in the classification results by increasing the number of PCs to 20. After that the classification improvement becomes less significant and sometimes decreases for instance for replicate 8. c) The generated average LC with standard deviation using 10, 20, and 30 PCs. d) The fitted models by inverse power law. For 10 PCs we can note that the classification performance still improves by increasing the number of training replicates. This explains the reason behind the non-convergence between the LC and Bayes error. In this case, the required number of training replicates is too high. At least 31758 biological replicates are needed to obtain classification mean sensitivity of 90.19%. This large estimate means that dimensionality of the model is too small for the given task. By using 20 or 30 PCs, the LCs converge faster to its asymptotes (Bayes error *c*). Only 7 biological replicates were required to build a PCA-LDA model with 20 PCs and classification mean sensitivity around 81%. This number increased to 24 biological replicates when we utilized 30 PCs.

II. WE-ASCA: The Weighted-Effect ASCA for Analyzing 3 Unbalanced Multifactorial Designs – A Raman Spectra Based-Example

The Reprinted [N. Ali, J. Jansen, A. Doel, G. H. Tinnevelt, and T. Bocklitz: WE-ASCA: The Weighted-Effect ASCA for Analyzing 3 Unbalanced Multifactorial Designs – A Raman Spectra Based-Example, Molecules, 2020, 26 (1), 66] is licensed under a Creative Commons Attribution-NonCommercial NoDerivatives 4.0 International License.

Erklärungen zu den Eigenanteilen des Promovenden sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation.

N. Ali, J. Jansen, A. Doel, G. H. Tinnevelt, and T. Bocklitz. *WE-ASCA: The Weighted-Effect ASCA for Analyzing 3 Unbalanced Multifactorial Designs – A Raman Spectra Based-Example*, Molecules, 2020, 26 (1), 66					
Beteiligt an (Zutreffendes ankreuzen)					
	1	2	3	4	5
Konzeption des Forschungsansatzes	×	×			×
Planung der Untersuchungen	×	×	×	×	×
Datenerhebung					
Datenanalyse und -interpretation	×	×	×	×	×
Schreiben des Manuskripts	×	×	×	×	×
Vorschlag Anrechnung Publikationsäquivalente	1.0				

molecules

MDPI

Article

WE-ASCA: The Weighted-Effect ASCA for Analyzing Unbalanced Multifactorial Designs—A Raman Spectra-Based Example

Nairveen Ali [1,2], Jeroen Jansen [3], André van den Doel [3,4], Gerjen Herman Tinnevelt [3] and Thomas Bocklitz [1,2,*]

[1] Institute of Physical Chemistry and Abbe Center of Photonics (IPC), Friedrich-Schiller-University, Helmholtzweg 4, D-07743 Jena, Germany; nairveen.ali@uni-jena.de
[2] Leibniz Institute of Photonic Technology (Leibniz-IPHT), Member of Leibniz Research Alliance Health Technologies, Albert-Einstein-Strasse 9, D-07745 Jena, Germany
[3] Institute for Molecules and Materials (IMM), Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands; jj.jansen@science.ru.nl (J.J.); h.vandendoel@science.ru.nl (A.v.d.D.); G.Tinnevelt@science.ru.nl (G.H.T.)
[4] TI-COAST, Science Park 904, 1098 XH Amsterdam, The Netherlands
[*] Correspondence: thomas.bocklitz@uni-jena.de; Tel.: +49-3541-948328

Abstract: Analyses of multifactorial experimental designs are used as an explorative technique describing hypothesized multifactorial effects based on their variation. The procedure of analyzing multifactorial designs is well established for univariate data, and it is known as analysis of variance (ANOVA) tests, whereas only a few methods have been developed for multivariate data. In this work, we present the weighted-effect ASCA, named WE-ASCA, as an enhanced version of ANOVA-simultaneous component analysis (ASCA) to deal with multivariate data in unbalanced multifactorial designs. The core of our work is to use general linear models (GLMs) in decomposing the response matrix into a design matrix and a parameter matrix, while the main improvement in WE-ASCA is to implement the weighted-effect (WE) coding in the design matrix. This WE-coding introduces a unique solution to solve GLMs and satisfies a constrain in which the sum of all level effects of a categorical variable equal to zero. To assess the WE-ASCA performance, two applications were demonstrated using a biomedical Raman spectral data set consisting of mice colorectal tissue. The results revealed that WE-ASCA is ideally suitable for analyzing unbalanced designs. Furthermore, if WE-ASCA is applied as a preprocessing tool, the classification performance and its reproducibility can significantly improve.

Keywords: ASCA; unbalanced experimental design; general linear model; weighted-effect coding; biomedical Raman spectra

check for updates

Citation: Ali, N.; Jansen, J.; van den Doel, A.; Tinnevelt, G.H.; Bocklitz, T. WE-ASCA: The Weighted-Effect ASCA for Analyzing Unbalanced Multifactorial Designs—A Raman Spectra-Based Example. *Molecules* **2021**, *26*, 66. https://dx.doi.org/10.3390/molecules26010066

Academic Editor: Ewa Sikorska
Received: 11 November 2020
Accepted: 22 December 2020
Published: 25 December 2020

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

An essential part of statistical analysis is the extraction of informative features that describe a specific phenomenon based on a limited number of samples. These samples are mostly collected by conducting either experiments or surveys [1,2]. In survey studies, a large number of individuals are involved to collect information without changing the existing conditions of the studied phenomenon. The other type of sampling is to conduct an experiment that tests the effect of one, or more than one, treatment on selected individuals. This experimental approach is widely applied in the fields of the physical and life sciences. In such studies, an experiment is carefully designed so that the obtained results are objective and valid [3,4]. Here, the term "design of experiment" (DoE) refers to statistical techniques that deal with planning and analyzing controlled tests, which investigate the effect of the studied treatments on selected individuals. There are several techniques for designing the experiments and one of the most common designs is the factorial design [5]. Therein, experiments are planned to extract information based on investigating the effect of at least two treatments in one experiment [6]. These treatments are termed the experimental factors, and the combinations of these factors define their interactions.

Within biomedical studies, e.g., testing the efficiency of a new drug or checking a new technique for disease detection, the effect of experimental factors and their interactions are translated into different types of variations that can be categorized into two main groups: informative (or interesting) variations and disturbing (or unwanted) variations. The informative variations highlight the differences between different states like sample properties or disease states. In contrast, disturbing variations may be assigned to systematic perturbations within the experiment, which might negatively affect the results of further analyses. The later variation is very difficult to be controlled, and it mostly arises when many devices or different individuals are considered to perform an experiment. In this discourse, multifactorial analysis methods were introduced as powerful techniques to understand and analyze the variations within the factorial experimental design. The basic idea here is built upon hypothesis testing of more than two groups referring to factor levels. These factorial analysis tests were established quite well for univariate data, and they are known as analysis of variance (ANOVA) tests [1,7,8]. In one of their classical versions, namely the one-way ANOVA test, the effect of one factor on selected observations is studied based on testing the mean differences between the factor levels. The multi-way ANOVA tests search in a multifactorial design for significant effects based on checking the differences between the levels of each factor and each factor interaction. However, if the response data set is described by multiple features, only a few methods of multifactorial design were developed, which typically feature some limitations. For instance, in the basic form of multivariate-ANOVA (MANOVA) tests, an ANOVA test is performed for several response variables, which allows for studying the effect of one or more than one factor on these response variables [9,10]. The main restriction here is that performing MANOVA tests require a large number of measurements, e.g., sample size, compared to the number of variables (features). Such data are often not available, especially in modern technologies which introduce high dimensional measurements like spectra or images. Therefore, combining principal component analysis (PCA) models with ANOVA tests provided a solution to deal with the high dimensionality of response matrices in multifactorial designs [11]. The PC-ANOVA starts by fitting the response matrix with a PCA model, then the obtained principal components (PCs) are analyzed using ANOVA tests. Although the PC-ANOVA does not have the limitation respecting the sample size and number of response variables, some information related to factor contributions might be missed during the PCA projection. Besides, ANOVA simultaneous component analysis (ASCA) was presented as a powerful tool to deal with multivariate data in multifactorial designs [12,13]. In brief, ASCA methods decompose the response matrix into different effect matrices, which characterize the contribution of each effect in the designed model. These contributions are measured by the amount of variance explained by each effect. Thereafter, ASCA checks which effect contributes significantly to the considered model, and finally, the dimensions of each effect matrix are reduced based on a PCA model or a SCA model [14,15]. Later improvements of ASCA were introduced based on scaling the response matrix first, then applying the classical ASCA pipeline [16]. In this reference, it was shown how the considered scaling approach can affect the interpretation of the ASCA results. However, the proposed design of ASCA and its improvements are valid only for balanced designs, in which the levels of each factor have equal numbers of measurements. This constraint creates an additional limitation to the application area of ASCA. Thus, the ASCA+ was introduced later as an extension of ASCA to deal with unbalanced designs [17]. It utilizes general linear models (GLMs) to decompose the response matrix into two main terms: The estimated response matrix and the residual matrix, which refers to the estimation error [8,17]. Within ASCA+, the levels of each effect are coded using the deviation coding that has a main advantage concerning the variance maximization produced if the classical ASCA is applied on unbalanced designs [18].

In the paper, we review the implementation of ASCA and ASCA+ in unbalanced designs, and we introduce an updated extension of ASCA based on weighted-effect (WE) coding as a powerful tool to deal with these unbalanced designs. This WE-coding is a type of dummy coding that offers an attractive feature in which the sum of all level effects of a categorical variable is equal to zero [19,20]. Additionally, the results of WE-coding are identical to those obtained by deviation coding in balanced designs. The new ASCA extension, named WE-ASCA, substitutes the deviation coding of the design matrix in the ASCA+ method by the WE-coding in order to estimate the contributions of experiment effects. The performance of WE-ASCA was evaluated based on a Raman spectral data set of 47 individual mice for two different applications. In the first task, we analyzed the complex multifactorial design presented within the Raman data set using WE-ASCA, and we compared its results with the obtained ones by ASCA and ASCA+. Thereafter, the WE-ASCA was implemented as a preprocessing technique to improve the tissue classification. This was accomplished by applying the WE-ASCA to exclude disturbing variations from the training set. Later, a classification model was constructed using the new training set only, i.e., without disturbing variations, while the classification results based on WE-ASCA were compared with those obtained by the same classifier trained without using WE-ASCA-based preprocessing.

2. Results

Two applications of WE-ASCA are demonstrated in this section based on an unbalanced multifactorial design of a Raman spectral data set comprising 387 colorectal tissue scans that were collected from 47 mice. Within this study, the activity of the P53 gene (active +, inactive -) and the mice gender (male, female) were recoded while the Raman spectra of each scan were annotated as different tissue “types” representing normal, hyperplasia (HP), adenoma, and carcinoma tissue. Later, the biological variations of these colorectal tissues extracted from mice rectum or colon were evaluated based on the acquired Raman spectral scans [21]. In the presented work, a mean spectrum per tissue type was calculated resulting in 485 Raman spectra acquired from 387 scans. The number of mean spectra and number of scans differ because in one scan multiple tissue types can be present, yielding a higher number of mean spectra per scan. For example, a scan may contain cancer tissue and normal tissue, leading to two mean spectra for this scan. Figure 1 depicts the mean spectra of the tissue types beside the design of our experiment. Therein, the factors exhibit the sample location, the mouse gender, and the activity of the P53 gene. Notably, the number of spectra within the levels of each factor is different; thus, the introduced WE-ASCA is ideal for analyzing this unbalanced design.

In the following, the contributions of experimental factors in addition to their interactions are estimated based on the classical ASCA and its extensions ASCA+ and WE-ASCA. Then, an evaluation for the variance explained by each effect was performed. In the second subsection, the WE-ASCA is applied as a preprocessing technique to assess its performance in improving the classification of colorectal tissues.

2.1. A Comparison between Analyses of Multifactorial Design Using ASCA, ASCA+ and WE-ASCA

Since 47 mice were included to perform this study, an additional variation connected to the biological differences between mice might be produced. We added therefore another factor featuring the individuals (mice) contribution to the experimental design. Consequently, the multifactorial model that describes the considered experiment was built upon the individual factor with 47 levels indicating the mice, the activity of the P53 gene, the mouse gender, and the sample location (colon or rectum) in addition to interactions between the last three factors. This mathematical model can be formulated as:

$$\mathbf{X} = \mathbf{M}_0 + \mathbf{Con}_{\mathbf{Individual}} + \mathbf{Con}_{\mathbf{Location}} + \mathbf{Con}_{\mathbf{P53}} + \mathbf{Con}_{\mathbf{Gender}} + \mathbf{Con}_{\mathbf{Loction:P53}} + \mathbf{Con}_{\mathbf{Location:Gender}} + \mathbf{Con}_{\mathbf{P53:Gender}} + \hat{\mathbf{E}}, \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^{485\times696}$ is the Raman spectral matrix, $\mathbf{M}_0$ denotes a matrix in which mean Raman spectra with respect to the wavenumbers are oriented in rows, and $\mathbf{Con}_f$ represents the matrix of an effect f. Based on this model, we evaluated the results of classical ASCA and its extensions ASCA+ and WE-ASCA using the obtained percentages of variances. As it is displayed in Table 1, the residual matrix of all analyses introduced the largest percentage of variance while the individual factor produced the largest factor contribution among all other effects. Here, the percentages of variance explained by the individuals are 37.97%, 33%, and 33.94% when ASCA, ASCA+, and WE-ASCA are applied, respectively. In contrast, the remaining factors and their interactions showed quite small contributions to the overall variance if any of the three multifactorial analyses were implemented. Nevertheless, the sum of percentages of variances by ASCA, ASCA+, and WE-ASCA is 108.62%, 93.3%, and 96.01%, respectively. This means that the classical ASCA maximized the effect contributions while the ASCA+ analysis minimized these contributions; however, the WE-ASCA analysis introduced the best estimations of effect contributions among the other analyses.

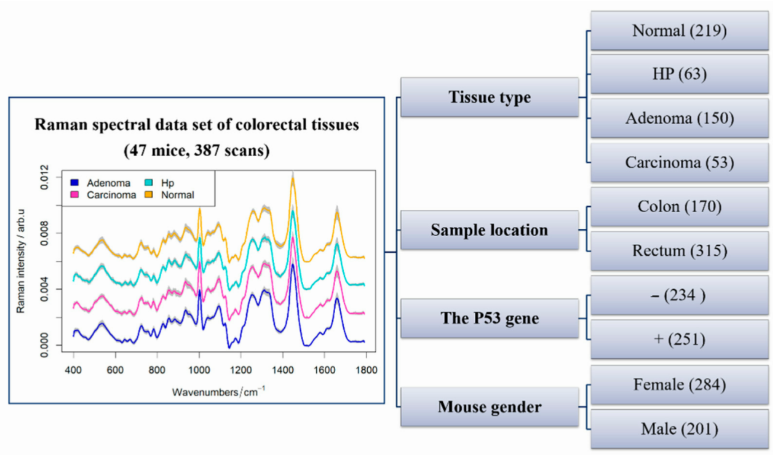


Figure 1. An overview of the experimental design of the Raman spectral data set. The studied data set consists of 47 mice and 387 scans that were collected from four different tissue types. The means spectra of these tissue types are shown in the left side. In this experiment, three factors were investigated: the activity of the P53 gene (active and inactive), the mouse gender, and the sample location (colon and rectum). It is observed that the number of scans of factor levels is different.

Table 1. The variance explained by the experimental effects in percentage (%) using ANOVA-simultaneous component analysis (ASCA), ASCA+ and WE-ASCA. A value of zero represents a percentage smaller than 10^{-9}.

Effect	Individual	Location	P53 Gene	Gender	Location: P53	Location: Gender	P53: Gender	Residuals	Sum (%)
ASCA	37.96	1.47	1.07	1.03	0.25	0.28	0.94	65.06	108.62
ASCA+	33.00	0.53	0	0	0.19	0.18	0	59.38	93.3
WE-ASCA	33.94	0.61	0	0	0.20	0.19	0	61.07	96.01

The obtained results in Table 1 show that only the individual factor and the sample location in addition to its interactions with the P53 gene and the mouse gender contributed to explain the variance in the considered design. To interpret the inner variance of these effects, a sperate PCA model was fitted to each of previous effect matrices, then the obtained PCA results were summarized in Table 2 and Figure 2. The column named "all data" in Table 2 shows the percentages of variance explained by the first two PCs if the spectral matrix **X** was mean centered and projected by a PCA model. These two PCs explained around 57.72% of the overall variance. For the PCA sub-model of the individual effect and the sample location, the three multifactorial analyses estimated approximately the same percentage of variance by the first two PCs. Therein, the first two PCs of the individual sub-models explained between 53.68% and 55.57% of the variance presented in the individual effect matrix, while the first PC estimated almost the whole variance of the sample location effect. Moving to the interaction effects, the first PC could describe almost 100% of the effect's variance if ASCA+ and WE-ASCA were applied. But this variance estimation was different in the case of classical ASCA, where the first PC explained around 3% variance less of this interaction effect. In Figure 2, the score plots of the PCA sub-models extracted from the effect matrix of the sample location and its interaction with the P53 gene and mouse gender are presented. The points in this figure depict the sum of the contribution of each effect and the projection of the residual matrix on the loadings obtained by the PCA sub-models (see [22] for more details). The bold points represent the group means and the ellipses are a representation of the covariance matrix. The WE-ASCA here shows slightly better separation between the group means in comparison to the results obtained by the ASCA and ASCA+.

After calculating the effect contributions and the PCA sub-models, we determined which effects contributed significantly to the experimental design using both ASCA extensions, i.e., ASCA+ and WE-ASCA, based on the described permutation test with $N = 1000$ iterations. The obtained p-values $\mathrm{p}(f)$ by these tests were combined and presented in Table 3. Whether ASCA+ or WE-ASCA are applied, the individual factor and the sample location factor caused a significant effect in the design of our experiment. Here, the obtained $\mathrm{p}(f)$ of the individual factor is almost zero by both analyses, while the $\mathrm{p}(f)$ of the sample location is 0.021% and 0.034% when WE-ASCA and ASCA+ are applied, respectively.

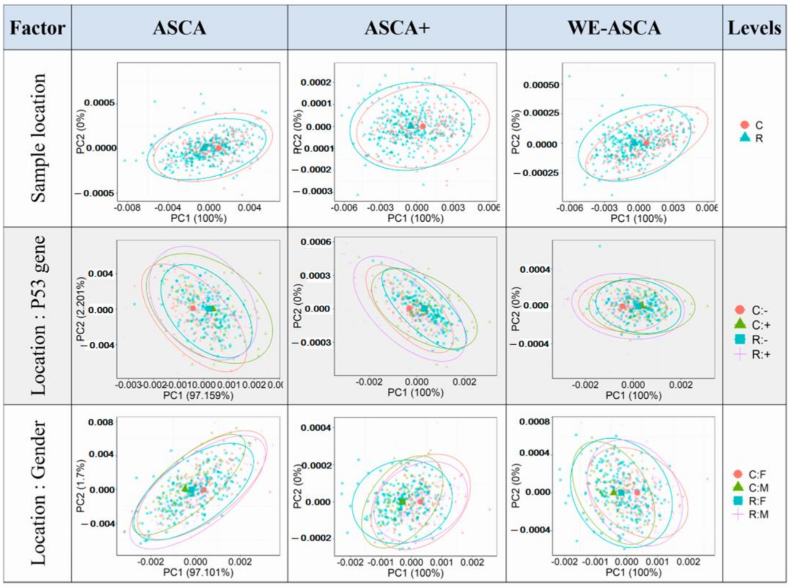


Figure 2. The score plots of first two principal components (PCs) of principal component analysis (PCA) sub-models using ASCA, ASCA+ and WE-ASCA. The WE-ASCA analysis provides better separation between the group means in comparison to the results obtained by the classical ASCA or its extension ASCA+.

Table 2. The percentage (%) of variance explained by the first two principal components of the PCA models. The mice data set and effect matrices obtained by ASCA, ASCA+, and WE-ASCA are fitted with a PCA model.

Effect		**All Data**	**Individuals**	**Location**	**Location: P53**	**Location: Gender**
ASCA	PC1 (%)	42.52	36.44	100	97.16	97.10
	PC2 (%)	15.20	17.74	0	2.20	1.7
ASCA+	PC1 (%)	42.52	38.78	100	100	100
	PC2 (%)	15.20	17.40	0	0	0
WE-ASCA	PC1 (%)	42.52	35.49	100	100	100
	PC2 (%)	15.20	18.19	0	0	0

Table 3. The obtained p-values $p(f)$ based on applying the permutation test on the results of ASCA+ and WE-ASCA analyses.

Effect		**Individuals**	**Location**	**Location: P53**	**Location: Gender**
$\boldsymbol{p(f)}$	ASCA+	0.000	0.034	0.399	0.453
	WE-ASCA	0.000	0.021	0.383	0.424

To conclude, the WE-ASCA improved the analysis of unbalanced multifactorial design within the considered Raman data set. This improvement was detected by comparing the sum of explained variance obtained by the classical ASCA and its extension ASCA+ with the results of the presented WE-ASCA. The analysis of this experiment yielded that the effect of the individual factor produced the highest variation and the most significant effect within the studied data set.

2.2. A WE-ASCA as Preprocessing Technique in Classification Models

The goal of this subsection is to assess the performance of the ASCA analyses as a preprocessing technique. Therein, the WE-ASCA was implemented within the cross-validation as a preprocessing tool in order to exclude disturbing variations. In our work, a leave-one-mouse-out cross-validation was performed to check the results of two classifiers, namely the combination of a principal component analysis with a linear discriminant analysis (PCA-LDA) and the combination of a partial least square regression with a linear discriminant analysis (PLS-DA). The classification and validation procedure starts by fixing spectra of a specific mouse $T_i : i = 1, 2, \ldots, 47$ as a test set and training the classifiers using the Raman spectra of the remaining mice, e.g., $\mathbf{X}(-T_i)$. This procedure is iterated until spectra of all mice are predicted once while the WE-ASCA is always applied on the training set of each cross-validation iteration (see Figure 3). It was shown in Section 2.1 that the individual factor produced the highest percentage of variance, which might negatively affect the classification results. Therefore, a new training set $\mathbf{X}_{\text{ASCA}}(-T_i)$ is estimated by excluding the variation of these individuals, then the training set $\mathbf{X}_{\text{ASCA}}(-T_i)$ can be defined as:

$$\mathbf{X}_{\mathbf{ASCA}}(-T_i) = \mathbf{X}(-T_i) - \mathbf{Con}_{\mathbf{Individual}}(-T_i); (i = 1, 2, \ldots, 47), \quad (2)$$

where $\mathbf{Con}_{\mathbf{Individual}}(-T_i)$ denotes to the individual effect matrix obtained by applying the WE-ASCA on $\mathbf{X}(-T_i)$. Using $\mathbf{X}_{\text{ASCA}}(-T_i)$ and $\mathbf{X}(-T_i)$, the PCA-LDA model and PLS-DA model are constructed. Then the tissue labels of each scan are predicted by both classifiers.

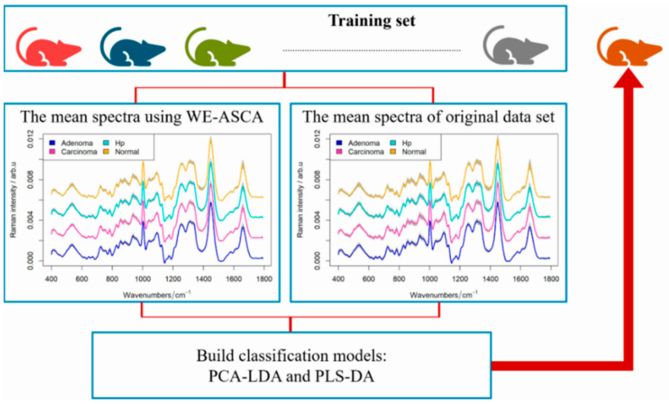


Figure 3. Overview of classification pipelines using WE-ASCA as a preprocessing step. Herein, a leave-one-mouse-out cross-validation (LOMO-CV) is implemented, while WE-ASCA is performed for each iteration within the CV loop. We can see that the variation of mean spectra after applying the WE-ASCA is smaller than the variation of mean spectra of the tissue types without preprocessing the spectral data. Based on training sets with or without applying WE-ASCA as preprocessing step, a PCA-LDA and PLS-DA are constructed, and their performance was determined.

The previous procedure was tested on four classification tasks, and the results are presented in Figure 4. The columns of this figure describe the results of PCA-LDA and PLS-DA models while the rows show the cross-validation results of each task for a different number of PCs or latent variables. If an LDA was trained on $\mathbf{X}_{\mathrm{ASCA}}(-T_i)$, the standard deviation and the mean sensitivity were presented by the yellow regions and the yellow lines. In the other case, the blue regions and blue lines indicate the standard deviation and the mean sensitivity of LDA models constructed by $\mathbf{X}(-T_i)$ without preprocessing by WE-ASCA. In Figure 4A, the classification means sensitivities of PCA-LDA and PLS-DA are presented in order to differentiate between the scans of normal, HP, adenoma, and carcinoma tissues. It is observed that both classifiers produced better results when they were trained by the $\mathbf{X}_{\mathrm{ASCA}}(-T_i)$. Here, the maximum mean sensitivity of the PCA-LDA model trained on $\mathbf{X}_{\mathrm{ASCA}}(-T_i)$ is 50.67%. If the same classifier was built on the training set $\mathbf{X}(-T_i)$, the mean sensitivity decreased to 42.93%. Constructing a PLS-DA model on a $\mathbf{X}_{\mathrm{ASCA}}(-T_i)$ or on $\mathbf{X}(-T_i)$ presented almost the same mean sensitivity, but it showed narrower standard deviation regions. For classifying adenoma and carcinoma tissues, it is clear in Figure 4B that using the WE-ASCA in preprocessing Raman spectra improved the LDA performance. While the maximum mean sensitivity of PCA-LDA and PLS-DA without implementing the WE-ASCA was 62.56% and 66%, respectively, the maximum mean sensitivity of PCA-LDA and PLS-LDA based on the training sets $\mathbf{X}_{\mathrm{ASCA}}(-T_i)$ increased to 67.98% and 68.47%, respectively. Moving to Figure 4C, which represents the classification results of the three suspicious tissues, WE-ASCA significantly improved the classification results, which can be observed as an increase in the mean sensitivity and a decrease of the standard deviation of PCA-LDA and PLS-DA models if they are trained by $\mathbf{X}_{\mathrm{ASCA}}(-T_i)$. In this case, the maximum mean sensitivity of PCA-LDA and PLS-LDA models increased at least 5% if they are constructed on the training sets $\mathbf{X}_{\mathrm{ASCA}}(-T_i)$. The last classification task aimed to differentiate between normal tissues and tumor tissues combining carcinoma and adenoma spectra in one class. The obtained results are presented in Figure 4D. Both classifiers provided almost the same maximum mean sensitivity based on the training sets $\mathbf{X}_{\mathrm{ASCA}}(-T_i)$ and $\mathbf{X}(-T_i)$. However, training the classification models on $\mathbf{X}_{\mathrm{ASCA}}(-T_i)$ decreased the standard deviation and required less latent variables for constructing PLS-LDA models. Nonetheless, utilizing WE-ASCA in preprocessing Raman spectra enhanced the results reproducibility of classification models. This reproducibility improvement is clearly seen in Figure 4 as narrower variation regions when the LDA models were trained by the $\mathbf{X}_{\text{ASCA}}(-T_i)$. Here, the standard deviations of both classifiers built on $\mathbf{X}_{\text{ASCA}}(-T_i)$ decreased significantly compared to the standard deviations of the same models trained by $\mathbf{X}(-T_i)$.

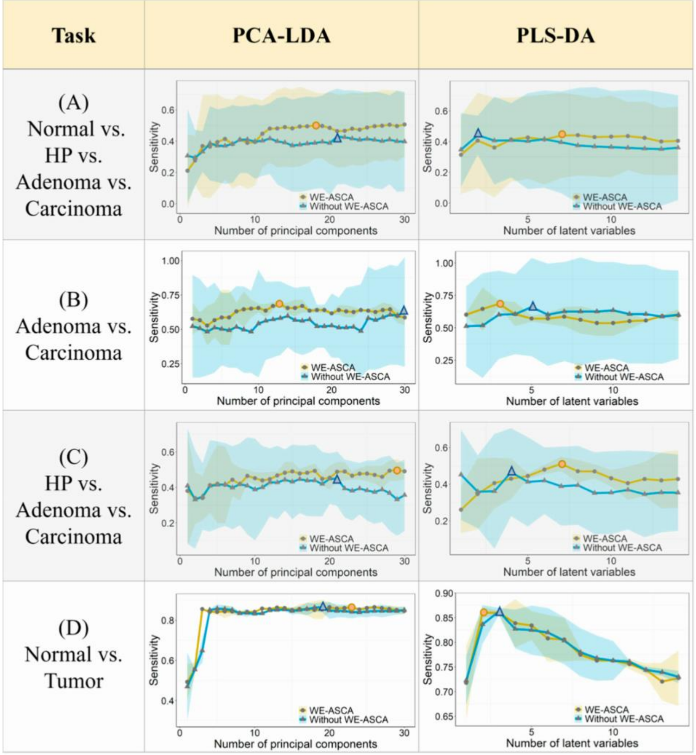


Figure 4. A comparison between the classification results of principal component analysis with a linear discriminant analysis (PCA-LDA) and partial least square regression with a linear discriminant analysis (PLS-DA) models based on leave-one-mouse-out cross-validation. Each classifier was trained twice with and without applying WE-ASCA-based preprocessing. The blue lines and the blue regions show the mean sensitivity and the standard deviation of a classifier constructed on the spectra without applying WE-ASCA on the training set. The yellow lines and the yellow regions depict the mean sensitivity and the standard deviation of a classification model trained on training sets that were preprocessed using WE-ASCA. The elimination of individual variations based on WE-ASCA improved the classification performance, and it significantly reduced the variance within the cross-validation results: (**A**) The maximum mean sensitivity for the differentiation between the scans of normal, HP, adenoma, and carcinoma tissues is 50.67%, and it was reached when training a PCA-LDA model on spectra processed by WE-ASCA. (**B**) For the classification of adenoma and carcinoma tissues, the maximum mean sensitivity of PCA-LDA (PLS-DA) is 67.98% (68.47%). These results were also achieved if the training sets were processed based on the WE-ASCA. (**C**) WE-ASCA-based preprocessing improved the differentiation between the three suspicious tissues. The maximum mean sensitivity of PCA-LDA model and PLS-DA are 49. 85% and 51.09%, respectively. (**D**) The results of differentiating the normal and tumor tissue. While, training an PCA-LDA model with or without spectra processed by WE-ASCA provided almost the same classification results, training a PLS-DA model on spectra processed by WE-ASCA improved the mean sensitivity and decreased thestandard deviation.

Overall, the WE-ASCA-based spectra al preprocessing allowed us to exclude the disturbing variation from the training data set, and it significantly improved the classification performance. This improvement can be detected as an increase in the classification mean sensitivities and a reduction of the variance in the cross-validation results. Furthermore, the WE-ASCA-based spectral preprocessing required a smaller number of principal components (or latent variables) to build the classification models since the data distortion in the training data set was eliminated.

3. Discussion

The weighted-effect ASCA (WE-ASCA) was introduced as a new extension of the classical ASCA to analyze multivariate data in unbalanced multifactorial designs. The core of this WE-ASCA is to use the weighted-effect (WE) coding in designing model matrices of GLMs instead of dummy coding and deviation coding considered in designing schemes of classical ASCA and its extension ASCA+, respectively. The main advantage of implementing the WE-coding is that the sum of all level effects of a categorical variable in the design matrix is equal to zero for unbalanced designs, which is not the case by the other coding schemes. Also, the WE-coding offers a unique estimation of the effect of a specific variable because it always codes a variable with α levels by $\alpha - 1$ columns within the design matrix. The described advantages convinced us to update the coding scheme utilized in the design matrix of ASCA+ with the WE-coding. The response matrix then can be estimated easily by a general linear model and using the new balanced design matrix and the parameter matrix. This estimated response can be decomposed linearly as different effect matrices representing the experimental factors and their interactions. Besides, the significant effects in a particular design are determined based on permutation tests while the dimensions of the effect matrices are reduced using PCA.

Using a Raman spectral data set consisting of four colorectal tissue types that were collected from 47 mice in 387 scans, two possible applications of WE-ASCA were checked. Here, the data set was acquired with respect to four factors describing the experimental design. These factors are the different individuals with 47 levels referring to the mice, the activity of the P53 gene, the mouse gender, and the location of samples (colon or rectum). In the first application, we aimed to understand and analyze the design of our experiment in addition to determining which of experimental factors contributed significantly to the considered experiment. This was achieved by applying ASCA, ASCA+, and WE-ASCA and comparing their results based on the explained variances by all effects. It tuned out that the classical ASCA overestimated the effect contributions, while the ASCA+ underestimated these contributions. In contrast, the presented WE-ASCA performed the best in estimating these effect contributions in term of the summation of percentage of explained variances. Nevertheless, the three versions of ASCA proved that the individual factor has the largest effect in our design. Therefore, we studied the influence of excluding such variations on the classification of colorectal tissues. This was demonstrated for four different classification tasks using two classifiers, i.e., PCA-LDA and PLS-DA, and leave-one-mouse-out cross-validation as a validation method. Our results showed that excluding the contribution of the individual factor from the training set introduced more robust classification results, and it improved the mean sensitivity in most classification tasks. Additionally, the training of an LDA model on spectra, when their individual effects were excluded, required a smaller number of principal components (or latent variables) and improved the reproducibility of the results.

4. Methods

4.1. ASCA for Crossed Balanced Design

The typical way of depicting the ASCA starts by introducing the ANOVA decomposition for the response of a single measurement described by a single variable [23]. Therein, the variation of each cell in a response matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, which has n measurements described by m variables (features), can be defined for a two-factor crossed design as:

$$x_{ijpq} = \mu_j + a_{jp} + b_{jq} + ab_{jpq} + \varepsilon_{ijpq},\ i = 1,\ 2,\ \ldots,\ n;\ j = 1,\ 2,\ \ldots,\ m;\ p = 1,\ 2,\ \ldots,\ \alpha; q = 1,\ 2,\ \ldots, \beta, \tag{3}$$

where x_{ijpq} is the observed response of a measurement, i, on the variable, j, that has the level p of factor A and the level q of factor B. The parameter μ_j indicates the global mean with respect to the variable j, and ε_{ijpq} refers to the error term which is supposed to be a Gaussian distributed random variable with a mean of 0. Under additional constrains of ANOVA described in [13,17,23], a unique estimation of the previous statistical model is:

$$x_{ijpq} = x_{.j..} + \left(x_{.jp.} - x_{.j..}\right) + \left(x_{.j.q} - x_{.j..}\right) + \left(x_{.jpq} - x_{.jp.} - x_{.j.q} + x_{.j..}\right) + \left(x_{ijpq} - x_{.jpq}\right). \tag{4}$$

The dot-notation in the previous equation's subscripts describes over which index the mean is calculated. Moving to multivariate data, the classical ASCA provides a direct generalization of the ANOVA tests in balanced designs. It calculates the effect contributions to the response matrix **X**:

$$\mathbf{X} = \mathbf{M}_0 + \mathbf{Con}_A + \mathbf{Con}_B + \mathbf{Con}_{AB} + \hat{\mathbf{E}}, \tag{5}$$

where $\mathbf{M}_0 \in \mathbb{R}^{n\times m}$ represents the global mean matrix (its rows are the means over the variables), $\mathbf{Con}_f \in \mathbb{R}^{n\times m}$ refers to the estimated effect contribution of f where $f \in \{A, B, AB\}$, and $\hat{\mathbf{E}} \in \mathbb{R}^{n\times m}$ estimates the residual matrix. When the experimental design is balanced, each estimated effect matrix in Equation (5) is orthogonal, and the summation of percentage of variances of these matrices equals to 100%. Consequently, the contributions of individual effects in the overall variance can be measured by the partitions of the following sums of squares:

$$\parallel \mathbf{X} \parallel^2 = \parallel \mathbf{M}_0 \parallel^2 + \parallel \mathbf{Con}_A \parallel^2 + \parallel \mathbf{Con}_B \parallel^2 + \parallel \mathbf{Con}_{AB} \parallel^2 + \parallel \hat{\mathbf{E}} \parallel^2, \tag{6}$$

where $\parallel . \parallel^2$ indicates the squared Frobenius norm.

4.2. General Linear Models and ASCA+

The general linear models (GLMs) usually refer to a multiple linear regression where a continuous response variable is given continuous and (or) categorical predictors. These GLMs are fundamentals for several statistical tests such as ANOVA and ANCOVA (analysis of covariance) [24]. In its multivariate version, GLMs aim to decompose a response matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ linearly into different contributions based on a design matrix $\mathbf{D} \in \mathbb{R}^{n\times p}$ and a parameter matrix $\boldsymbol{\beta} \in \mathbb{R}^{p\times m}$. These contributions are related to the experimental factors and their interactions while the linear decomposition can be formulated as the following equation:

$$\mathbf{X} = \mathbf{D}\boldsymbol{\beta} + \mathbf{E}, \tag{7}$$

where $\mathbf{E} \in \mathbb{R}^{n\times m}$ denotes to the residual matrix. In formula (7), the matrix $\boldsymbol{\beta}$ relies only on the data features like intensity or pixel values while the method of coding the design (model) matrix **D** performs a critical part in GLMs decomposition. In principle, the matrix **D** can be designed using different coding techniques; however, a unique estimation of effect contributions can be obtained only if a factor with α levels is coded by $\alpha - 1$ columns in the design matrix [17]. By ASCA+, the deviation coding was utilized to design the matrix **D**, where a factor with α levels is coded by $\alpha - 1$ columns with values of 0 and 1 for the first $\alpha - 1$ levels and with the value of -1 for the last level of this factor [17]. Then the parameter matrix $\boldsymbol{\beta}$ is estimated using the ordinary least square method as $\hat{\boldsymbol{\beta}} = \left(\mathbf{D}^{\mathrm{T}}\mathbf{D}\right)^{-1}\mathbf{D}^{\mathrm{T}}\mathbf{X}$, and the data matrix **X** is approximated as $\hat{\mathbf{X}} = \mathbf{D}\ \hat{\boldsymbol{\beta}}$. The error of this estimation is determined by the residual matrix $\hat{\mathbf{E}}$ which can be written mathematically as:

$$\hat{\mathbf{E}} = \mathbf{X} - \mathbf{D}\hat{\boldsymbol{\beta}}. \tag{8}$$

Nevertheless, the main advantage of the previous coding is that the sub-design matrix of each effect is orthogonal for a balanced design. For instance, suppose a balanced two-factor crossed design is considered in which six measurements are affected by the factors $A : (a_1,\ a_2)$ and $B :\ (b_1,\ b_2,\ b_3)$. A response matrix $\mathbf{X} \in \mathbb{R}^{6\times m}$ of this balanced design can be described with respect to these factor levels as:

$$\mathbf{X} = \left[X_{1,a_1b_1}\ X_{2,a_1b_2}\ X_{3,a_1b_3}\ X_{4,a_2b_1}\ X_{5,a_2b_2}\ X_{6,a_2b_3}\right]^{\mathrm{T}}. \tag{9}$$

$X_{i,a_kb_l} \in \mathbb{R}^m$ indicates a measurement of level $a_p : p \in \{1,\ 2\}$ and level $b_q : q \in \{1,\ 2,\ 3\}$, which is oriented in the row $i \in \{1,\ 2, \ldots,\ 6\}$. Based on the ASCA+ and the deviation coding, the design matrix $\mathbf{D}$ and the sub-design matrix of factor B can be visualized by:

$$\mathbf{D} = \left[\begin{array}{cccccc} \mathrm{M}_0 & A & b_1 & b_2 & A:b_1 & A:b_2 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{array}\right]; \mathbf{D}_{/B} = \left[\begin{array}{cccccc} \mathrm{M}_0 & A & b_1 & b_2 & A:b_1 & A:b_2 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \end{array}\right].$$

Then, the effect matrix of factor B, namely $\mathbf{Con}_B$, is simply estimated by multiplying the matrix $\mathbf{D}_{/B}$ and the parameter matrix $\hat{\boldsymbol{\beta}}$. Likewise, all sub-design matrices and effect matrices can be calculated, and the response matrix $\mathbf{X}$ is subsequently decomposed according to formula (3) while the contribution of each effect into overall variance is estimated using a squared Frobenius norm (see Equation (6)).

4.3. Weighted-Effect ASCA (WE-ASCA)

The deviation coding introduced by ASCA+ provides an orthogonal design matrix only if an experimental design is balanced. In this case, the traditional constrains of analysis of variance models are perfectly satisfied, and the summation of percentage of variance equals to 100%. For unbalanced multifactorial designs, the design matrices introduced by any of ASCA or ASCA+ are non-orthogonal, and the estimated experiment effects are biased; therefore, it is desirable to remove, or at least reduce, these estimation biases. Another coding scheme, which can be considered to design the model matrix of GLMs in unbalanced data, is the weighted-effect (WE) coding. This WE-coding is a type of dummy coding which can be used to facilitate the inclusion of categorical variables in GLMs [19,20]. Thereby, the effect of each level of a categorical variable represents the level deviation from the weighted mean instead of using the grand mean in deviation coding. The WE-coding offers an attractive property related to the constrain in which the sum of all level effects of a categorical variable is equal to zero, which is not fulfilled by other coding schemes in unbalanced designs [19]. Moreover, the results of WE-coding are identical with those obtained by deviation coding if an experiment's design is balanced. Beside this, the estimated effect of a specific variable provided by the WE-coding are unique because a variable of α levels is coded by $\alpha - 1$ columns within the model matrix, and the interaction between two variables of α and $\boldsymbol{\beta}$ levels is coded by $(\alpha - 1) \times (\boldsymbol{\beta} - 1)$ columns in this model matrix. Using the previous advantages, the WE-coding can be used to improve the performance of ASCA models in unbalanced multifactorial designs. In the following, the implementation of WE-coding in unbalanced multifactorial designs will be described in detail for a two-factor crossed design. However, this coding scheme is still valid for higher multifactorial designs. Let $\mathbf{X} \in \mathbb{R}^{7\times m}$ be a response matrix collected from an unbalanced two-factor crossed design and presented as:

$$\mathbf{X} = \left[X_{1,a_1b_1}\ X_{2,a_1b_2}\ X_{3,a_1b_3}\ X_{4,a_2b_1}\ X_{5,a_2b_2}\ X_{6,a_2b_3}\ X_{7,a_2b_3}\right]^{\mathrm{T}}, \tag{10}$$

where $X_{i,a_k b_l} \in \mathbb{R}^m$ indicates a measurement of level $a_p : p \in \{1,\ 2\}$ and level $b_q : q \in \{1,\ 2,\ 3\}$, which is oriented in rows $i \in \{1,\ 2, \ldots,\ 7\}$. According to ASCA+, the design matrix **D** and the sub-design matrix with respect to factor B can be depicted by:

$$\mathbf{D} = \left[\begin{array}{cccccc} \mathrm{M}_0 & A & b_1 & b_2 & A:b_1 & A:b_2 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{array}\right]; \mathbf{D}_{/B} = \left[\begin{array}{cccccc} \mathrm{M}_0 & A & b_1 & b_2 & A:b_1 & A:b_2 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \end{array}\right].$$

We can note that the design matrix **D** is non-orthogonal and that the sum of level effects of the factors and their interaction does not equal to zero, which introduces biased estimators of experimental effects. The previous design matrix can be converted into a balanced design matrix if the WE-coding described by the coding matrix in Table 4 is applied. Based on this table, the obtained balanced design matrix **BD** and the balanced sub-design matrix $\mathbf{BD}_{/B}$ are determined as the following:

$$\mathbf{BD} = \left[\begin{array}{cccccc} \mathrm{M}_0 & A & b_1 & b_2 & A:b_1 & A:b_2 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -2/3 & -2/3 & -1/1 & -1/1 \\ 1 & -3/4 & 1 & 0 & -1/1 & 0 \\ 1 & -3/4 & 0 & 1 & 0 & -1/1 \\ 1 & -3/4 & -2/3 & -2/3 & 1/2 & 1/2 \\ 1 & -3/4 & -2/3 & -2/3 & 1/2 & 1/2 \end{array}\right];$$

$$\mathbf{BD}_{/B} = \left[\begin{array}{cccccc} \mathrm{M}_0 & A & b_1 & b_2 & A:b_1 & A:b_2 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -2/3 & -2/3 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -2/3 & -2/3 & 0 & 0 \\ 0 & 0 & -2/3 & -2/3 & 0 & 0 \end{array}\right].$$

Here, each effect A, B, and AB is estimated in **BD** by $p-1=1$, $q-1=2$, and $(p-1) \times (q-1) = 2$ columns, respectively. Clearly, the considered WE-coding estimates the levels of effects A, B, and AB in a way that the sum is always equal to zero.

Table 4. The coding matrix of a two-factor crossed design based on the weighted-effect coding. In this design, the effects of factors $A : (a_1,\ a_2)$, $B : (b_1,\ b_2,\ b_3)$ and their interaction AB are presented.

Possible Combinations	**Factors**			**Interaction**	
	$\boldsymbol{A}$	$\boldsymbol{b_1}$	$\boldsymbol{b_2}$	$\boldsymbol{A{:}b_1}$	$\boldsymbol{A{:}b_2}$
$a_1\&b_1$	1	1	0	1	0
$a_1\&b_2$	1	0	1	0	1
$a_1\&b_3$	1	$-n_{b_1}/n_{b_3}$	$-n_{b_2}/n_{b_3}$	$-n_{a_1,\ b_1}/n_{a_1,\ b_3}$	$-n_{a_1,\ b_2}/n_{a_1,\ b_3}$
$a_2\&b_1$	$-n_{a_1}/n_{a_2}$	1	0	$-n_{a_1,\ b_1}/n_{a_2,\ b_1}$	0
$a_2\&b_2$	$-n_{a_1}/n_{a_2}$	0	1	0	$-n_{a_1,\ b_2}/n_{a_2,\ b_2}$
$a_2\&b_3$	$-n_{a_1}/n_{a_2}$	$-n_{b_1}/n_{b_3}$	$-n_{b_2}/n_{b_3}$	$n_{a_1,\ b_1}/n_{a_2,\ b_3}$	$n_{a_1,\ b_2}/n_{a_2,\ b_3}$

In this paper, we update the ASCA+ by replacing the deviation coding of the design matrix in GLMs by the WE-coding. This updated version, namely weighted-effect ASCA

(WE-ASCA), provides a new extension of ASCA in unbalanced multifactorial designs. Thereby, a balanced design matrix **BD** is estimated using the WE-coding [19,20], then the parameter matrix $\boldsymbol{\beta}$ is estimated based on the ordinary least square method:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{BD}^{\mathbf{T}}\mathbf{BD}\right)^{-1} \mathbf{BD}^{\mathbf{T}}\,\mathbf{X}. \tag{11}$$

The response matrix **X** can be thereafter approximated as $\hat{\mathbf{X}} = \mathbf{BD}\,\hat{\boldsymbol{\beta}}$, while the error of this estimation is given by:

$$\hat{\mathbf{E}} = \mathbf{X} - \mathbf{BD}\,\hat{\boldsymbol{\beta}}. \tag{12}$$

Because the WE-coding estimates each factor of α levels by $\alpha - 1$ columns in the design matrix **BD**, the presented WE-ASCA analysis provides a unique solution to solve the statistical models of any multifactorial design. Additionally, it reduces the bias introduced by unbalanced designs. Consequently, the effect matrix of any factor or interaction in multifactorial designs can be uniquely estimated by:

$$\mathbf{Con}_f = \mathbf{BD}_{/f}\,\hat{\boldsymbol{\beta}}, \tag{13}$$

where $\mathbf{BD}_{/f}$ and $\mathbf{Con}_f$ denotes the balanced sub-design matrix and the effect matrix of factor (or interaction) f, respectively. In case of a two-factor crossed design, the response matrix **X** is decomposed linearly into different effects of the factors A and B and their interaction:

$$\mathbf{X} = \mathbf{M}_0 + \mathbf{Con}_A + \mathbf{Con}_B + \mathbf{Con}_{AB} + \hat{\mathbf{E}}. \tag{14}$$

The previous estimated effect matrices have the same size of matrix **X**. Thus, it is useful to reduce the dimensions of these matrices using PCA models in order to highlight the variations between different effect levels. In the presented two-factor crossed design, the statistical model based on the PCA sub-models can be presented by the decomposition:

$$\mathbf{X} = \mathbf{M}_0 + \mathbf{T}_A\mathbf{P}_A^{\mathrm{T}} + \mathbf{T}_B\mathbf{P}_B^{\mathrm{T}} + \mathbf{T}_{AB}\mathbf{P}_{AB}^{\mathrm{T}} + \hat{\mathbf{E}}, \tag{15}$$

where the matrix $\mathbf{T}_f$ represents the score matrix, which highlights the variations between the levels of effect f. The matrix $\mathbf{P}_f$ denotes the loadings matrix of the same effect.

4.4. The Percentage of Variance

One of the main goals of multifactorial design analysis is to study how different factors influence a particular experiment based on estimating their contributions to the overall variance. In balanced designs, a factor contribution is approximated simply using the type I sum squares. This method of sum squares sequentially computes the factor contributions with respect to their order in the designed model [25]. For the unbalanced multifactorial design, the factor levels have different numbers of measurements, which provides overestimation of some factor contributions. It is recommended according to [8,25,26] to calculate these contributions based on the type III sum squares. Thereby, the effect of one factor is evaluated after all other factors have been considered. This type of sum squares offers identical estimations of factor contributions with those obtained by type I sum squares when the considered design is balanced. In Table 5, the type III sum squares of each effect contribution in a two-factor crossed design is presented. Herein, the mean model decomposes the response matrix **X** based on the global mean matrix $\mathbf{M}_0$ and the overall variance, named $\hat{\mathbf{E}}_1$. Then, the response matrix **X** is decomposed by a reduced model that does not consider the contribution of a specific effect f. Subsequently, the residual matrix $\hat{\mathbf{E}}_f$ of this reduced model is estimated, and the sum squares of the effect f is calculated by the difference between $\| \hat{\mathbf{E}}_f \|^2$ and $\| \hat{\mathbf{E}} \|^2$. The explained variance by each effect can be approximated finally as a percentage of the sum squares of that effect to the sum squares of the residual matrix of the mean model, i.e., $\mathrm{SS}\left(\hat{\mathbf{E}}_1\right)$. Mathematically, the percentage of variance explained by an effect $f \in \{A,\ B,\ AB\}$ and the percentage of variance explained by the residual $\hat{\mathbf{E}}$ of a two-factor crossed model can be formulated as:

$$\%\boldsymbol{Var}_f = \frac{\| \hat{\mathbf{E}}_f \|^2 - \| \hat{\mathbf{E}} \|^2}{\| \mathbf{X} - \mathbf{M}_0 \|^2} \times 100 \text{ and } \%\boldsymbol{Var}_{\hat{\mathbf{E}}} = \frac{\| \hat{\mathbf{E}} \|^2}{\| \mathbf{X} - \mathbf{M}_0 \|^2} \times 100. \tag{16}$$

Table 5. Type III sum squares for a two-factor crossed design.

Model		Type III Sum Squares
Mean model	$\mathbf{X} = \mathbf{M}_0 + \hat{\mathbf{E}}_1$	$SS\left(\hat{\mathbf{E}}_1\right) = \| \mathbf{X} - \mathbf{M}_0 \|^2$
Two-factor cross design	$\mathbf{X} = \mathbf{M}_0 + \mathbf{Con}_A + \mathbf{Con}_B + \mathbf{Con}_{AB} + \hat{\mathbf{E}}$	$SS\left(\hat{\mathbf{E}}\right) = \| \hat{\mathbf{E}} \|^2$
Without the effect of A	$\mathbf{X} = \mathbf{M}_0 + \mathbf{Con}_B + \mathbf{Con}_{AB} + \hat{\mathbf{E}}_A$	$SS(\mathbf{Con}_A) = \|_{2} \hat{\mathbf{E}}_A \|^2 - \| \hat{\mathbf{E}} \|$
Without the effect of B	$\mathbf{X} = \mathbf{M}_0 + \mathbf{Con}_A + \mathbf{Con}_{AB} + \hat{\mathbf{E}}_B$	$SS(\mathbf{Con}_B) = \| \hat{\mathbf{E}}_B \|^2 - \| \hat{\mathbf{E}} \|^2$
Without the effect of AB	$\mathbf{X} = \mathbf{M}_0 + \mathbf{Con}_A + \mathbf{Con}_B + \hat{\mathbf{E}}_{AB}$	$(\mathbf{Con}_{AB}) = \| \underset{2}{\hat{\mathbf{E}}}_{AB} \|^2 - \| \hat{\mathbf{E}} \|$

In our study, we compare the results of the WE-ASCA with the classical ASCA and its extension ASCA+ based on the summation of percentage of variances in an unbalanced design.

4.5. Permutation Tests

The basic idea of permutation tests is to check whether a specific effect f in ANOVA models contributes significantly to the variation of an experiment or it has a random influence [14,17]. For a two-factor crossed design, the procedure of permutation test starts by calculating the Frobenius sum squares of the first l principal components of the score matrix $\mathbf{T}_f$ for each effect $f \in \{A,\ B,\ AB\}$:

$$SS(f) = \sum_{i=1}^{n} \sum_{j=1}^{l} \left(\boldsymbol{T}_f\right)_{i,j}^2, \tag{17}$$

where n denotes the number of measurements in a response matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$. Then, we generate N random permutations of the rows of $\mathbf{X}$, and the Frobenius norm $SS_r(f)$ of each effect is computed for each permutation $r = 1,\ 2, \ldots, N$ with respect to the considered ASCA extension, i.e., ASCA+ and WE-ASCA. The last step of this test is to calculate the p-value $\mathrm{p}(f)$ as:

$$\mathrm{p}(f) = \frac{\#\{SS_r(f) \geq SS(f)\}}{N}. \tag{18}$$

This p-value determines whether an effect, *f*, explains a random variation or shows a significant contribution within a considered experiment.

4.6. Data and Software

All computational parts were carried out based on in-house written functions in R version 3.4.2. The utilized Raman spectral data and R functions are freely available via Zenodo through the following links:

- Raman spectra of colon cancer in a mice model: https://zenodo.org/deposit/3975464

- Weighted-effect ASCA (WE-ASCA) codes: https://zenodo.org/deposit/3975471

5. Conclusions

The presented WE-ASCA provides an updated version of the ASCA and ASCA+ that suits the analysis of variance in unbalanced multifactorial designs. WE-ASCA proved its potential in understanding and analyzing the influence of experimental factors in a complex multifactorial design. Furthermore, the WE-ASCA was presented as a powerful preprocessing tool that can improve the classification performance and increase the classification reproducibility. The current implementations of WE-ASCA were checked only for Raman spectra and for tissue classification tasks; however, the application field is not limited to these previous applications. It can be extended to cover the analysis of variance of any type of multivariate data and any statistical modeling task.

Author Contributions: Conceptualization, T.B., N.A., and J.J.; methodology, T.B., G.H.T., A.v.d.D., J.J., and N.A.; software, N.A. and T.B.; validation, N.A. and A.v.d.D.; formal analysis, G.H.T., T.B., and N.A.; writing—original draft preparation, N.A. and T.B.; writing—review and editing, T.B., J.J., A.v.d.D., N.A., and G.H.T.; visualization, N.A.; supervision, J.J. and T.B.; project administration, J.J. and T.B.; funding acquisition, T.B. and J.J. All authors have read and agreed to the published version of the manuscript.

Funding: The funding from the DFG for the project BO4700/4-1 is highly appreciated. Financial support by the BMBF for the project Uro-MDD (FKZ 03ZZ0444J) is highly acknowledged. This research received additionally funding from the Netherlands Organization for Scientific Research (NWO) in the framework of the Programmatic Technology Area PTA-COAST3 of the Fund New Chemical Innovations. This publication reflects only the authors' views and NWO is not liable for any use that may be made of the information contained herein. We acknowledge support by the German Research Foundation and the Open Access Publication Fund of the Thueringer Universitaets- und Landesbibliothek Jena Projekt-Nr. 433052568.

Data Availability Statement: The data of Vogler et al. [21] and scripts are available in the following repositories: Raman spectra of colon cancer in a mice model: https://zenodo.org/deposit/3975464; Weighted-effect ASCA (WE-ASCA) codes: https://zenodo.org/deposit/3975471.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Sample Availability: Not available.

References

1. Sampford, R.M.; Cochran, W.G. Sampling Techniques. *Biometrics* **1978**, *34*, 332–333. [CrossRef]
2. Smith, T.M.F.; Sugden, R.A. Sampling and Assignment Mechanisms in Experiments, Surveys and Observational Studies, Correspondent Paper. *Int. Stat. Rev.* **1988**, *56*, 165–180. [CrossRef]
3. Steinberg, D.M.; Hunter, W.G. Experimental Design: Review and Comment. *Technometrics* **1984**, *26*, 71–97. [CrossRef]
4. Winer, B.J. *Statistical Principles of Experimental Design*; McGraw-Hill: New York, NY, USA, 1962; Volume 3, pp. 381–385.
5. Mead, R.; Curnow, R.N.; Hasted, A.M. *Statistical Methods in Agriculture and Experimental Biology*, 3rd ed.; Chapman and Hall, CRC Press: Boca Raton, FL, USA, 2017; pp. 1–472.
6. Durakovic, B. Design of experiments application, concepts, examples: State of the art. *Period. Eng. Nat. Sci.* **2017**, *5*, 421–439. [CrossRef]
7. St»hle, L.; Wold, S. Analysis of variance (ANOVA). *Chemom. Intell. Lab. Syst.* **1989**, *6*, 259–272.
8. Christensen, R. Analysis of Variance and Generalized Linear Models. In *International Encyclopedia of the Social & Behavioral Sciences*; Smelser, N.J., Baltes, P.B., Eds.; Pergamon Press: Oxford, UK, 2001; pp. 473–480.
9. Mardia, K.V.; Kent, J.T.; Bibby, J.M. *Multivariate Analysis*; Academic Press: London, UK, 1979; Volume 37.
10. Martens, H.M. Multivariate Analysis of Quality. An Introduction. *Meas. Sci. Technol.* **2001**, *12*, 1746. [CrossRef]
11. Bratchell, N. Multivariate response surface modelling by principal components analysis. *J. Chemom.* **1989**, *3*, 579–588. [CrossRef]
12. Jansen, J.J.; Hoefsloot, H.C.J.; van der Greef, J.; Timmerman, M.E.; Westerhuis, J.; Smilde, A.K. ASCA: Analysis of multivariate data obtained from an experimental design. *J. Chemom.* **2005**, *19*, 469–481. [CrossRef]
13. Smilde, A.K.; Jansen, J.J.; Hoefsloot, H.C.J.; Lamers, R.-J.A.N.; van der Greef, J.; Timmerman, M.E. ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics* **2005**, *21*, 3043–3048. [CrossRef] [PubMed]
14. Anderson, M.; ter Braak, C. Permutation tests for multi-factorial analysis of variance. *J. Stat. Comp. Simul.* **2003**, *73*, 85–113. [CrossRef]

15. Bertinetto, C.; Engel, J.; Jansen, J.J. ANOVA simultaneous component analysis: A tutorial review. *Anal. Chim. Acta* **2020**, *6*, 100061. [CrossRef]
16. Timmerman, M.E.; Hoefsloot, H.C.J.; Smilde, A.K.; Ceulemans, E. Scaling in ANOVA-simultaneous component analysis. *Metabolomics* **2015**, *11*, 1265–1276. [CrossRef] [PubMed]
17. Thiel, M.; Féraud, B.; Govaerts, B. ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. *J. Chemom.* **2017**, *31*, e2895. [CrossRef]
18. Madsen, H.; Thyregod, P. *Introduction to General and Generalized LInear Models*; CRC Press: Boca Raton, FL, USA, 2010.
19. Nieuwenhuis, R.; Grotenhuis, M.; Pelzer, B. Weighted Effect Coding for Observational Data with wec. *R. J.* **2017**, *9*, 477–485. [CrossRef]
20. te Grotenhuis, M.; Pelzer, B.; Eisinga, R.; Niuwenhuis, R.; Schimdt-Catran, A.; Konig, R. A novel method for modelling interaction between categorical variables. *Int. J. Public Health* **2017**, *62*, 427–431. [CrossRef] [PubMed]
21. Vogler, N.; Bocklitz, T.; Salah, F.S.; Schmidt, C.; Bräuer, R.; Cui, T.; Mireskandari, M.; Greten, F.R.; Schimidt, M.; Stallmach, A.; et al. Systematic evaluation of the biological variance within the Raman based colorectal tissue diagnostics. *J. Biophotonics* **2016**, *9*, 533–541. [CrossRef] [PubMed]
22. Zwanenburg, G.; Hoefsloot, H.C.J.; Westerhuis, J.A.; Jansen, J.J.; Smilde, A.K. ANOVA–principal component analysis and ANOVA–simultaneous component analysis: A comparison. *J. Chemom.* **2011**, *25*, 561–567. [CrossRef]
23. Smilde, A.K.; Hoefsloot, H.C.J.; Westerhuis, J.A. The geometry of ASCA. *J. Chemom.* **2008**, *22*, 464–471. [CrossRef]
24. Neter, J.; Wasserman, W.; Kutner, M. *Applied Linear Statistical Models*, 4th ed.; Irwin: Chicago, IL, USA, 1996.
25. Shaw, R.G.; Mitchell-Olds, T. Anova for Unbalanced Data: An Overview. *Ecology* **1993**, *74*, 1638–1645. [CrossRef]
26. Langsrud, Ø. ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Stat. Comput.* **2003**, *13*, 163–167. [CrossRef]

III. Predictive modeling of antibiotic susceptibility in *E. coli* strains based on the U-Net network and one-class classification

The Reprinted [N. Ali, J. Kirchhoff, P. I. Onoja, A. Tannert, U. Neugebauer, J. Popp, T. Bocklitz: Predictive modeling of antibiotic susceptibility in *E. coli* strains based on the U-Net network and one-class classification, IEEE Access, 2020, 8, 167711-167720] is licensed under a Creative Commons Attribution-NonCommercial NoDerivatives 4.0 International License.

Erklärungen zu den Eigenanteilen des Promovenden sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation.

N. Ali, J. Kirchhoff, P. I. Onoja, A. Tannert, U. Neugebauer, J. Popp, T. Bocklitz. *Predictive modeling of antibiotic susceptibility in E. coli strains based on the U-Net network and one-class classification*, IEEE Access, 2020, 8, 167711-167720							
Beteiligt an (Zutreffendes ankreuzen)							
	1	2	3	4	5	6	7
Konzeption des Forschungsansatzes	×				×	×	×
Planung der Untersuchungen	×	×	×	×	×	×	×
Datenerhebung		×	×	×			
Datenanalyse und -interpretation	×						×
Schreiben des Manuskripts	×	×	×	×	×	×	×
Vorschlag Anrechnung Publikationsäquivalente	1.0						

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 21, 2020, accepted August 29, 2020, date of publication September 8, 2020, date of current version September 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3022829

Predictive Modeling of Antibiotic Susceptibility in *E. Coli* Strains Using the U-Net Network and One-Class Classification

NAIRVEEN ALI[1,2], **JOHANNA KIRCHHOFF**[1,2,3,4], **PATRICK IGOCHE ONOJA**[1,2],
ASTRID TANNERT[2,3], **UTE NEUGEBAUER**[1,2,3,4], **JÜRGEN POPP** [1,2,3,4],
AND THOMAS BOCKLITZ[1,2]
[1]Institute of Physical Chemistry (IPC), Friedrich Schiller University Jena, 07743 Jena, Germany
[2]Leibniz Institute of Photonic Technology (Leibniz-IPHT), 07745 Jena, Germany
[3]Center for Sepsis Control and Care (CSCC), Jena University Hospital, 07747 Jena, Germany
[4]InfectoGnostics Forschungscampus Jena, 07743 Jena, Germany

Corresponding author: Thomas Bocklitz (thomas.bocklitz@uni-jena.de)

This work was supported in part by the Bundesministerium für Bildung und Forschung through the Project Uro-MDD and the Center for Sepsis Control and Care (CSCC, Integrated Research and Treatment Center) under Grant FKZ 03ZZ0444J and Grant FKZ 01EO1502, in part by the Deutsche Forschungsgemeinschaft through the Core Facility Jena Biophotonic and Imaging Laboratory (JBIL) under Project BO4700/4-1, in part by the InfectoGnostics Research Campus under Grant FKZ 13GW0096F, in part by the Leibniz ScienceCampus InfectoOptics which was financed by the funding line Strategic Networking of the Leibniz Association, in part by the Ministry for Economics, Sciences and Digital Society of Thuringia (TMWWDG) through the framework of the Landesprogramm ProDigital under Grant DigLeben-5575/10-9, and in part by the Free State of Thuringia which was co-financed by the European Union funds within the framework of the European Social Fund (ESF) under Grant 2019 FGR 0083. The publication of this article was funded by the Open Access Fund of the Leibniz Association.

ABSTRACT The antibiotic resistance of bacterial pathogens has become one of the most serious global health issues due to misusing and overusing of antibiotics. Recently, different technologies were developed to determine bacteria susceptibility towards antibiotics; however, each of these technologies has its advantages and limitations in clinical applications. In this contribution, we aim to assess and automate the detection of bacterial susceptibilities towards three antibiotics; *i.e.* ciprofloxacin, cefotaxime and piperacillin using a combination of image processing and machine learning algorithms. Therein, microscopic images were collected from different *E. coli* strains, then the convolutional neural network U-Net was implemented to segment the areas showing bacteria. Subsequently, the encoder part of the trained U-Net was utilized as a feature extractor, and the U-Net bottleneck features were utilized to predict the antibiotic susceptibility of *E. coli* strains using a one-class support vector machine (OCSVM). This one-class model was always trained on images of untreated controls of each bacterial strain while the image labels of treated bacteria were predicted as control or non-control images. If an image of treated bacteria is predicted as control, we assume that these bacteria resist this antibiotic. In contrast, the sensitive bacteria show different morphology of the control bacteria; therefore, images collected from these treated bacteria are expected to be classified as non-control. Our results showed 83% area under the receiver operating characteristic (ROC) curve when OCSVM models were built using the U-Net bottleneck features of control bacteria images only. Additionally, the mean sensitivities of these one-class models are 91.67% and 86.61% for cefotaxime and piperacillin; respectively. The mean sensitivity for the prediction of ciprofloxacin is only 59.72% as the bacteria morphology was not fully detected by the proposed method.

INDEX TERMS Antibiotic resistance, *E. coli* strains, U-Net convolutional neural network, one-class SVM.

I. INTRODUCTION

Escherichia coli (*E. coli*) is a large and diverse bacterial species that can be found almost everywhere. This bacterial species shows a high degree of biological variance, where many of *E. coli* strains are essential in the digestive tract while other strains exhibit pathogenic properties and can cause many complications in the urinary tract or in the intestinal tract. On order to cure such infections, antibiotics are utilized. Their selection is becoming increasingly complicated due

---

The associate editor coordinating the review of this manuscript and approving it for publication was Haluk Eren.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

to the overuse and misuse of these drugs yielding resistant bacteria [1]. The extensive and often unnecessary application of antibiotics both in health care as well as in agriculture increases the evolutionary pressure on these bacteria and leads to the development of new mechanisms to resist the existing antibiotics, and subsequently the antibiotics lose their ability to treat bacterial infections [2]. Consequently, the impact of antibiotic resistance is increasing dangerously to extreme levels all over the world.

To select an effective antibiotic for treating severe infections, the determination of the susceptibility profile of the causing pathogen is required. This can be achieved via antibiotic susceptibility testing (AST) which should in an ideal case be rapid, accurate and quantitative. In this context, most AST in clinical praxis relies on culturing the pathogen in the presence of antibiotics and therefore are slow, demanding an initial therapy of a patient with broad-spectrum (and sometimes ineffective) drugs, which might later be changed to a narrow spectrum antibiotic featuring the appropriate mechanism of action to cover the bacterial sensitivity profile. Traditionally, AST was performed by disk diffusion (Kirby-Bauer) methods, where the size of the growth-free zone determines the susceptibility reaction of bacterial pathogens towards a particular antibiotic [3]. Later studies recommended determining the minimal inhibitory concentration (MIC) of an antimicrobial drug. This MIC offers a precise determination of the lowest concentration (in μg/mL) of a drug that inhibits visible growth of bacteria. The classical method to identify the MIC of a specific antibiotic is still the broth micro dilution (BMD) test. Thereby, a defined volume of liquid medium is mixed with a defined concentration of the antibiotic drug and incubated for 16 to 20 h with the bacteria. Then, the MIC is read as the lowest concentration that prevents the visible bacterial growth [4]. Recently, many novel techniques for fast estimation and prediction of antibiotic susceptibility have arisen. These are mainly so-called genotypic methods including polymerase chain reaction (PCR)-based techniques [5] matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) [6] and whole-genome sequencing [7]. Here, the existence of genes or gene products that induce resistance against certain antibiotics is detected, requiring knowledge of the resistance mechanism and the underlying gene product. Though these genotypic methods are quite fast, not all resistances will be detected, especially when they are caused by new spontaneous mutations. Innovative approaches to accelerate phenotypic AST rely on a reduced culturing period in the presence of antibiotics and a subsequent appropriate readout of phenotypic changes caused by antibiotics to susceptible bacteria. These approaches often use microfluidics [8] or microarrays [9] in addition to more sensitive detection methods like Raman spectroscopy [4], [10], [11] or real-time imaging of single cells, where in addition to the detection of the cell count, often an altered cellular morphology upon interaction with antibiotics can be detected in sensitive strains [12]. A number of morphological changes induced by antibiotics in sensitive strains have been described including -among others- filamentation, spheroplast formation, ovoid cell formation, swelling of cells and blebbing (see [13] for a review on this topic). Filamentation can be caused by several mechanisms including an inhibition of DNA synthesis, of protein synthesis and an inhibited peptidoglycan synthesis. The latter can further lead to spheroplast formation or cell lysis [13]. Each of the previously described methods for antimicrobial susceptibility detection feature its advantages and limitations regarding the type of resistance, costs and time requirements to analyze.

Nowadays, machine learning (ML) algorithms are widely implemented in several biomedical studies including the detection of the antibiotic susceptibility of bacteria. Therein, ML algorithms are designed to automate the resistance analysis for a certain AST. In this context, many applications were established to predict antimicrobial MICs [14] or to identify the bacterial resistance towards a specific antibiotic [15], [16] based on whole genome sequence (WGS) data. Also, image-based identification was often utilized to detect the morphological changes in treated bacteria using ML algorithms [12], [17], [18]. Likewise, ML approaches showed quite promising results in automating bacteria susceptibility detection based on their Raman spectra [4], [19].

In this contribution, we present an image-based approach to identify the susceptibility of *E. coli* strains with different susceptibility patterns towards the following antibiotics: ciprofloxacin, cefotaxime, piperacillin (see figure S1). Hereby, microscopic images of one *E. coli* laboratory strain and 11 clinical *E. coli* isolates were acquired, where a part was untreated and used as control bacteria while other parts were treated with different antibiotics for a short period of time (90 min). Then a combination of image processing and ML algorithms were applied to detect the morphological changes caused by these antibiotics. In our analysis, an anomaly detection approach was implemented to find the morphological changes in treated bacteria based on their images. In terms of machine learning, the task is to detect anomalous objects of a certain class, which can be performed by a one-class classifier after training it on normal objects of the same class. Using the previous property, we could train a one-class support vector machine (OCSVM) model on images of only untreated bacteria, which were utilized as control. Then image labels of treated bacteria with antibiotics were predicted as control or non-control. The detection results of *E. coli* susceptibility were presented for two types of image features and for two training methods to construct OCSVM models.

II. SAMPLE PREPARATION AND COLLECTION

Bacteria were obtained from the strain collection of the Institute of Medical Microbiology at the Jena University Hospital. AG100 is a laboratory strain derived from *E. coli* K12 and the other strains (*E. coli* 407, *E. coli* 416, *E. coli* 422, *E. coli* 455, *E. coli* 500, *E. coli* 544, *E. coli* 545, *E. coli* 554, *E. coli* 579, *E. coli* 673, *E. coli* 683) are clinical isolates from sepsis

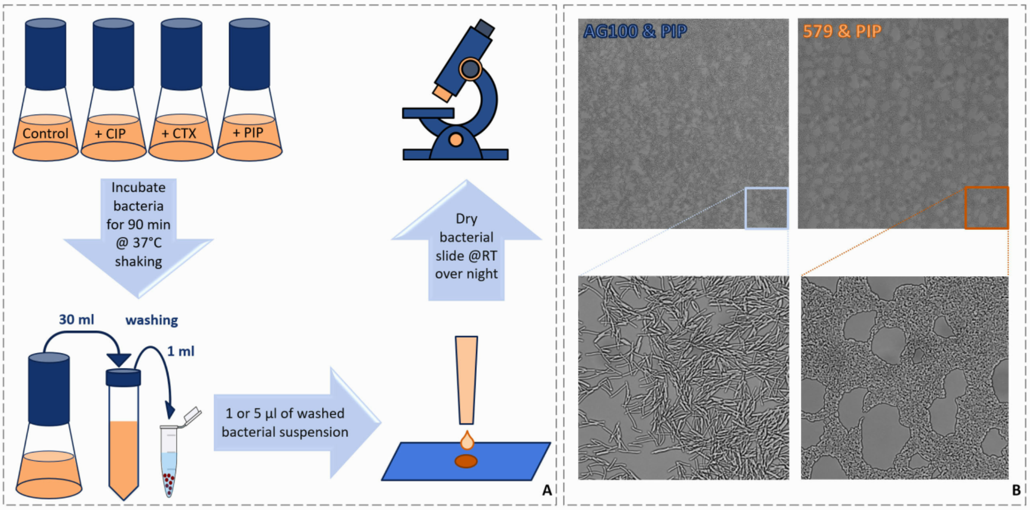

Control
+ CIP
+ CTX
+ PIP
Incubate bacteria for 90 min @ 37°C shaking
30 ml
washing
1 ml
1 or 5 µl of washed bacterial suspension
Dry bacterial slide @RT over night
A
AG100 & PIP
579 & PIP
B

FIGURE 1. **Sample preparation and image collection methods. (A) The bacteria are inoculated with three antibiotics; namely ciprofloxacin, cefotaxime and piperacillin. After inoculating, the bacteria were incubated for 90 min at 37°C, then the bacteria are washed twice in water to remove the antibiotic. Five μl of these washed bacteria are pipetted onto a slide and left to dry at room temperature (RT) over night. Finally, images of these dried bacteria are collected using a bright field microscope. (B) An example of piperacillin (PIP) interaction with *E. coli* bacteria of strains *E. coli* AG100 and *E. coli* 579. Treating bacteria of strain *E. coli* AG100 with PIP prevents bacterial growth and causes the observed morphological changes. Treating bacteria of strain *E. coli* 579 with the same antibiotic does not affect the bacterial growth.**

patients at Jena University Hospital. Quantitative MIC values were determined using VITEK-2 system (bioMeìrieux) or E-Test (Liofilchem MIC Test stripes) and susceptibility categorization in sensitive (S) or resistant (R) is based on the EUCAST clinical breakpoints [20]. The upper EUCAST clinical breakpoint (R>) which categorizes resistance if the corresponding MIC is higher, was selected as test concentration (see table S1). More detailed information on the strains, their reference MIC values, and categorization are given in Kirchhoff *et al.* [4].

For each experiment, a fresh overnight culture was prepared from a −80°C bacterial cryo stock. Four culture flasks were prepared with 30 ml CASO broth (Roth GmbH); in three flasks antibiotic was added to give a final antibiotic concentration of 0.5 mg/l ciprofloxacin (ciprofloxacin hydrochloride, AppliChem), 2 mg/l cefotaxime (cefotaxime sodium, Sigma-Aldrich) or 16 mg/l piperacillin (piperacillin sodium, Sigma-Aldrich); respectively. The fourth flask served as a control without antibiotic treatment. Flasks were pre-warmed until inoculation. The overnight cultures were diluted for measuring the optical density with a spectrophotometer (Spark, Tecan) and inoculated into the pre-warmed flasks to adjust a final inoculum of 5×10^5 bacteria/ml. The cultures with and without antibiotic treatment were incubated for 90 min at 37°C while shaking at 160 rpm in an incubator (Infors HT Ecotron). After 90 min the bacterial suspensions were transferred into a tube and centrifuged for 5 min with a relative centrifugal force of 4,000 g (Universal 320R, Hettich). The bacterial pellet was re-suspended in 1 mL deionized water and washed twice by centrifuging them for 1.5 min with a relative centrifugal force of 11,500 g (MiniSpin®, Eppendorf AG). Finally, the washed pellet was suspended in 20 μl of deionized water. 1 μl and 5 μl of this suspension were pipetted onto a glass slide and allowed to dry at room temperature until the microscopic analysis. Microscopic images were acquired within 5 days after sample preparation. For each sample, a tile scan of 5 × 5 bright field images was recorded using an Axiobserver.Z1 (Carl Zeiss AG, Oberkochen, Germany) equipped with an LD Plan Neofluar 63x/0.75 Korr objective (Zeiss) and an Orca Flash 4.0 camera (Hamamatsu). The total imaged area per sample was 972 × 972 μm. On order to compensate for focal variations within the sample, 5 different focal planes with a distance of 1 μm were collected. Overall, the collected number of replicates for the strains *E. coli* 579, *E. coli* AG100 and *E. coli* 673 is four, three and two independent biological replicates; respectively, while the remaining strains were cultivated in a single biological replicate.

In this experiment, the considered centrifugation protocol is a standard technique in microbiology, and it was applied on order to concentrate the samples and wash the bacteria. These centrifugation protocols are well established and applied in numerous studies [4], [21], [22]. Nevertheless, viable bacteria have been obtained after centrifugation without any observed alteration in the bacteria behavior or in the cell morphology if

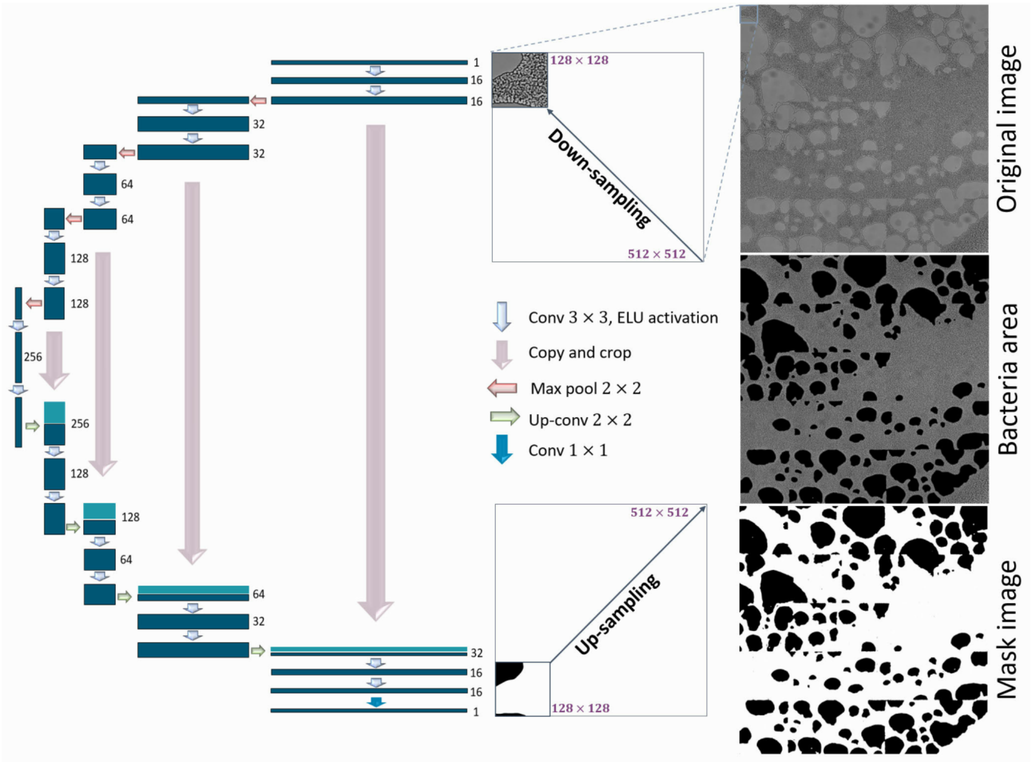


FIGURE 2. A schematic diagram of the segmentation process of the bacteria area using the U-Net network. Each bacteria image is enhanced and sliced into patches of the size 512 × 512 pixels, which are down-sampled and fed into the U-Net network. The up-sampled binary patches are stitched together to create a mask, that can be overlaid with the enhanced image on order to get the segmented image.

they were compared to samples without prior centrifugation steps (see [23]).

III. IMAGE PROCESSING AND MACHINE LEARNING

A. COMPUTATIONAL ANALYSIS

All computations were carried out based on in-house written functions in the programming language Python version 3.6.5 and the statistical programing language R version 3.4.2. The utilized packages are Scikit-learn 0.22 [24], Numpy 1.17.4 [25], OpenCV 4.1.3 [26], Pandas 0.25.3 [27], TensorFlow 2.00, Imager 0.41.2 [28] and Radiomics 0.1.3 [29]. All these functions are available upon request.

B. SEGMENTATION OF BACTERIA AREA

On order to improve the prediction of the antibiotic susceptibility and to exclude artefacts due to the drying process, only image areas with a high bacterial density were included in the analysis. Therefore, each image was segmented into a region with a high density of bacteria and a background region based on the convolutional neural network U-Net [30]. This network showed exceptional performances in semantic segmentation tasks in biomedical applications. The utilized U-Net network consists of an encoder and a decoder with four blocks in each (see Figure 2). The encoder blocks are composed of two convolutional layers, a dropout layer and a max pooling layer, while each decoder block contains an up-convolutional layer, a concatenation layer, two convolutional layers and a dropout layer. The input of the first layer of the encoder is a grayscale-image of the size 128 × 128 pixels and the output of the decoder is a binary image of the same size.

In our work, the collected bacterial images were resized into 9216 × 9216 pixels, and the image contrast was adjusted based on the contrast limited adaptive histogram equalization algorithm (CLAHE) [31]. Thereafter, the enhanced images were sliced into patches of the size 512× ×512 pixels, and the bacterial areas of the obtained patches were predicted using the presented U-Net network. This area predication was performed for all image patches, whether they were acquired from untreated control bacteria or treated bacteria, after down-sampling the patches with a factor of four. The obtained binary patches from the U-Net were up-sampled again by

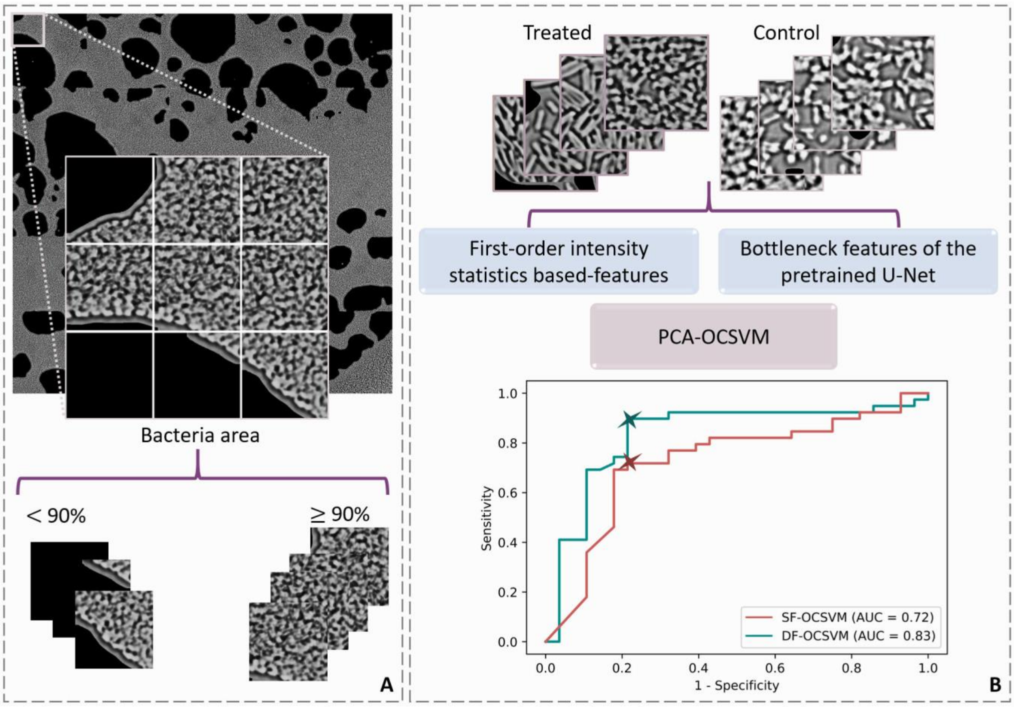


FIGURE 3. Overview of the considered patch selection method and the utilized machine learning techniques. (A) Only image patches that have 90% of their area covered by bacteria are selected to predict the antibiotic susceptibility. (B) Two types of features (SF, DF) are extracted from all selected image patches. Then the predication of the antibiotic susceptibility is performed using a once-class SVM model constructed based on features of the control patches. The obtained classifier is utilized afterwards to predict patch labels of treated bacteria.

a factor of four, and the up-sampled patches were stitched together to reconstruct a binary image with the original size of 9216 × 9216 pixels. This binary image separates the enhanced image into two regions; *e.g.* a bacteria containing area and a background region. Nevertheless, the training procedure of the presented U-Net network was accomplished based only on images of the strains *E. coli* AG100 and *E. coli* 579 while the bacteria area of the remaining image was not used for training the U-Net network. The selection of training set was done due to a pre-experiment, in which the antibiotic susceptibilities of stains *E. coli* AG100 and *E. coli* 579 were checked. In this per-experiment, the enhanced images of both strains were manually converted into binary images using the Java-based image processing program ImageJ [32]. Thereafter, these enhanced images and binary images were portioned with the ratio 2:1 into a training set and a validation set; respectively. Lastly, the U-Net network was trained based on the binary and the enhanced images for 50 epochs using a mini-batch of 50 patches and the Adam optimizer with a learning rate of 0.001 to minimize the binary-cross entropy loss function on the validation set. After training the U-Net, the best model was saved and utilized to segment the bacteria containing area of all remaining images. Here, the default value of learning rate for Adam optimizer was considered while the other hyperparameters; *i.e.* batch size and number of epochs, were manually selected due to the complexity of the presented segmentation task.

C. PATCH SELECTION AND FEATURE EXTRACTION

The segmented images based on the U-Net network were cut into patches of the size 256 × 256 pixels. Then, image patches that have at least 90% of their area covered by bacteria were selected. The previous selection of the bacteria threshold was considered to ensure approximately the same foreground areas in all selected patches (see Figure 3-A). Thereafter, the texture of the selected patches was quantified based on two types of image features. These image features are the first-order statistics-based features (SFs) of the intensity and the bottleneck features of the trained U-Net network. The latter features are indicated later as DFs. The SFs characterize texture properties of the area of interest of an image, and they measure the spatial distribution of intensity values for image pixels [33], [34]. In our work, the energy, entropy, skewness, uniformity, kurtosis, variance, mean deviation, root mean square, mean, median, minimum and maximum were calculated for each selected patch. In table S2, the utilized statistical features were presented. Here, each feature describes a specific property of the gray level distribution of a selected patch *I(x, y)* that has the size 256 × 256 pixels [34]. The other type of features; *i.e.* the DFs, can be simply extracted from the trained U-Net model after removing the decoder layers. The encoder in this case represents an image feature extractor where 256 features per patch can be extracted as it is shown in Figure 2.

D. MACHINE LEARNING FOR SUSCEPTIBILITY DETECTION

Based on the extracted features, the anomaly detection was performed to identify the susceptibility of *E. coli* strains towards the considered antibiotics. This anomaly detection is usually implemented to identify anomalous objects of a specific class [35]. The basic idea is to let a classification model learn on an available dataset in which all objects belong to a same class. Then, this learnt model is utilized to identify normal and anomalous objects of a new dataset with respect to that considered class.

For the presented study, the images of untreated control bacteria were always considered as normal objects while the treated bacterial images were predicted as normal or anomalous patches. This prediction was accomplished by comparing the intrinsic and control-specific morphology of bacterial strains with the morphological changes caused by antibiotics. So, if a particular antibiotic affects the cultivated bacteria, it changes their morphology and let these bacteria look anomalous as compared to untreated control ones (see Figure S2-A). In contrast, when the bacteria resist an antibiotic, they keep growing as untreated bacteria doing (see Figure S2-B) [36]. Under this assumption, an OCSVM model is ideal to detect the sensitivity of treated bacteria to an antibiotic drug. This detection was performed based on a principal component analysis (PCA) based dimension reduction of the feature matrix. The PCA space is formed by new uncorrelated features, i.e. the principal components, which maximize the data variance and often increase the interpretability. Then an OCSVM model was constructed using principal components (PCs) that include 99% of the variation within the untreated control bacteria patches. The obtained classifier was finally utilized to predict labels of treated bacteria as normal (control) or anomalous (non-control).

IV. RESULTS

The susceptibility identification of *E. coli* strains towards the considered antibiotics was accomplished based on a dataset comprising microscopic images collected from 12 *E. coli* strains. Within this dataset, the strains *E. coli* 579, *E. coli* AG100 and *E. coli* 673 were cultivated in four, three and two independent biological replicates; respectively, while the other *E. coli* strains were grown in a single replicate. From each replicate, images of control and treated bacteria were collected using a bright field microscope. After data acquisition, the described image processing pipeline, a patch extraction and a patch selection were applied. The selected image patches were afterwards utilized to identify the antibiotic resistance based on an OCSVM model that was trained on features extracted either from the first-order intensity statistics or from the trained U-Net network, as it is described in Figure S3.

A. THE IDENTIFICATION OF ANTIBIOTIC RESISTANCE IN E. COLI STRAINS

We present in this subsection the obtained results for predicting the antibiotic susceptibility of *E. coli* strains within each biological replicate. In Figure 3-B, a schematic view of the utilized features extraction methods and machine learning techniques is presented. For each biological replicate, the SFs and DFs were extracted, then a feature mean centering was applied with respect to the features of the control patch of each replicate. Later, two PCA models were constructed based on the extracted features from the selected untreated control patches; *i.e.* the PCA model based on the SFs and the PCA model based on DFs. Using the PCs that describes 99% of the variation within the control patch features, two OCSVM models were built. These OCSVM models represent the OCSVM based on the statistical features (SFs) named SF-OCSVM and the OCSVM based on the bottleneck features of the trained U-Net network (DFs) termed DF-OCSVM. For both models, a radial kernel was optimized for the regularization parameter $\vartheta \epsilon$ {0.001, 0.01, 0.1, 0.25, 0.50, 0.75, 0.90, 0.99} and the kernel coefficient $\gamma \epsilon$ {0.001, 0.01, 0.1, 0.25, 0.50, 0.75, 0.90, 1}. This hyperparameters optimization was accomplished via a grid search using the previous noted values of ϑ and γ. The hyperparameter values, that performed the best identification results, were selected to construct a final OCSVM model. This model was used later to predict patch labels of treated bacteria cultivated within the same replicate. As we mentioned earlier, if a specific bacterial pathogen resists an antibiotic, image patches of this pathogen are predicted as control. In contrast, if an antibiotic prevents the growth of a specific bacterial pathogen, it can change the bacteria's morphology. Therefore, the images of bacteria sensitive to this antibiotic are expected to be identified as non-control; *e.g.* dissimilar to bacteria grown without antibiotics that represent here the control bacteria. In the latest case, the untreated control and treated bacteria were cultivated in same experiment hence we are sure that any changes in the bacteria's morphology were caused only by the antibiotics. Based on these assumptions, we predicted labels for all selected patches, and we calculated the percentage of patches predicted as control for each treatment within each replicate. This percentage was denoted in the following by *CP*.

After calculating the *CP* values for bacterial images, the prediction performance of the SF-OCSVM and DF-OCSVM models was evaluated and compared using receiver operating characteristic (ROC) curves. In Figure 3-B, the ROC curves of SF-OCSVM and DF-OCSVM models were depicted. It is clear that the OCSVM model trained on the bottleneck features of U-Net network shows larger area under the curve (AUC) than the AUC of the ROC curve obtained by SF-OCSVM models. Here, the AUC of SF-OCSVM and DF-OCSVM is 72% and 83%; respectively. In Table S1, the susceptibility predictions of bacterial image slides are presented based on two indicated thresholds of the ROC curves in Figure 3-B. These thresholds are 78.46% and 99.07%, and they are corresponding to the highest sensitivity and specificity introduced by classification models. One can note that both selected thresholds describe high values of the ROC curves, which can be interpreted that treated bacteria are predicted as control if a large percentage of image patches captured from these bacteria were classified as control patches; *i.e.* high percentages of *CP*. Nevertheless, within Table S1, antibiotic MIC values are presented beside the predictions of the *E. coli* susceptibility of both SF-OCSVM and DF-OCSVM models. Also, the antibiotic breakpoints and the reference antibiotic susceptibilities are shown with respect to each *E. coli* strain. It is observed that the OCSVM based on DFs could predict the susceptibility of *E. coli* strains toward piperacillin and cefotaxime quite well in comparison to predictions provided by the SF-OCSVM models. In contrast, neither the SF-OCSVM model nor the DF-OCSVM model could predict the susceptibility of ciprofloxacin in good manner based on the selected thresholds. In Table 1, a summary of the predicted susceptibility is presented as confusion matrices with respect to each antibiotic and each OCSVM model. For the susceptibility predictions of piperacillin, the mean sensitivity of OCSVM model increased from 41.07% to 86.61% when this classifier was trained on the bottleneck features of the U-Net network instead of using the SFs. Also, the mean sensitivity of cefotaxime improved around 4% when the DF-OCSVM was considered as the mean sensitivity of SF-OCSVM and DF-OCSVM are 87.5% and 91.67%; respectively. The mean sensitivities of SF-OCSVM and DF-OCSVM models for ciprofloxacin are only 70.14% and 59.72%. These results exhibit that the changes in bacteria morphology is not sufficient to predict the resistance towards ciprofloxacin.

TABLE 1. **The confusion matrices using local OCSVM models. For each antibiotic, the reference susceptibility, and the predicted susceptibilities (S: sensitive, R: resistant) based on SF-OCSVM and DF-OCSVM models are presented, then the mean sensitivities of each antibiotic and each classifier are calculated.**

Antibiotic		SF-OCSVM		M. Sens.	SD-OCSVM		M. Sens.
		R	S		R	S	
PIP	R	2	6	**41.07%**	7	1	**86.61%**
	S	3	4		1	6	
CTX	R	5	1	**87.5%**	5	1	**91.67%**
	S	1	11		0	12	
CIP	R	5	3	**70.14%**	6	2	**59.72%**
	S	2	7		5	4	

Overall, the DF-OCSVM models have better identification performance than the obtained predictions using the SF-OCSVM. These prediction results are strongly influenced by two main factors: the selected threshold by the ROC curve and the changes in bacteria shape when a particular antibiotic was applied. In case of ciprofloxacin, quite small changes in bacteria morphology were detected after incubating the bacteria while the selected threshold seems to be not suitable.

B. STUDYING THE PREDICTION PERFORMANCE OF ANTIBIOTIC SUSCEPTIBILITY BASED ON LOCAL-TRAINED/GLOBAL-TRAINED OCSVM MODEL

The main goal of the following study is to check the prediction quality of the OCSVM models based on two training techniques: the local-training and the global-training. Here, we denote by local-trained OCSVM an OCSVM model that is trained and tested on one individual replicate, while global-trained models describe the OCSVM models that are trained on a larger number of replicates and tested on other independent replicates. For local-trained models, the control bacteria images from a specific replicate are utilized to train an OCSVM model. This model is implemented to predict the resistance of treated bacteria cultivated in the same replicate but were not used for model training. A global-trained OCSVM model is built upon control images of a number of replicates, then this classifier can be utilized to predict antibiotic susceptibilities of bacteria images acquired from other replicates. In both cases, the prediction of newly acquired test data is possible and linked to the estimated accuracy. To perform such comparison, the different replicates of strain *E. coli* 579 and strain *E. coli* AG100 were considered. Therein, the bottleneck features of the trained U-Net network were extracted for all selected patches of both *E. coli* strains, and a feature mean centering was applied as was explained previously. Finally, a leave-one-replicate-out cross-validation (LORO-CV) was performed based on the PCs extracted from control patches. For model construction based on LORO-CV, we always exclude one replicate and optimize a radial kernel for the regularization parameter ϑ and the kernel coefficient γ using patches extracted from the remaining replicates. This procedure is iterated until the susceptibility of all patches selected from all replicates are identified once.

Based on the ROC curves, a comparison between the performance of local-trained OCSVM and global-trained OCSVM was performed. First of all, we calculated the *CP* values for each treatment and for all replicates. Figure 4-A presents the ROC curves of the OCSVM models using both training methods. Our results showed that the susceptibility prediction using a local-trained OCSVM model is much better than the predictions by global-trained models. Thereby, around 31% increase in the AUC can be observed by local-trained OCSVM models. In Table 2, the identification results of *E. coli* susceptibility using local-trained and global-trained models are presented based on selected thresholds of ROC curves in Figure 4-A. It turned out that a local-training for the

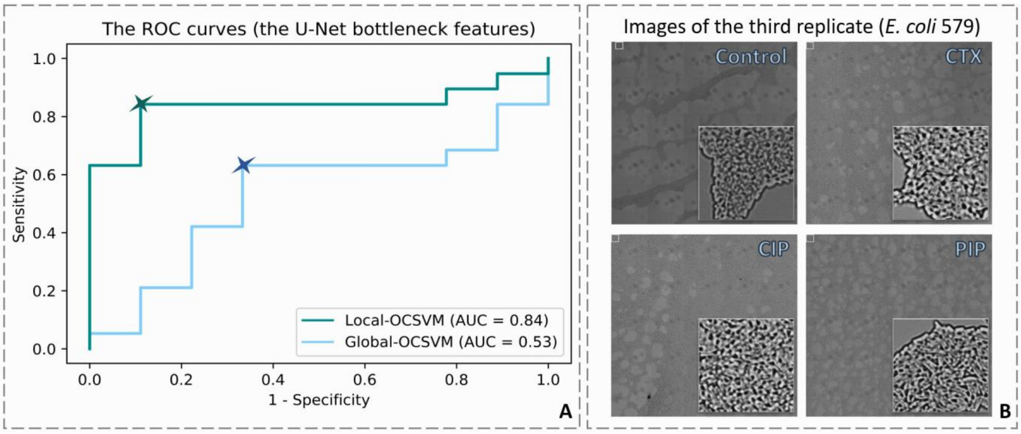


FIGURE 4. **(A) The obtained ROC curves of local-OCSVM and global-OCSVM models. These models were constructed based on the bottleneck features of the trained U-Net network. The antibiotic susceptibilities were determined for the percentage of predicted patches as control (*CP*) (B) Images collected from the third replicate of strain *E. coli* 579. The reference susceptibility of *E. coli* 579 is resistant, but the image patches of the treated bacteria are obviously different to the control image patches of the strain *E. coli* 579.**

TABLE 2. **A comparison of local-OCSVM and global-OCSVM models. The predicted susceptibilities (S: sensitive, R: resistant) were determined based on selected thresholds. It turns out that the identification of antibiotic resistance using local-OCSVM models is much better than the predictions by global-OCSVM models.**

Antibiotic & strain		Replicate	Local-OCSVM S < 80.77% & R≥ 80.77%		Global-OCSVM S < 55.20% & R ≥ 55.20%	
			CP(%)	Pred.	*CP(%)*	Pred.
Piperacillin	579 (R)	1	86.64	R	55.20	R
		2	85.76	R	65.33	R
		3	19.53	S	11.39	S
		4	80.96	R	2.57	S
	AG100 (S)	1	12.35	S	78.3	R
		2	33.39	S	45.06	S
		3	8.15	S	4.01	S
Cefotaxime	579 (R)	1	93.33	R	86.87	R
		2	88.77	R	74.22	R
		3	11.03	S	8.45	S
		4	80.77	R	0.0	S
	AG100 (S)	1	33.51	S	78.33	R
		2	29.58	S	42.72	S
		3	41.01	S	20.59	S
Ciprofloxacin	579 (R)	1	97.50	R	97.5	R
		2	90.96	R	92.4	R
		3	7.29	S	4.17	S
		4	82.82	R	0.0	S
	AG100 (S)	1	74.84	S	98.11	R
		2	85.90	R	88.83	R
		3	64.59	S	27.21	S

OCSVM models introduced a better identification of strain sensitivity compared to the identification by global-trained models. In detail, only the treated bacteria of strain *E. coli* 579 in the third replicate, and the bacteria of strain *E. coli* AG100 treated with ciprofloxacin in second replicate were misidentified when a local-trained OCSVM was considered. However, 10 images of 21 images were misclassified when a global-training for OCSVM models was applied.

The results presented above showed that the local-training of OCSVM models provide, in most cases, more stable identification results of the *E. coli* susceptibility towards antibiotics in comparison to global-trained models. These results were expected because of the high biological variations between the replicates which can confuse classifiers in case of global-trained models. Another reason for these results is that some pathogens might change their growing behavior; *e.g.* stop duplicating or interacting differently with a particular antibiotic drug. In our study, the control bacteria cultivated in the third replicate of *E. coli* 579 stopped duplicating while the treated bacteria started elongating (see Figure 4-B). Therefore, the treated patches within this replicate were mostly misclassified and were predicted as sensitive bacteria as compared to untreated control ones, even though the EUCAST clinical breakpoints indicate a resistance.

V. SUMMARY

We presented in this article the results of an image-based identification approach to detect the antibiotic susceptibilities of *E. coli* strains. The chosen antibiotics cause a strong morphological alteration in sensitive strains leading to a cell elongation (filamentation) while resistant strains retain their normal morphological properties upon treatment. In the presented work, different image processing techniques were combined with machine learning algorithms on order to automate the susceptibility detection. We started the analysis by enhancing the image contrast, and we segmented the high-bacterial density areas based on the U-Net network. The segmented images were afterwards cut into patches, and the patches that have at least 90% of their area covered by bacteria were selected for further analyses composed of feature extraction and modeling. In our work, the first-order statistics-based features of the intensity (SFs) and the bottleneck features of the trained U-Net network (DFs) were extracted and used to train a one-class classification model; specifically, an OCSVM model. This type of classification is designed to detect anomalous objects of a particular class after training the model only on normal objects of this considered class.

Based on the described data analysis pipeline, we performed two comparisons to identify the *E. coli* susceptibility using OCSVM models. In the first comparison, the antibiotic sensitivity of each bacterial replicate was predicted using a local-OCSVM model that was built on both types of image features. The second comparison was performed to check the prediction quality of the OCSVM models using two training methods and using the DFs only. The results of the first comparison showed that using the DFs to train local-OCSVM models introduced larger area under the ROC curve than the SF-OCSVM models. Also, for selected thresholds of ROCs, the classification mean sensitivities of piperacillin and cefotaxime increased from 41.07% to 86.61% and from 87.5% to 91.67%; respectively, when OCSVM models were constructed on the bottleneck features instead of using the SFs. In contrast, both classifiers showed low identification results based on the selected thresholds when the bacterial pathogens were treated by ciprofloxacin. To investigate this behavior, the DFs were utilized to perform the second comparison. Therein, two training techniques; namely local-training and global-training, were compared. While a local-OCSVM model was trained and tested on untreated control and treated bacteria patches of the same replicate, different independent replicates were utilized to train and test the global-OCSVM models. The evaluation of these models proved that locally trained one-class models feature a great potential in identifying the antibiotic sensitivity as compared to global-trained OCSVM models.

VI. CONCLUSION

It was shown that the combination of bottleneck features of the trained U-Net and the local trained OCSVM models introduced quite promising results in identifying the susceptibility of *E. coli* strains towards antibiotics. These local models are correcting for the biological variations between different replicates or patients and yielding better predictions of individual patient's susceptibility towards antibiotics. Therefore, the presented local-one-class classification approach can be easily implemented to predict other antibiotic susceptibilities, and an easy image-based antibiotic susceptibility tests (ASTs) can be generated. Since the morphological changes appear already after short incubation times of antibiotics with bacteria (90 min), this image-based method might be used for the development of fast phenotypic AST, maybe in combination with statistical parameters from other readout methods like Raman spectroscopy.

ACKNOWLEDGMENT

Support information: Figure S1, S2, S3 and Table S1, S2.

REFERENCES

[1] (2020). *Antibiotic Resistance*. [Online]. Available: https://www.who.int/topics/antimicrobial_resistance/en/
[2] C. L. Ventola, "The antibiotic resistance crisis: Part 1: Causes and threats. P & T : A peer-reviewed," *J. Formulary Manage.*, vol. 40, no. 4, pp. 277–283, 2015.
[3] C. Giuliano, "A guide to bacterial culture identification and results interpretation. P & T : A Peer-Reviewed," *J. Formulary Manage.*, vol. 44, no. 4, pp. 192–200, 2019.
[4] J. Kirchhoff, U. Glaser, J. A. Bohnert, M. W. Pletz, J. Popp, and U. Neugebauer, "Simple ciprofloxacin resistance test and determination of minimal inhibitory concentration within 2 h using Raman spectroscopy," *Anal. Chem.*, vol. 90, no. 3, pp. 1811–1818, Feb. 2018.
[5] M. R. Pulido, "Progress on the development of rapid methods for antimicrobial susceptibility testing," *J. Antimicrob Chemother*, vol. 68, no. 12, p. 2710, 2013.
[6] V. Belkum, "Matrix-assisted laser desorption ionization time-of-flight mass spectrometry in clinical microbiology: What are the current issues," *Ann. Lab. Med.*, vol. 37, no. 6, pp. 475–483, 2017.
[7] M. J. Ellington, "The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: Report from the EUCAST subcommittee," *Clin. Microbiol. Infection*, vol. 23, no. 1, pp. 2–22, Jan. 2017.
[8] J. Campbell, C. McBeth, M. Kalashnikov, A. K. Boardman, A. Sharon, and A. F. Sauer-Budge, "Microfluidic advances in phenotypic antibiotic susceptibility testing," *Biomed. Microdevices*, vol. 18, no. 6, p. 103, Dec. 2016.
[9] A. Srinivasan, G. C. Lee, N. S. Torres, K. Hernandez, S. D. Dallas, J. Lopez-Ribot, C. R. Frei, and A. K. Ramasubramanian, "High-throughput microarray for antimicrobial susceptibility testing," *Biotechnol. Rep.*, vol. 16, pp. 44–47, Dec. 2017.
[10] K. Chang, "Antibiotic susceptibility test with surface-enhanced Raman scattering in a microfluidic system," *Anal. Chem.*, vol. 91, no. 17, pp. 10988–10995, 2019.
[11] A. Tannert, R. Grohs, J. Popp, and U. Neugebauer, "Phenotypic antibiotic susceptibility testing of pathogenic bacteria using photonic readout methods: Recent achievements and impact," *Appl. Microbiol. Biotechnol.*, vol. 103, no. 2, pp. 549–566, Jan. 2019.
[12] J. Choi, J. Yoo, M. Lee, E.-G. Kim, J. S. Lee, S. Lee, S. Joo, S. H. Song, E.-C. Kim, J. C. Lee, H. C. Kim, Y.-G. Jung, and S. Kwon, "A rapid antimicrobial susceptibility test based on single-cell morphological analysis," *Sci. Transl. Med.*, vol. 6, no. 267, pp. 174–267, Dec. 2014.
[13] T. P. T. Cushnie, N. H. O'Driscoll, and A. J. Lamb, "Morphological and ultrastructural changes in bacterial cells as an indicator of antibacterial mechanism of action," *Cellular Mol. Life Sci.*, vol. 73, no. 23, pp. 4471–4492, Dec. 2016.
[14] M. Nguyen, "Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal salmonella," *J. Clin. Microbiol.*, vol. 67, no. 2, 2019, Art. no. e01260.
[15] E. S. Kavvas, E. Catoiu, N. Mih, J. T. Yurkovich, Y. Seif, N. Dillon, D. Heckmann, A. Anand, L. Yang, V. Nizet, J. M. Monk, and B. O. Palsson, "Machine learning and structural analysis of mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance," *Nature Commun.*, vol. 9, no. 1, p. 4306, Dec. 2018.
[16] M. W. Pesesky, T. Hussain, M. Wallace, S. Patel, S. Andleeb, C.-A.-D. Burnham, and G. Dantas, "Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in gram-negative bacilli from whole genome sequence data," *Frontiers Microbiol.*, vol. 7, p, 5, Nov. 2016.
[17] M. Marschal, "Evaluation of the accelerate pheno system for fast identification and antimicrobial susceptibility testing from positive blood cultures in bloodstream infections caused by gram-negative pathogens," *J. Clin. Microbiol.*, vol. 55, no. 7, p. 2116, 2017.
[18] H. Yu, W. Jing, R. Iriya, Y. Yang, K. Syal, M. Mo, T. E. Grys, S. E. Haydel, S. Wang, and N. Tao, "Phenotypic antimicrobial susceptibility testing with deep learning video microscopy," *Anal. Chem.*, vol. 90, no. 10, pp. 6314–6322, May 2018.
[19] C.-S. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. A. E. Saleh, S. Ermon, and J. Dionne, "Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning," *Nature Commun.*, vol. 10, no. 1, p. 4927, Dec. 2019.

[20] D. Kusic, "Raman spectroscopic characterization of packaged L. Pneumophila strains expelled by T. Thermophila," *Anal. Chem.*, vol. 88, no. 5, pp. 2533–2537, 2016.

[21] U. Schröder, "On-chip spectroscopic assessment of microbial susceptibility to antibiotics within 3.5 hours," *J. Biophoton.*, vol. 10, no. 11, pp. 1547–1557, 2017.

[22] T. Ursell, T. K. Lee, D. Shiomi, H. Shi, C. Tropini, R. D. Monds, A. Colavin, G. Billings, I. Bhaya-Grossman, M. Broxton, B. E. Huang, H. Niki, and K. C. Huang, "Rapid, precise quantification of bacterial cellular dimensions across a genomic-scale knockout library," *BMC Biol.*, vol. 15, no. 1, p. 17, Dec. 2017.

[23] J. Choi, H. Y. Jeong, G. Y. Lee, S. Han, S. Han, B. Jin, T. Lim, S. Kim, D. Y. Kim, H. C. Kim, E.-C. Kim, S. H. Song, T. S. Kim, and S. Kwon, "Direct, rapid antimicrobial susceptibility test from positive blood cultures based on microscopic imaging analysis," *Sci. Rep.*, vol. 7, no. 1, Dec. 2017, Art. no. 1148.

[24] F. Pedregosa, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 1, pp. 2825–2830, Nov. 2011.

[25] T. E. Oliphant, *A Guide to NumPy*, vol. 1. New York, NY, USA: Trelgol Publishing, 2006.

[26] G. Bradski, "The OpenCV library," *Dr. Dobb's J. Softw. Tools*, 2000.

[27] W. McKinney, "Data structures for statistical computing in Python," in *Proc. 9th Python Sci. Conf.*, 2010, pp. 1–5.

[28] S. Barthelmé and D. Tschumperlé, "Imager: An r package for image processing based on CImg," *J. Open Source Softw.*, vol. 4, no. 38, p. 1012, Jun. 2019.

[29] J. Carlson, "Radiomics: 'Radiomic' image processing toolbox," *R Package Version 0.1*, vol. 2, 2016.

[30] O. Ronneberger and P. T. F. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Cham, Switzerland: Springer, 2015.

[31] A. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," *J. VLSI Signal Process. Syst. Signal, Image Video Technol.*, vol. 38, no. 1, pp. 35–44, 2004.

[32] J. Schindelin, "Fiji: An open-source platform for biological-image analysis," *Nat. Methods*, vol. 9, no. 7, pp. 676–682, 2012.

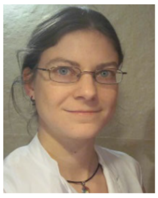
[33] R. M. Haralick and K. I. Shanmugam Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[34] C. Parmar, E. Rios Velazquez, R. Leijenaar, M. Jermoumi, S. Carvalho, R. H. Mak, S. Mitra, B. U. Shankar, R. Kikinis, B. Haibe-Kains, P. Lambin, and H. J. W. L. Aerts, "Robust radiomics feature quantification using semiautomatic volumetric segmentation," *PLoS ONE*, vol. 9, no. 7, Jul. 2014, Art. no. e102107.

[35] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, Oct. 2004.

[36] J. L. Martínez and F. Baquero, "Interactions among strategies associated with bacterial infection: Pathogenicity, epidemicity, and antibiotic resistance," *Clin. Microbiol. Rev.*, vol. 15, no. 4, p. 647, 2002.

NAIRVEEN ALI received the B.Sc. and M.Sc. degrees in mathematical statistics from Damascus University, in 2009 and 2014, respectively. She is currently pursuing the Ph.D. degree with the Research Group of Statistical Modeling and Image Analysis, University of Jena, supervised by PD. Dr. Bocklitz. Her research interests include multivariate data analysis, image processing, and machine learning for biomedical applications.



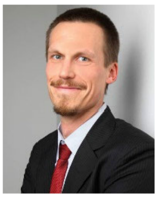
JOHANNA KIRCHHOFF received the Diploma degree in biology and the Ph.D. degree in physical chemistry from Friedrich Schiller University Jena, in 2012 and 2019, respectively. In 2014, she started her Ph.D. research in the group of Prof. Neugebauer with the Center for Sepsis Control and Care (CSCC) and the Leibniz Institute of Photonic Technology (IPHT), Jena. Her Ph.D. project was focused on micro-Raman spectroscopic characterization of the interactions of antibiotics with sepsis pathogens. Her research interests include new optical-spectroscopic procedures for microbiological diagnosis and antimicrobial susceptibility testing strategies.

PATRICK IGOCHE ONOJA received the B.Sc. degree in physics from Benue State University, Nigeria, in 2006, and the M.Sc. degree in biomedical engineering from Heidelberg University, Germany, in 2014. He is currently pursuing the second M.Sc. degree in the research group of Dr. Popp and under the supervision of PD. Dr. Christoph Krafft. From 2015 to 2016, he worked as a Student Research Assistant in the group of Dr. G. Glatting with the Mannheim University Hospital. In 2018, he joined with Dr. Ute Neugebauer, where he studied microscopic characterization of antibiotic bacteria interaction. His research interests include radiation therapy, biomedical imaging, spectroscopy, and microscopy of biological samples.

ASTRID TANNERT received the degree in biochemistry from Martin-Luther-University Halle-Wittenberg, the University of Wales, Cardiff University, and the Free University of Berlin, and the Ph.D. degree in biophysics from the Humboldt University of Berlin, in 2003. She is currently coordinating the Jena Biophotonic and Imaging Laboratory. Her research interest includes microscopic solutions for biomedical research and diagnosis especially in infectious diseases.

UTE NEUGEBAUER studied chemistry in Jena, Germany, and Chapel Hill, NC, USA. After her Ph.D. degree, she joined the Biomedical Diagnostics Institute, Dublin, Ireland. From 2011 to 2016, she was leading the Junior Research Group Spectroscopic Pathogen Detection at the Center for Sepsis Control and Care (CSCC), Jena University Hospital. Since 2016, she has been a Professor with the University of Jena, the Department Leader with the Leibniz Institute of Photonic Technology, Jena, and the Head of the Core Unit Biophotonics, CSCC. Her research interests include novel spectroscopic tools and methods for medical diagnostics and the characterization of physiological interactions with a special focus on infection and sepsis.

JÜRGEN POPP received the degree in chemistry from the Universities of Erlangen and Würzburg, the Ph.D. degree in chemistry, and the Habilitation degree from the University of Würzburg, in 2002. After his Ph.D. degree in chemistry, he joined Yale University for his postdoctoral studies. Since then, he has held a Chair of physical chemistry with Friedrich Schiller University Jena. He has also been the Scientific Director of the Leibniz Institute of Photonic Technology, Jena, since 2006. His research interest includes biophotonics, in particular with the development and application of innovative Raman techniques for biomedical diagnosis.



THOMAS BOCKLITZ received the Diploma degree in theoretical physics, in 2007, the Ph.D. degree in chemometrics, in 2011, and the degree in physics and mathematics from Friedrich Schiller University Jena. He is currently the Junior Research Group Leader of statistical modeling and image analysis with the University of Jena. His research interest includes the translation of physical information, obtained by AFM, TERS, Raman spectroscopy, CARS, SHG, or TPEF, into medically or biologically relevant information.

• • •

IEEE Access
Multidisciplinary ⁞ Rapid Review ⁞ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Supporting Information for:

Predictive Modeling of Antibiotic Susceptibility in *E. coli* Strains Using the U-Net Network and One-Class Classification

Nairveen Ali [1,2], Johanna Kirchhoff [1,2,3,4], Patrick Igoche Onoja [1,2], Astrid Tannert [2,3], Ute Neugebauer [1,2,3,4], Jürgen Popp [1,2,3,4] and Thomas Bocklitz [1,2]

[1]Institute of Physical Chemistry (IPC), Friedrich Schiller University, Jena, Germany
[2]Leibniz Institute of Photonic Technology (Leibniz-IPHT), Member of Leibniz Research alliance "Health technologies", Jena, Germany
[3]Center for Sepsis Control and Care (CSCC), Jena University Hospital, Jena, Germany
[4]InfectoGnostics, Forschungscampus Jena, Jena, Germany

Corresponding author: Thomas Bocklitz (e-mail: Thomas.bocklitz@uni-jena.de).

Financial support by the Bundesministerium für Bildung und Forschung for the project Uro-MDD (FKZ 03ZZ0444J) and via the Integrated Research and Treatment Center "Center for Sepsis Control and Care" (CSCC, FKZ 01EO1502) is highly acknowledged. Funding of the Deutsche Forschungsgemeinschaft via the Core Facility Jena Biophotonic and Imaging Laboratory (JBIL) and for the project BO4700/4-1 is highly appreciated. Furthermore, the work was performed within the Research Campus InfectoGnostics (FKZ 13GW0096F) and the Leibniz ScienceCampus InfectoOptics which is financed by the funding line Strategic Networking of the Leibniz Association. 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) and by the Free State of Thuringia under the number 2019 FGR 0083 which was co-financed by European Union funds within the framework of the European Social Fund (ESF).

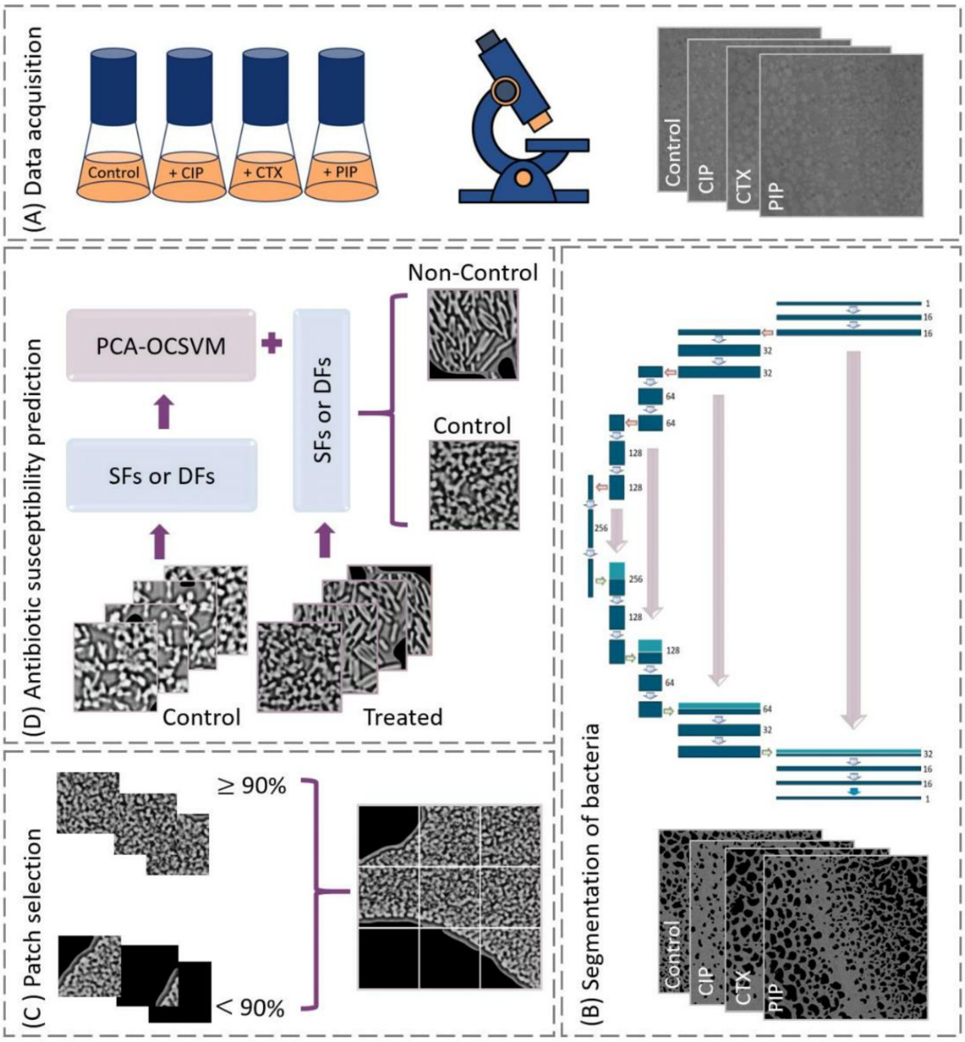


FIGURE S1. A schematic diagram of the proposed image processing and statistical analysis for predicting the antibiotic resistances. (A) Images of cultivated bacteria, untreated control and treated with three antibiotics, are acquired using bright field microscopy. (B) The bacteria images are segmented using the U-Net network into high intensity bacteria regions and background. (C) The segmented bacteria images are sliced into patches of the size 265×256 pixels, and patches that have 90% of their area covered by bacteria are considered further. (D) The selected image patches of control bacteria are utilized to build one-class SVM (OCSVM) models based on two types of features; *i.e.* statistical features and the bottleneck features of the pre-trained U-Net network. The constructed OCSVM models are implemented to predict bacteria susceptibility towards the antibiotics using the extracted features from the selected patches of treated bacteria.

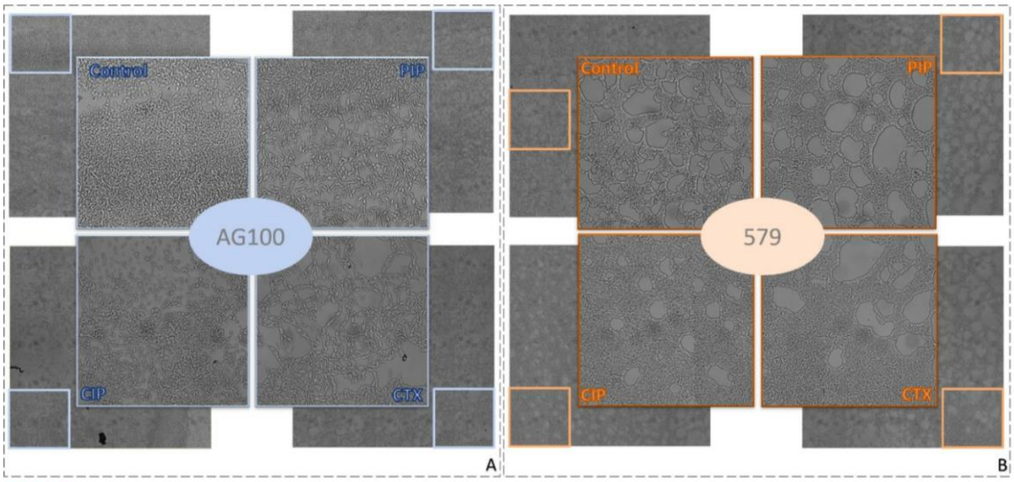


FIGURE S2. Image examples of sensitive and resistance *E. coli* strains. (A) Bacteria of *E. coli* AG100 strain interact sensitively with the selected antibiotics. This interaction introduces specific morphological changes can be seen in this example by the bacterial elongation. (B) The selected antibiotics do not affect *E. coli* 579; therefore, these treated bacteria look similar to the untreated control bacteria.

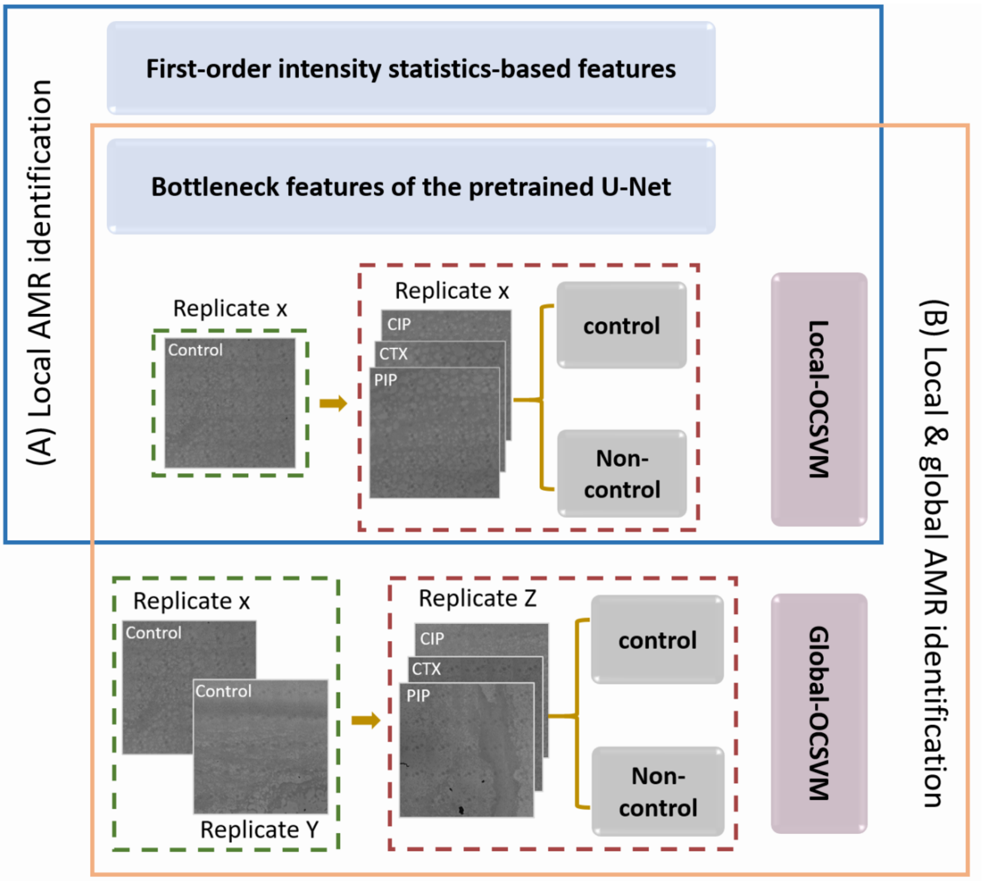

First-order intensity statistics-based features
Bottleneck features of the pretrained U-Net
(A) Local AMR identification
Replicate x
Control
Replicate x
CIP
CTX
PIP
control
Non-control
Local-OCSVM
Replicate x
Control
Control
Replicate Y
Replicate Z
CIP
CTX
PIP
control
Non-control
Global-OCSVM
(B) Local & global AMR identification

FIGURE S3. An overview of the presented analyses and comparisons using OCSVM models. (A) The local-trained models are constructed either based on the statistical features or based on the bottleneck features of the trained U-Net network. Here, OCSVM models are trained on untreated control image of a specific replicate and tested on the treated images collected from the same replicate. (B) A comparison between the identification performance of the local-trained and global-trained OCSVM models using the bottleneck features of the trained U-Net network.

TABLE S1. The prediction results of antibiotic resistance using local-OCSVM models. For esch replicate and for each antibiotic, the predicted susceptibilities (S: sensitive, R: resistant) based on SF-OCSVM and DF-OCSVM models are presented then compared with the reference susceptibility according to the MICs of each *E. coli* strain and the EUCAST breakpoints.

Antibiotic and EUCAST Breakpoint	Strain	Replicate	MIC [mg/l] & Susceptibility Categorization according to EUCAST		SF-OCSVM S < 99.07% & R ≥ 99.07%		DF-OCSVM S < 78.46% & R ≥ 78.46%	
					CP(%)	Pred.	*CP(%)*	Pred.
Piperacillin S ≤ 8 **mg/l** & R > 16 **mg/l**	407	1	≥ 256	**R**	98.91	S	81.52	R
	416	1	≥ 128	**R**	98.19	S	78.46	R
	422	1	≤ 4	**S**	99.88	R	63.01	S
	500	1	≥ 128	**R**	100	R	93.93	R
	544	1	≥ 256	**R**	99.88	R	92.36	R
	579	1	≥ 256	**R**	7.01	S	86.64	R
		2	≥ 256	**R**	61.30	S	85.76	R
		3	≥ 256	**R**	20.25	S	19.53	S
		4	≥ 256	**R**	32.61	S	80.96	R
	673	1	1	**S**	100	R	13.77	S
		2	1	**S**	64.27	S	22.67	S
	683	1	2	**S**	100	R	87.50	R
	AG100	1	4	**S**	40.21	S	12.35	S
		2	4	**S**	65.15	S	33.39	S
		3	4	**S**	88.13	S	8.15	S
Cefotaxime S ≤ 1 **mg/l** & R > 2 **mg/l**	407	1	8	**R**	99.07	R	88.81	R
	416	1	≤ 1	**S**	94.30	S	45.53	S
	422	1	≤ 1	**S**	98.81	S	42.93	S
	455	1	≤ 1	**S**	99.87	R	52.36	S
	500	1	32	**R**	100	R	93.50	R
	544	1	≤ 1	**S**	85.02	S	52.36	S
	545	1	≤ 1	**S**	31.23	S	23.84	S
	554	1	≤ 1	**S**	45.74	S	70.54	S
	579	1	≥ 256	**R**	100	R	93.33	R
		2	≥ 256	**R**	100	R	88.77	R
		3	≥ 256	**R**	6.73	S	11.03	S
		4	≥ 256	**R**	100	R	80.77	R
	673	1	≤ 1	**S**	98.41	S	26.07	S
		2	≤ 1	**S**	31.05	S	11.95	S
	683	1	0.125	**S**	77.98	S	51.81	S
	AG100	1	4	**S**	3.99	S	33.51	S
		2	4	**S**	13.62	S	29.58	S
		3	4	**S**	0	S	41.01	S
Ciprofloxacin S ≤ 0.25 **mg/l** & R > 0.50 **mg/l**	407	1	0.016	**S**	92.66	S	94.92	R
	416	1	1	**R**	93.22	S	91.77	R
	422	1	1	**R**	99.62	R	64.40	S
	455	1	≤ 0.25	**S**	99.10	R	86.83	R
	544	1	≥ 32	**R**	99.88	R	86.83	R
	545	1	0.125	**S**	96.56	S	34.90	S
	554	1	≥ 32	**R**	35.07	S	85.30	R
	579	1	≥ 256	**R**	100	R	97.50	R
		2	≥ 256	**R**	99.87	R	90.96	R
		3	≥ 256	**R**	9.20	S	7.29	S
		4	≥ 256	**R**	99.39	R	82.82	R
	673	1	0. 032	**S**	100	R	100	R
		2	0.032	**S**	15.52	S	94.83	R
	683	1	0.032	**S**	75.63	S	52.94	S
	AG100	1	0.032	**S**	12.58	S	74.84	S
		2	0.032	**S**	98.67	S	85.90	R
		3	0.032	**S**	48.36	S	64.59	S

TABLE S2. The definitions of the first-order intensity statistics-based features. Here, *I(x, y)* denotes to the intensity value of the gray distribution on the position *x, y* of where *x*ϵ {1, 2, …, X} and *y*ϵ {1, 2, …, Y}, *P* is the first order histogram and *P*(*i*) refers to the fraction of intensity value with gray level *i*.

Feature	Formula	Property
Energy	$E = \sum_{y=1}^{Y}\sum_{x=1}^{X} I(x,y)^2$	The value is lowest of coarse texture
Entropy	$H = -\sum_{i=1}^{N_g} P(i)log_2P(i)$	Variability of intensity levels within an image patch
Skewness	$\gamma_2 = \frac{1}{XY}\sum_{y=1}^{Y}\sum_{x=1}^{X}\left[\frac{I(x,y)-\mu}{\sigma}\right]^3$	The asymmetry degree of histogram
Kurtosis	$\gamma_2 = \frac{1}{XY}\sum_{y=1}^{Y}\sum_{x=1}^{X}\left[\frac{I(x,y)-\mu}{\sigma}\right]^4 - 3$	Sharpness of the histogram
Uniformity	$U = \sum_{i=1}^{N_g} P(i)^2$	Maximum for all equal intensities
Variance	$Var = \frac{1}{XY-1}\sum_{y=1}^{Y}\sum_{x=1}^{X}[I(x,y)-\mu]^2$	Average the contrast of an image patch
Mean deviation	$MD = \frac{1}{XY}\sum_{y=1}^{Y}\sum_{x=1}^{X}\lvert I(x,y)-\mu\rvert$	Average of the absolute intensity deviations of an image patch from the intensity mean
Maximum	$I_{max} = \max(I(x,y))$	Maximum intensity level of an image patch
Minimum	$I_{min} = \min(I(x,y))$	Minimal intensity level of an image patch
Mean	$\mu = \frac{1}{XY}\sum_{y=1}^{Y}\sum_{x=1}^{X} I(x,y)$	Average intensity level of an image patch
Median	The median is the value that separates the lower and upper half of the sorted array of pixel values	Gives a rough idea about the shape of the histogram
Root mean square	$RMS = \sqrt{\frac{1}{XY}\sum_{y=1}^{Y}\sum_{x=1}^{X} I(x,y)^2}$	Measures of differences of intensity level of an image patch

---

IV. Deep learning-based classification of blue light cystoscopy imaging during transurethral resection of bladder tumors

Reprinted with permission from [N. Ali, C. Bolenz, T. Todenhöfer, A. Stenzel, P Deetmar, M. Kriegmair, T. Knoll, S. Porubsky, A. Hartmann, J. Popp, M. C. Kriegmair, and T. Bocklitz: Deep learning-based classification of blue light cystoscopy imaging during transurethral resection of bladder tumors, Scientific Reports, 2021, 11, 1169] Copyright © 2021, Springer Nature, The Author(s)

Erklärungen zu den Eigenanteilen des Promovenden sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation.

N. Ali, C. Bolenz, T. Todenhöfer, A. Stenzel, P Deetmar, M. Kriegmair, T. Knoll, S. Porubsky, A. Hartmann, J. Popp, M. C. Kriegmair, and T. Bocklitz. *Deep learning-based classification of blue light cystoscopy imaging during transurethral resection of bladder tumors*, Scientific reports, 2021, 11, 1169

Beteiligt an (Zutreffendes ankreuzen)												
	1	2	3	4	5	6	7	8	9	10	11	12
Konzeption des Forschungsansatzes	×	×	×	×	×	×	×	×	×	×	×	×
Planung der Untersuchungen	×										×	×
Datenerhebung		×	×	×	×	×	×	×	×	×	×	
Datenanalyse und -interpretation	×											×
Schreiben des Manuskripts	×	×	×	×						×	×	×
Vorschlag Anrechnung Publikationsäquivalente	1.0											

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

Deep learning-based classification of blue light cystoscopy imaging during transurethral resection of bladder tumors

Nairveen Ali[1,2], Christian Bolenz[3], Tilman Todenhöfer[4], Arnulf Stenzel[4], Peer Deetmar[5], Martin Kriegmair[6], Thomas Knoll[7], Stefan Porubsky[8], Arndt Hartmann[9], Jürgen Popp[1,2], Maximilian C. Kriegmair[10]✉ & Thomas Bocklitz[1,2]✉

Bladder cancer is one of the top 10 frequently occurring cancers and leads to most cancer deaths worldwide. Recently, blue light (BL) cystoscopy-based photodynamic diagnosis was introduced as a unique technology to enhance the detection of bladder cancer, particularly for the detection of flat and small lesions. Here, we aim to demonstrate a BL image-based artificial intelligence (AI) diagnostic platform using 216 BL images, that were acquired in four different urological departments and pathologically identified with respect to cancer malignancy, invasiveness, and grading. Thereafter, four pre-trained convolution neural networks were utilized to predict image malignancy, invasiveness, and grading. The results indicated that the classification sensitivity and specificity of malignant lesions are 95.77% and 87.84%, while the mean sensitivity and mean specificity of tumor invasiveness are 88% and 96.56%, respectively. This small multicenter clinical study clearly shows the potential of AI based classification of BL images allowing for better treatment decisions and potentially higher detection rates.

Bladder cancer is among the most common cancers and the leading cause of death in western countries[1]. The primary diagnosis and treatment of bladder cancer is based on endoscopic procedures. Here, the standard of health care is white light (WL) cystoscopy, which offers an excellent sensitivity and specificity to detect papillary tumors, but it misses a significant fraction of small and flat lesions[2,3]. To increase the detection rate of these lesions, modern imaging technologies such as photodynamic diagnosis (PDD) are highly recommended. According to a couple of meta-analyses, 40% of flat cancerous lesions are only detected in BL cystoscopy[4,5]. Consequently, the implementation of PDD can result in a change of the respective bladder cancer risk classification, and thus a more accurate therapy[6]. However, PDD harbors some significant drawbacks concerning its low specificity which ranges from 35 to 60%[4,5]. For instance, it is difficult to distinguish flat cancerous lesions from inflammable alterations following transurethral resection or instillation[6]. Moreover, the interpretation of PDD findings is highly subjective and may vary between observers. This accounts especially for less experienced endoscopists, where the rate of false positives is particularly high[7]. Finally, PDD supports the distinction of malignant and benign tissues but does not offer diagnostic information regarding tumor stage and grading.

Currently, brain-inspired deep neural networks (DNNs) have been revolutionizing artificial intelligence, and they have shown their potential for computer-aided diagnostic systems in various fields such as radiology[8], histopathology[9] and computational neuroscience[10]. In terms of image processing, deep neural networks (DNNs) exhibit the best performing models for object recognition and yield human performance levels for object categorization[11]. Typically, these DNNs mimic the mechanism of human brains by letting DNNs learn specific image features that improve the identification performance on new unlabeled data sets. In the basic architecture

[1]Institute of Physical Chemistry and Abbe Center of Photonics (IPC), Friedrich-Schiller-University, Jena, Germany. [2]Leibniz Institute of Photonic Technology (IPHT), Jena, Germany. [3]Department of Urology, University of Ulm, Ulm, Germany. [4]Department of Urology, University Hospital Tübingen, Tübingen, Germany. [5]Pathology Munich-Nord, Munich, Germany. [6]Urological Hospital Munich-Planegg, Munich, Germany. [7]Department of Urology, Hospital Sindelfingen-Böblingen, University of Tübingen, Sindelfingen, Germany. [8]Institute of Pathology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany. [9]Institute of Pathology, University of Erlangen, Erlangen, Germany. [10]Department of Urology, University Medical Centre Mannheim, Mannheim, Germany. ✉email: Maximilian.Kriegmair@medma.uni-heidelberg.de; thomas.bocklitz@uni-jena.de

Scientific Reports | (2021) 11:11629 | https://doi.org/10.1038/s41598-021-91081-x

nature portfolio

Histology	n (%)
Benign	74 (34.26%)
CIS	17 (7.87%)
Ta, LG	72 (33.33%)
Ta, HG	28 (10.18%)
T1, LG	1 (0.46%)
T1, HG	13 (06.02%)
≥T2, any grade	11 (05.09%)
Low-grade	73 (33.80%)
High-grade	69 (31.94%)
Malignant	142 (65.74%)

Table 1. Distribution of pathological staging after TUR-BT of the respective PDD positive lesions. The separation into low-grad and high-grade was made according to the WHO 2004 classification and malignant was defined as all samples diagnosed with any kind of bladder cancer.

of a DNN, the neural network is trained by passing a data set of labeled images through multiple layers that consist of simple units called neurons. These neurons compute different linear combinations of specific image features captured from the labeled data set and pass the results into the next layer through a static nonlinearity, e.g., replacing negative values by zeros. The previous nonlinear layer is usually known as activation layer, and it is followed by pooling layers that aim to reduce the spatial dimension of the image features. Then, DNNs process the images as a sequence of visual representations in which each neuron detects a specific local region of the feature map in the previous layer while similar feature detectors exist across locations in the feature map[10]. Nonetheless, the term "Deep" in deep neural networks indicates that multiple layers of neurons are utilized in DNNs and improve their identification performance. Such training procedures are usually time consuming and require a large sample size of labeled images, which is rarely available for biomedical applications. Therefore, the concept of transfer learning of DNNs was introduced to deal with classification tasks on small data set. Thereby, the identification knowledge gained via training DNNs on a large annotated data set can be transferred to solve another classification task based on a new and small data set[12,13]. These strategies have shown a great potential for diagnostic classifications of biomedical images using relatively small sample sizes[14–17]. Further, implementing such deep learning models in biomedicine may increase sensitivity and specificity of diagnostic procedures and reduce inter-observer variance[18,19]. However, respective solutions in endoscopy are rare. Recently, Shkolyar and colleagues introduced a deep learning automated image processing platform for cystoscopy. The software was able to identify papillary lesions in videos from WL cystoscopy with a high sensitivity and specificity[20]. Similarly, a recent study established a classification system based on 233 images of bladder wall lesions that was able to identify cancerous formations with a very high sensitivity, but with a low specificity of 50%[21]. Although these preliminary findings are promising, further developments in automated image processing are highly appreciated in urological endoscopy. In this context, PDD was considered as an effective modern imaging technique that offers characteristic information about tumor morphology. This technology utilizes the fluorescence properties of an extrinsic metabolic substrate, which is differently metabolized in cancerous and healthy tissues[22]. Consequently, PDD images contain more comprehensive information as compared to WL images.

The aim of this study is to test the classification of a small BL image data set consisting of bladder tumor and healthy urothelium images. This test was accomplished using an automated image processing pipelines and deep convolutional neural networks (CNNs) as a first step to implement computer-aided diagnosis in urological endoscopy. Our workflow started by preprocessing the BL images to include regions containing bladder tissue only. Then, the identification performance of different pre-trained CNNs in predicting bladder cancer malignancy, invasiveness and grading was investigated via a fine-tuning-based transfer learning strategy. Shortly, a comparison between the implemented CNN models and bladder cancer ratings of two experienced urologists was performed on the basis of the classification sensitivity.

Results

We present in this section the classification results of BL images using the previously explained fine-tuned CNNs. Overall, images from 216 different lesions were included and three classification tasks based on these BL images were established. In Table 1, the pathological results of the biopsied lesions are shown. We can see that the numbers of collected images per class are different; thus, class weights were considered to correct the unbalance within the data set while the CNNs were trained.

Identification of malignant bladder tumors lesions.

The goal of this task is to evaluate the performance of the fine-tuned CNNs in identifying malignant lesions within the BL images, which were collected in multiple centers. Therein, the prediction results obtained by the proposed CNNs were compared with the physician ratings and summarized in Table S1 and Fig. 1. In Fig. 1A, the classification sensitivities of CNNs based on the leave-10-patients-out cross-validation (L10PO-CV) and the physician ratings were visualized as bar charts. For this binary task, the highest classification sensitivity for malignant lesions and for benign lesions are 95.77% and 87.84%; respectively. Here, the fine-tuned MobileNetV2 network provided the best identification of malig-

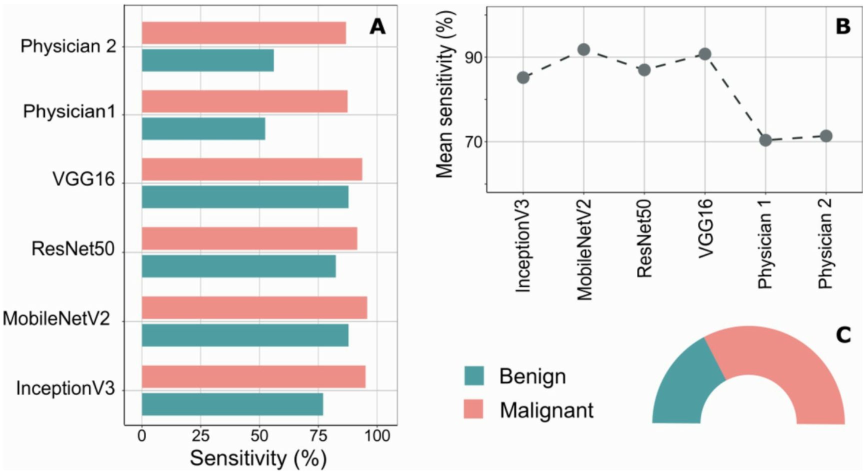


Figure 1. The identification results of the malignant tumor lesions in bladder. (**A**) The classification sensitivities of benign and malignant images based on the considered fine-tuned CNNs and the physician ratings. All CNNs could predict the malignant images quit well with sensitivity of at least 91% and specificity larger than 77%. (**B**) Comparison between the mean sensitivities of the fine-tuned CNNs and the physician ratings. The MobileNetV2 network followed by VGG16 network showed the best classification results with a mean sensitivity of 91.81% and 90.75%; respectively. (**C**) The class distribution of the BL image data set with respect to the percentage of malignant and benign images in the data set. Clearly, the number of malignant images is much larger than the number of the images collected from benign legions.

nant lesions while both fine-tuned MobileNetV2 network and fine-tuned VGG16 network introduced the highest sensitivity for the prediction of benign images. Moving to Fig. 1B, the mean sensitivities of all fine-tuned CNNs and both physicians are visualized. Clearly, the fine-tuned MobileNetV2 network features the highest mean sensitivity with the value of 91.81% followed by the fine-tuned VGG16 network with a mean sensitivity of 90.75%. It is also obvious that the performance of any fine-tuned CNN is at least 15% superior in their classification mean sensitivity as compared to the mean sensitivities obtained by both physician ratings. Nevertheless, the detailed confusion matrices of the pervious binary task are presented in Table S1 while the percentage of class sample sizes to all data set size is shown in Fig. 1C as pie chart.

The identification of bladder cancer invasiveness (T-stage). Because each stage of bladder cancer required a specific medical treatment, we were interested in comparing the classification results of the bladder cancer stages using the considered deep learning models with those obtained by the urologists. To do so, the acquired BL images were categorized into five classes representing benign tissue, carcinoma in situ (CIS) and the following three bladder cancer stages: Ta, T1 and ≥ T2. Table S2 shows the classification results of the bladder cancer invasiveness as a single confusion matrix reflecting all results of each model, i.e., results of CNNs and physician ratings.

In Fig. 2A, the class sensitivities and specificities of each model were plotted as a two-direction bar chart while in Fig. 2B the mean sensitivities and the mean specificities of the respective models were compared based on a point chart. Figure 2C visualizes the proportion of sample size for each class to the whole data set size as a pie chart. Clearly, the class sample sizes are quite different, and the sample size is quite small for the T1 and T2 cancer stages (≤ 15 images). Based on Fig. 2A, the best identification results of benign images were provided by the fine-tuned InceptionV3 network. Thereby, the observed sensitivity and specificity of benign images is 83.78% and 94.37%, respectively. Regarding image classification of CIS and the bladder cancer stage T1, the fine-tuned MobileNetV2 network introduced the best predictions with a class sensitivity of 76.47% for the CIS images and 100% for the T1 images. While the fine-tuned ResNet50 network presented the best identification of the T2 images with 100% classification sensitivity, the highest classification sensitivity for Ta images was obtained again using fine-tuned MobileNetV2 network. Here, the observed sensitivity of bladder cancer stage Ta is 93%. In contrast, both urologists misidentified almost all images of class CIS, T1 and T2, but they could assess the first stage of bladder cancer, i.e., Ta cancer stage, well. Overall, the highest mean sensitivity and the highest mean specificity were reached by the MobileNetV2 network based on the L10PO-CV as it is shown in Fig. 2B. The observed mean sensitivity and mean specificity of the previously mentioned CNN are 88% and 96.56%; respectively. However, the classification mean sensitivities dropped at least 50% when the BL images were accessed by any of both urologists.

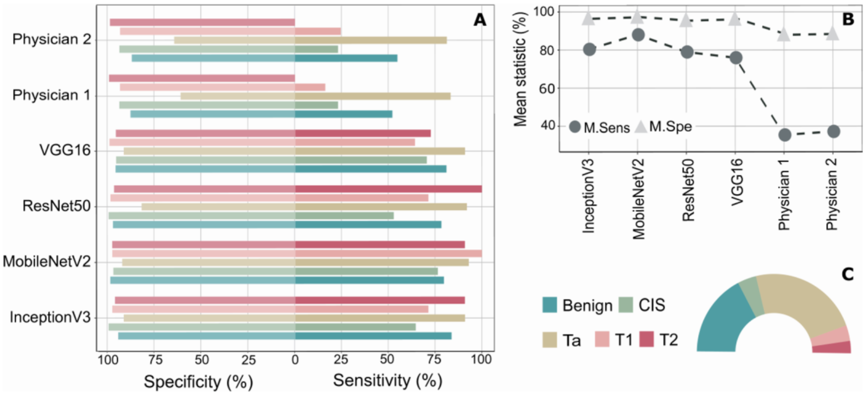


Figure 2. The identification of bladder tumor stage using the fine-tuned CNNs and urologist ratings. (**A**) The class sensitivities and specificities of the considered CNNs and both physician ratings as a two-directions bar chart. While both urologists could not assess well the last two invasive stages of bladder cancer, the detection of these tumor stages was quite good based on all CNNs (**B**) The obtained mean sensitivities and mean specificities for all classification models. The best classification results were achieved by the MobileNetV2 network with a mean sensitivity of 88% followed by the mean sensitivity obtained by the InceptionV3 network. The classification mean sensitivity decreased at least 50% when the same images were assessed by the urologists. (**C**) The class distribution of the BL image data set. The number of involved images for this task varies a lot from one class to another class.

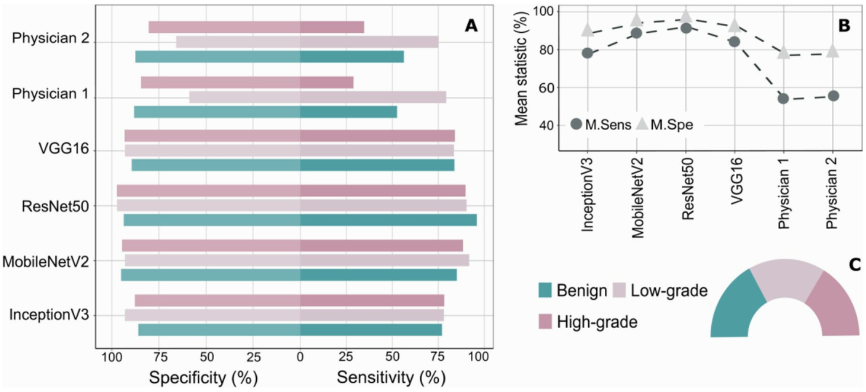


Figure 3. The classification results of bladder cancer grading. (**A**) Overview of the individual class sensitivity and specificity with respect to both physician ratings and the fine-tuned CNNs. (**B**) Summary plot of the mean sensitivities and mean specificities obtained by the fine-tuned CNNs and both physician's ratings. An increase between 25 and 40% of the classification mean sensitivity can be observed if the fine-tuned CNNs were considered to identify bladder cancer grading based on the collected BL images. (**C**) Image class distribution with respect to the whole data size. Almost similar number of images were acquired from benign, low-grade, and high-grade lesions.

The differentiation of the bladder cancer grading.

We present in this subsection the results of deep learning models and physician ratings in identifying the bladder cancer grading based on the BL image data set that was collected from a study involving multiple centers. In Table S3 and Fig. 3, the classification results obtained by the proposed fine-tuned CNNs and by the physician ratings are presented. Table S3 describes the detailed confusion matrices of all previous models and ratings.

The results showed that the fine-tuned ResNet50 network based on L10PO-CV is the best in predicting cancer grading followed by the fine-tuned MobileNetV2 network. Figure 3A depicts the sensitivities and specificities of all models as a two-directions bar chart while Fig. 3B shows a summary plot of the mean sensitivities and mean specificities of all previous models. We see in this figure that the ResNet50 network shows the highest-class sensitivity for high-grade cancer and benign images among the other models, i.e., CNNs and physician ratings. Thereby, the observed sensitivity of the high-grade images and the benign images has the values of 89.86% and 95.95%; respectively. For the identification of low-grade images, the MobileNetV2 network introduced the highest sensitivity in comparison to the other models. Here, the classification sensitivity of low-grade images using the MobileNetV2 network is 91.78%. The overall results indicate that the best performing model is the fine-tuned ResNet50 network while the lowest classification mean sensitivities were obtained by the urologist ratings. Therein, the mean sensitivity and mean specificity of the fine-tuned ResNet50 networks is 92.07% and 96.04%; respectively. On the other side, the mean sensitivity of the first and the second urologist is 53.71% and 55.17%, respectively.

Discussion

In this contribution, we introduced the identification results of bladder cancer using BL endoscopic images acquired from four different urological departments. The data set consists of 216 BL images, that were recorded prior to resection of the respective lesions. The collected BL images were utilized to evaluate the ability of deep learning models in automating the classification of the endoscopic lesions and predicting histopathological results. The bladder cancer identification was demonstrated based on four deep CNNs, and the results were compared with those obtained by two experienced urologists. This comparison was evaluated to predict cancer malignancy, cancer invasiveness and cancer grading. For all these tasks, pre-trained versions of InceptionV3, MobileNetV2, ReNet50, and VGG16 networks were fine-tuned, and then they were evaluated using a L10PO-CV.

The results of the previous named CNNs showed that the fine-tuned MobileNetV2 network has the best performance in detecting images of malignant lesions with a sensitivity of 95.77% and a specificity of 87.84%. The detection performance of this MobileNetV2 network exceeds the performance of other imaging technologies typically used for enhancing bladder cancer detection. For example, probe-based techniques such as CLE or OCT provide sensitivity levels between 80 and 90% for the detection of malignant lesions[23,24]. This underlines the potential of automated image analysis systems in urological endoscopy. For such a classification task, i.e., malignancy identification, an increasing sensitivity should be the primary target. Therefore, the detection of all cancerous lesions during TUR-BT is important for correct identification of bladder cancer staging and adjuvant therapy stratification. Moreover, the failure to remove all tumor tissues is a main reason for high residual tumor rates in patients with intermediate and high-risk NMIBC requiring a 2nd TUR-BT[25]. Recently, an image analysis platform, named: CystoNet, was constructed and evaluated resulting in a sensitivity of 90.9% in detecting papillary bladder tumors[20]. However, unlike the fine-tuned MobileNetV2 network in this study, its specificity was low. Accordingly, Gosnell and colleagues introduced an endoscopy image-based classification system with high sensitivity, but it also suffered of the low specificity (~ 50%)[21]. Consequently, the image analysis technology used in this study has not only shown promising results regarding sensitivity, but also for specificity levels.

Moving to the classification of the cancer invasiveness (T stage), the fine-tuned MobileNetV2 network performed quite well for the presented multiclass task. Thereby, the classification sensitivity of each of the tumor stage T1 and T2 is 100% and 90.91% even though the image sample size per class was quite small (< 15 images per class). Overall, the mean sensitivity and the mean specificity of the fine-tuned MobileNetV2 based on the cross-validation is 88.02% and 96.56%; respectively. For the identification of bladder cancer grading, the fine-tuned ResNet50 network provided the best classification results compared to other CNNs. The observed mean sensitivity and mean specificity of the ResNet50 using the considered cross-validation strategy is 92.07% and 96.04%, respectively. However, the identification performances of both urologists were much worse (mean sensitivity between 35 and 37%) than the classification performance of any of the considered deep learning models.

Beside the challenges presented for identifying bladder cancer invasiveness and grading, flat malignant lesions constitute a challenging situation for urologists. Due to its flat growth pattern, the carcinoma in-situ (CIS) of the urinary bladder is hard to be detected in WL imaging[4]. In contrast to other organs, CIS of the urinary bladder has high-grade characteristics and is potentially invasive. Therefore, it is important to correctly identify and hence cure this cancer type. Although PDD can assist physicians and significantly increase the detection rate of CIS, such flat lesions remain hard to be characterized for urologists[5]. Furthermore, scar tissue and inflammation can mimic CIS characteristics; especially in BL cystoscopy resulting in a high number of false negative biopsies[26]. In this discourse, artificial intelligence-based cancer identification might be an effective tool for better classification of the respective urothelial lesions. Consequently, we were interested in testing the prediction quality of deep learning models in differentiating flat bladder lesions, that include images of benign and CIS lesions. This was achieved using the proposed fine-tuned CNNs on the collected images of the respective bladder lesions, e.g., CIS, and benign tissue. Similar hyperparameters and similar network architectures were used for the aforementioned CNNs with the L10PO-CV being the validation method. In Table S4, the results of the previous binary classification obtained by all considered CNNs were summarized with respect to the class sensitivities, then they were compared with the class sensitivities resulted from the classification of bladder tumor invasiveness, i.e., multiclass models. It turned out that the fine-tuned MobileNetV3 network based on the multiclass training performed the best in the differentiation between benign and CIS lesions with a mean sensitivity of 78.10%. Remarkably, the specificity of CIS lesions in the binary model was high reaching 90% using the InceptionV3 and ReNet50 networks. To improve and verify these findings, further clinical research is needed to enhance the low specificity, which introduces one of the major drawbacks of PDD imaging. Similar drawbacks exist in other imaging techniques such as Optical Coherence Tomography[27]. The advantage of improving the detection sensitivity and specificity of flat lesions would be of particular interest when PDD is utilized in an outpatient setting, where biopsies and resection are not possible[28]. Indeed, a recent study revealed that refuting suspicious lesions is one of the major motivations physicians to use adjunct imaging modalities[28]. Thus, cancer diagnosis based on the combination of imaging technologies and artificial intelligence would be highly appreciated in clinics. Nevertheless, our classification results of flat lesions were highly influenced by the sample size when the CNNs were considered to perform the binary classification, i.e., CIS vs. benign lesions. Therein, the fine-tuned CNNs were trained and validated on 91 BL images with unequal class sizes. However, we expect to achieve better classification performance if the previous mentioned challenges are addressed in future studies.

In summary, transfer learning based on pre-trained CNNs enabled us to identify bladder cancer despite the small sample size and the unbalance in data set. For all tasks, the fine-tuned CNNs provided promising results. Moreover, the misclassification of BL images in most cases was expected due to the high variations between the images and due to other systematic errors. Figure 4 presents examples of correctly classified and misclassified images related to two of the considered classification tasks, i.e., the classification of malignant lesions and the classification of cancer grading. It is clear in this figure that the fluorescence of some images is very low while it is very spotty in others. Additionally, the urine fluorescence within some images may drown out the red fluorescence; therefore, these images were mostly incorrectly identified. Beside the fluorescence issues, some images depicted flat lesions while others were not close enough to capture the suspicious lesions. As a result, these images were also misclassified.

Nevertheless, the main limitations of the current study its retrospective design. Further, the image identification results provided by both urologists are not the typical procedure considered for such cancer diagnosis in clinical practice. Due to the low morbidity of additional biopsies, urologists tend to biopsy some to all suspicious lesions, which results in a high sensitivity and low specificity. Another limitation is the low number of BL image for all lesions, specifically for CIS lesions. These small sample sizes allowed for exploratory analyses on lesions. However, only BL images with obvious histopathological correlates were included while all four centers used equivalent clinical and technical set-ups.

Conclusion

The results in this study demonstrated the potential of artificial intelligence-based classification models for the diagnostics of bladder cancer based on BL cystoscopic images. In this context, further studies need to be performed in order to build an automatic BL cystoscopic platform that assists physicians in identifying and classifying potential lesions of interest. Applying such platforms into clinical routines aims to assist surgeons and aids the cancer diagnoses. Potentially this technology could increase the detection rates of cancer and improve the relative low specificity of BL imaging. However, the system will not be considered to substitute the opinion of endo-urologists or pathologists for clinical decision making.

Material and methods

Image acquisition and pathological evaluation.

A total of 216 BL images acquired during PDD-guided transurethral resection of bladder tumor (TUR-BT) were collected from four urological departments retrospectively and one image was taken for an individual patient. Therefore, every image represents a patient. Routine pathological evaluation was performed, and all tumors were classified according to the world health organization (WHO) classification 2004. Only endoscopic images recorded prior to the resection of the respective lesions were used while the distance from the endoscopy to the region of interest was not standardized. For PDD, intravesical instillation of 85 mg Hexaminolevulinathydrochlorid (Hexvix®, IPSEn Pharma, Boulogne, France) was performed 60 min prior to PDD. Imaging was performed using the Tricam II® system and a 30-degree Hopkins II optic (Karl Storz, Tuttlingen, Germany) in all centers. Further, two experienced urologists (CB and MCK, both > PDD 300 TUR-BTs) assessed the endoscopic images. Therein, the following distinctions were requested from the urologists and subsequently performed by different deep CNNs based on the PDD images only:

(i) Malignancy: malignant vs benign lesions
(ii) Tumor invasiveness (T-Stage): benign lesions vs carcinoma in situ (CIS) vs Ta vs T1 vs ≥ T2
(iii) Tumor grading: benign vs high-grade vs low-grade cancer
(iv) Flat lesions: benign vs CIS

For the previous classification tasks, the number of collected image per class is not equal; therefore, class weights were considered to correct this unequal class sizes within the data set while the classification models were trained.

Nevertheless, the data was collected retrospectively. Written informed consent was obtained, if possible. Data was analyzed and forwarded anonymized from the respective clinical center to all other study participants. This study was approved by the local ethical committee of the leading study side Mannheim (Ethics Committee II of the University of Heidelberg at the Medical Faculty Mannheim, Ref Number: 2015 549N MA) and in accordance with the Declaration of Helsinki.

Image region of interest.

To improve the identification results of bladder cancer, only the regions of interest (ROI) in the PDD images were included in the data modeling (see Fig. 5). Here, the ROI of an image refers to the image area containing the bladder tissue. This determination of ROIs was performed automatically based on image processing techniques, and it started by enhancing the contrast of the red and blue channels of all PDD images using the contrast limited adaptive histogram equalization algorithm[29]. Thereafter, the tissue area of each image was extracted by fitting a disk in order to remove background areas. Finally, the ROI of an image was acquired as an inscribed square region within the extracted image disk. Applying the previous image preprocessing pipeline returns images of the size of 384 × 384 pixels.

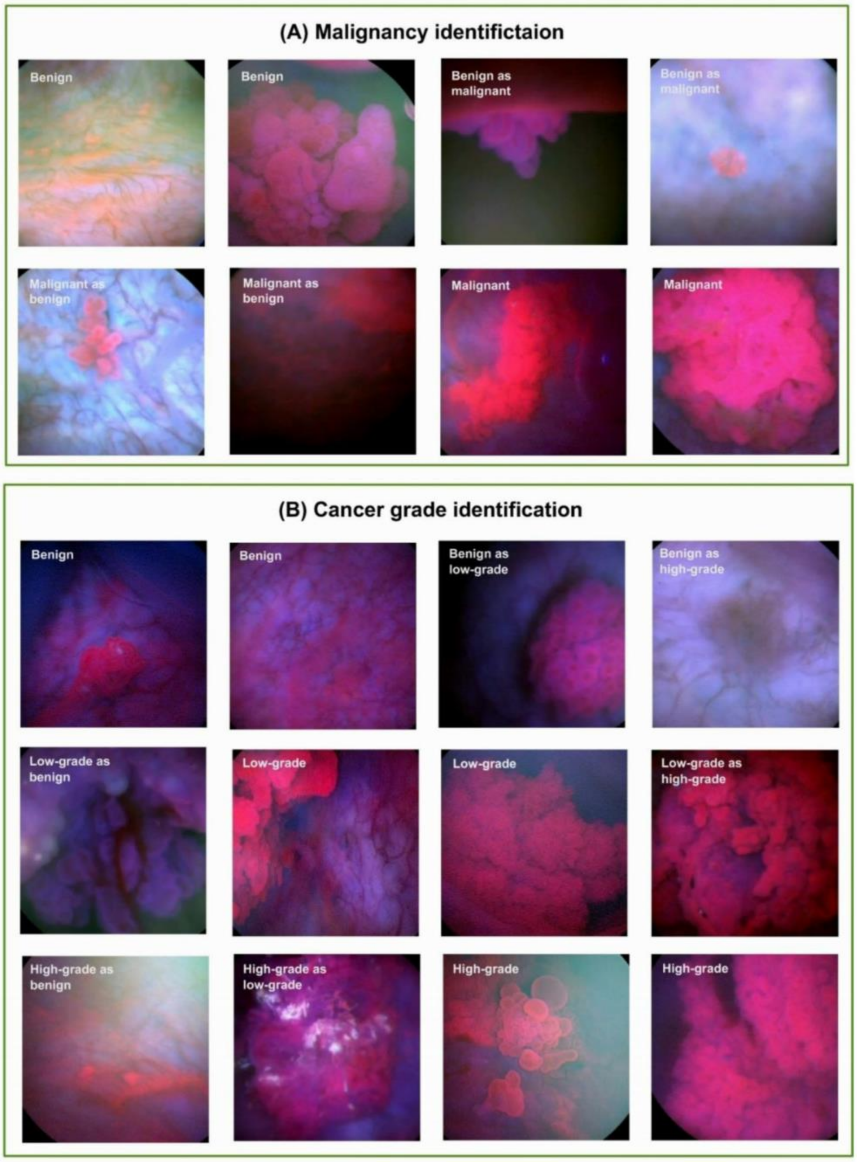


Figure 4. Examples of correctly predicted images and misclassified images using the fine-tuned CNNs.

Transfer learning based on deep convolutional neural networks. The basic idea of transfer learning for deep learning models is to utilize knowledge gained by training a deep neural network on a large and annotated data set to solve another classification task[12,13]. In this contribution, we proposed a transfer learning strategy in which an ensemble of different pre-trained deep CNNs were fine-tuned to improve the classification of BL images. The respective CNNs are the InceptionV3 network[30], MobileNetV2 network[31], ResNet50 network[32] and VGG16 network[33], and they represent common freely available fully connected CNNs that were pre-trained on the ImageNet dataset. As we mentioned, the identification ability of the previous pre-trained

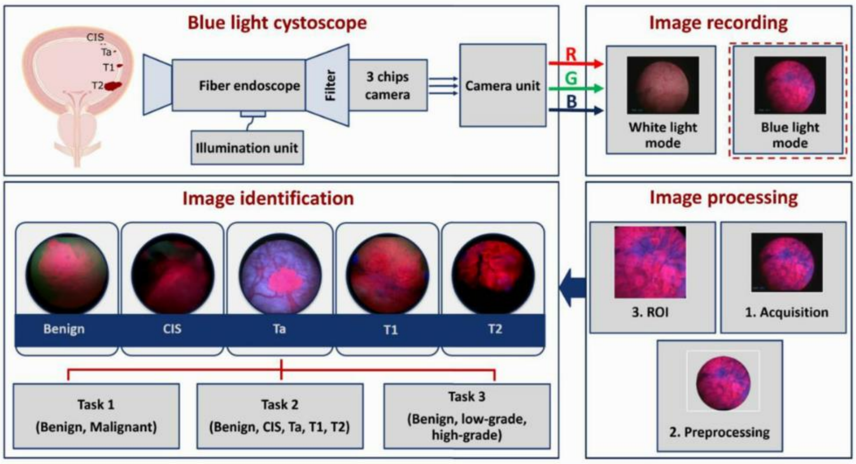


Figure 5. Overview of image acquisition and image processing using the blue light cystoscope.

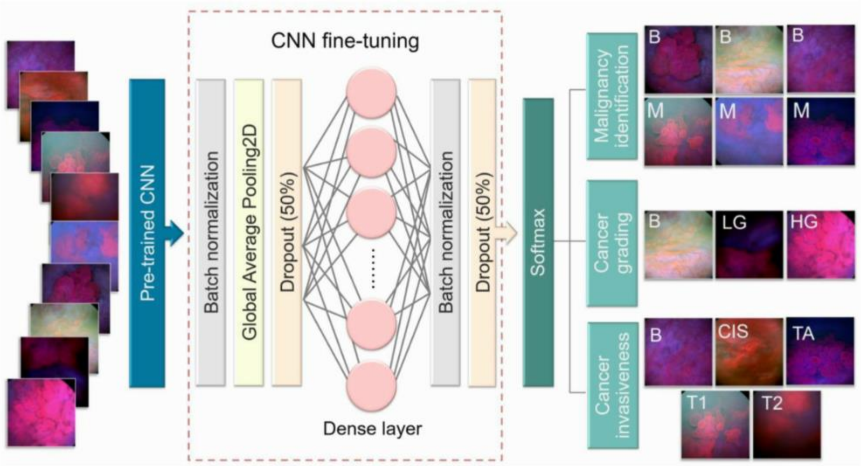


Figure 6. Schematic diagram of the CNN fine-tuning considered for identifying bladder cancer. Each pre-trained CNN was fine-tuned by appending two batch normalization layers, a global average pooling layer, dropout layers with the probability of 50%, a dense layer to improve cancer identification, and Softmax activation layer. The last layer delivers different label probabilities for each input image with respect to each classification task.

CNNs can be transferred via a fine-tuning approach, which was accomplished by appending additional layers on top of each network. These additional layers are two batch normalization layers, a global average pooling layer, dropout layers with the probability of 50%, and a dense layer to find the best combinations of the already learnt features that improve bladder cancer identification. The last additional layer is a Softmax activation layer, which provides label probabilities for each image with respect to the considered classification task (see Fig. 6 for more details). The parameters of these layer were optimized by an Adam optimizer, which was trained for 100 epochs based on a mini-batch of 5 patches. The optimization hyperparameters were a learning rate of 0.001 and the categorical-cross entropy loss function. Class imbalance was tackled by the SckiKit Learn function *class_weight.compute_class_weight ('balanced')*.

Image augmentation and cross validation. The input image size of all previous CNNs was fixed to be 224 × 224 pixels; therefore, all endoscopic images were first down-sampled to that size. Then, class labels of all down-sampled images were predicted four times based on the four fine-tuned CNNs. These CNN were evaluated using the L10PO-CV as a validation strategy. Therein, we always fixed 10 images (from 10 patients) as test set and 10 images as validation set while the remaining images were utilized to train the considered fine-tuned CNN. The last procedure was repeated 22 times until labels of all images (and patients) were predicted by all fine-tuned CNNs. Because the training set has a maximum size of 196 images per each cross-validation iteration, the BL images utilized to train the CNNs within each iteration were augmented automatically using random rotations by steps of 10° degrees within the range of 0° to 180°. Thereafter, each fine-tuned CNN was trained for 100 epochs based on a mini-batch of 5 patches and using the Adam optimizer with a learning rate of 0.001 to minimize the categorical-cross entropy loss function.

Data modeling and models evaluation. For interpretation of the cross-validation results, we calculated the confusion matrix and the classification sensitivity and specificity with respect to all tested classification tasks and all fine-tuned CNNs. For ratings of the urologists, confusion matrices were also computed by comparing these ratings with the image ground truth, e.g., the pathological annotation of the biopsied tissue sample. Lastly, the resulting mean sensitivities and mean specificities were calculated for all CNNs results and for both urologists' results.

All computations in this work were accomplished based on in-house written functions using the programming language Python version 3.7[34] and the statistical programing language R version 3.6[35].

Received: 11 December 2020; Accepted: 18 May 2021
Published online: 02 June 2021

References

1. Antoni, S. *et al.* Bladder cancer incidence and mortality: A global overview and recent trends. *Eur. Urol.* **71**(1), 96–108 (2017).
2. Burger, M. *et al.* Epidemiology and risk factors of urothelial bladder cancer. *Eur. Urol.* **63**(2), 234–241 (2013).
3. Cina, S. J. *et al.* Correlation of cystoscopic impression with histologic diagnosis of biopsy specimens of the bladder. *Hum. Pathol.* **32**(6), 630–637 (2001).
4. Burger, M. *et al.* Photodynamic diagnosis of non-muscle-invasive bladder cancer with hexaminolevulinate cystoscopy: A meta-analysis of detection and recurrence based on raw data. *Eur. Urol.* **64**, 846–854 (2013).
5. Rink, M. *et al.* Hexyl aminolevulinate-guided fluorescence cystoscopy in the diagnosis and follow-up of patients with non-muscle-invasive bladder cancer: A critical review of the current literature. *Eur. Urol.* **64**(4), 624–638 (2013).
6. Daneshmand, S. *et al.* Blue light cystoscopy for the diagnosis of bladder cancer: Results from the US prospective multicenter registry. *Urol. Oncol.* **36**(8), 361–361 (2018).
7. Gravas, S. *et al.* Is there a learning curve for photodynamic diagnosis of bladder cancer with hexaminolevulinate hydrochloride?. *Can. J. Urol.* **19**(3), 6269–6273 (2012).
8. Saba, L. *et al.* The present and future of deep learning in radiology. *Eur. J. Radiol.* **114**, 14–24 (2019).
9. Srinidhi, C., Ciga, O., & Martel, A. *Deep neural network models for computational histopathology: A survey.* (2019). http://arxiv.org/abs/1912.12378.
10. Kietzmann, T. C., McClure, P. & Kriegeskorte, N. *Deep Neural Networks in Computational Neuroscience* (Oxford University Press, 2019).
11. Cichy, R. M. *et al.* Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* **6**(1), 27755 (2016).
12. Shin, H. C. *et al.* Deep convolutional neural networks for computer-aided detection: CNN Architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **35**(5), 1285–1298 (2016).
13. Pradhan, P. *et al.* Deep learning a boon for biophotonics?. *J. Biophoton.* **1**, e201960186 (2020).
14. Ali, N. *et al.* Automatic label-free detection of breast cancer using nonlinear multimodal imaging and the convolutional neural network ResNet50. *Transl. Biophoton.* **1**(1–2), e201900003 (2019).
15. Aubreville, M. *et al.* Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning. *Sci. Rep.* **7**(1), 11979 (2017).
16. Wu, E. *et al.* Deep learning approach for assessment of bladder cancer treatment response. *Tomography* **5**(1), 201–208 (2019).
17. Xu, H. *et al.* Using transfer learning on whole slide images to predict tumor mutational burden in bladder cancer patients. *BioRxiv* **1**, 554527 (2019).
18. Tokas, T. *et al.* A 12-year follow-up of ANNA/C-TRUS image-targeted biopsies in patients suspicious for prostate cancer. *World J. Urol.* **36**(5), 699–704 (2018).
19. Xue, J. *et al.* Deep learning-based detection and segmentation-assisted management of brain metastases. *Neuro Oncol.* **22**(4), 505–514 (2019).
20. Shkolyar, E. *et al.* Augmented bladder tumor detection using deep learning. *Eur. Urol.* **76**(6), 714–718 (2019).
21. Gosnell, M. E. *et al.* Computer-assisted cystoscopy diagnosis of bladder cancer. *Urol. Oncol.* **36**(1), e9–e15 (2018).
22. Mari, A. *et al.* Novel endoscopic visualization techniques for bladder cancer detection: A review of the contemporary literature. *Curr. Opin. Urol.* **28**(2), 214–218 (2018).
23. Goh, A. C. *et al.* Optical coherence tomography as an adjunct to white light cystoscopy for intravesical real-time imaging and staging of bladder cancer. *Urology* **72**(1), 133–137 (2008).
24. Wu, J. *et al.* Optical biopsy of bladder cancer using confocal laser endomicroscopy. *Int. Urol. Nephrol.* **51**(9), 1473–1479 (2019).
25. Soria, F. *et al.* The rational and benefits of the second look transurethral resection of the bladder for T1 high grade bladder cancer. *Transl. Androl. Urol.* **8**(1), 46–53 (2019).
26. Nassiri, N. *et al.* Detecting invisible bladder cancers with blue light cystoscopy. *Urology* **139**, e8–e9 (2020).
27. Schmidbauer, J. *et al.* Fluorescence cystoscopy with high-resolution optical coherence tomography imaging as an adjunct reduces false-positive findings in the diagnosis of urothelial carcinoma of the bladder. *Eur. Urol.* **56**(6), 914–919 (2009).
28. Lotan, Y. *et al.* Blue light flexible cystoscopy with hexaminolevulinate in non-muscle-invasive bladder cancer: Review of the clinical evidence and consensus statement on optimal use in the USA: Update 2018. *Nat. Rev. Urol.* **16**(6), 377–386 (2019).
29. Reza, A. M. Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. *J. VLSI Signal Process.* **38**(1), 35–44 (2004).

30. Szegedy, C. *et al. Rethinking the inception architecture for computer vision*. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
31. Sandler, M. *et al. MobileNetV2: Inverted residuals and linear bottlenecks*. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018).
32. He, K. *et al. Deep residual learning for image recognition*. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
33. Simonyan, K. & Zisserman, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition* (2014). http://arxiv.org/abs/1409.1556.
34. Van Rossum, G.A.D. & Fred, L. *Python 3 Reference Manual*. (CreateSpace, 2009).
35. R Core Team, *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2019).

Acknowledgements

The funding from the DFG for the Project BO4700/4-1 is highly appreciated. Financial support by the BMBF for the Project Uro-MDD (FKZ 03ZZ0444J) is highly acknowledged.

Author contributions

Manuscript writing and review: N.A., C.B., T.T., A.S., P.D., M.K., T.K., S.P., A.H., J.P., M.C.K., T.B. Data generation: C.B., T.T., A.S., P.D., M.K., T.K., S.P., A.H., J.P., M.C.K. Medical information and diagnostics: C.B., T.T., A.S., P.D., M.K., T.K., S.P., A.H., M.C.K. Data analysis: N.A., T.B. Project conception: A.H., J.P., M.C.K., T.B.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-91081-x.

Correspondence and requests for materials should be addressed to M.C.K. or T.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021

Supplementary Materials for:

Deep learning-based classification of blue light cystoscopy imaging during transurethral resection of bladder tumors

Nairveen Ali [1,2], Christian Bolenz [3], Tilman Todenhöfer [4], Arnuf Stenzel [4], Peer Deetmar [5],
Martin Kriegmair [6], Thomas Knoll [7], Stefan Porubsky [8], Arndt Hartmann [9], Jürgen Popp [1,2],
Maximilian C. Kriegmair [10,*], Thomas Bocklitz [1,2,*]

[1] Institute of Physical Chemistry and Abbe Center of Photonics (IPC), Friedrich-Schiller-University, Jena, Germany

[2] Leibniz Institute of Photonic Technology (IPHT), Jena, Germany

[3] Department of Urology, University of Ulm, Ulm, Germany

[4] Department of Urology, University Hospital Tübingen, Tübingen, Germany

[5] Pathology Munich-Nord, Munich, Germany

[6] Urological Hospital Munich-Planegg, Germany

[7] Department of Urology, Hospital Sindelfingen-Böblingen, University of Tübingen, Sindelfingen, Germany

[8] Institute of Pathology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

[9] Institute of Pathology, University of Erlangen, Erlangen, Germany

[10] Department of Urology, University Medical Centre Mannheim, Mannheim, Germany

* Corresponding authors

Table S1. The confusion tables of bladder cancer malignancy identification. The results here were obtained by the fine-tuned CNNs based on the cross- validation and by physician ratings.

Model	**True**	**Prediction**		**Sens.**	**Model**	**True**	**Prediction**		**Sens.**
		Benign	Malignant				Benign	Malignant	
InceptionV3	Benign	57	17	77.03%	**ResNet50**	Benign	61	13	82.43%
	Malignant	7	135	95.07%		Malignant	12	129	91.55%
MobileNetV2	Benign	65	9	87.84%	**VGG16**	Benign	65	9	87.84%
	Malignant	6	136	95.77%		Malignant	9	133	93.66%
Physician 1	Benign	39	35	52.70%	**Physician 2**	Benign	41	33	56.41%
	Malignant	17	125	88.03%		Malignant	18	124	87.32%

Table S2. The prediction results of cancer stage based on the fine-tuned CNNs and physician ratings.

Model	True	Prediction					Sens.	Spec.
		Benign	Ta	T1	T2	CIS		
IncepctionV3	Benign	62	5	3	4	0	83.78%	94.37%
	Ta	4	91	2	2	1	91%	91.38%
	T1	1	3	10	0	0	71.43%	97.52%
	T2	1	0	0	10	0	90.91%	96.10%
	CIS	2	2	0	2	11	64.71%	99.50%
MobileNetV2	Benign	59	7	3	1	4	79.72%	98.59%
	Ta	1	93	2	2	2	93%	92.24%
	T1	0	0	14	0	0	100%	97.52%
	T2	0	1	0	10	0	90.91%	97.56%
	CIS	1	1	0	2	13	76.47%	96.98%
ResNet50	Benign	58	13	0	2	1	78.38%	97.18%
	Ta	3	92	2	3	0	92%	81.90%
	T1	0	4	10	0	0	71.43%	98.51%
	T2	0	0	0	11	0	100%	96.58%
	CIS	1	4	1	2	9	52.94%	99.50%
VGG16	Benign	60	5	0	6	3	81.08%	95.77%
	Ta	2	91	2	1	4	91%	91.37%
	T1	1	2	9	1	1	64.29%	99.01%
	T2	0	2	0	8	1	72.72%	95.61%
	CIS	3	1	0	1	12	70.59%	95.48%
Physician 1	Benign	39	25	2	1	7	52.70%	88.03%
	Ta	6	84	7	0	3	84.00%	61.21%
	T1	3	6	2	0	1	16.67%	93.63%
	T2	1	9	3	0	0	0%	99.50%
	CIS	7	5	1	0	4	23.53%	94.47%
Physician 2	Benign	41	21	2	2	8	55.41%	87.32%
	Ta	7	82	7	0	4	82.00%	64.66%
	T1	3	6	3	0	0	25.00%	93.62%
	T2	1	9	3	0	0	0%	99.01%
	CIS	7	5	1	0	4	23.53%	93.96%

Table S3. Confusion matrices obtained by the deep learning models and physician ratings.

Model	True	Prediction			Sens.	Spec.
		Benign	Low-grade	High-grade		
IncepctionV3	Benign	57	7	10	77.03%	85.92%
	Low-grade	8	57	8	78.08%	93.01%
	High-grade	12	3	54	78.26%	87.76%
MobileNetV2	Benign	63	5	6	85.14%	95.07%
	Low-grade	4	67	2	91.78%	93.01%
	High-grade	3	5	61	88.44%	94.56%
ResNet50	Benign	71	2	1	95.95%	93.66%
	Low-grade	4	66	4	90.41%	97.20%
	High-grade	5	2	62	89.86%	97.27%
VGG16	Benign	62	7	5	83.80%	89.43%
	Low-grade	7	61	5	83.56%	93.01%
	High-grade	8	3	58	84.06%	93.20%
Physician 1	Benign	39	23	12	52.70%	88.02%
	Low-grade	4	58	11	79.45%	58.74%
	High-grade	13	36	20	28.99%	84.35%
Physician 2	Benign	41	19	14	56.40%	87.32%
	Low-grade	3	55	15	75.34%	65.73%
	High-grade	15	30	24	34.78%	80.27%

Table S4. A comparison between binary CNN models and multiclass CNN models in image identification of benign and carcinoma in situ (CIS) bladder lesions.

Binary model				Multiclass model			
Model	**Class**	**Sens.**	**M. Sens.**	**Model**	**True**	**Sens.**	**M. Sens.**
IncepctionV3	Benign	90.78%	57.03%	**IncepctionV3**	Benign	83.78%	74.35%
	CIS	23.52%			CIS	64.71%	
MobileNetV2	Benign	56.75%	60.73%	**MobileNetV2**	Benign	79.72%	78.10%
	CIS	64.70%			CIS	76.47%	
ResNet50	Benign	89.19%	56.35%	**ResNet50**	Benign	78.38%	65.66%
	CIS	23.53%			CIS	52.94%	
VGG16	Benign	97.29%	51.59%	**VGG16**	Benign	81.08%	75.84%
	CIS	05.88%			CIS	70.59%	

V. Automatic label-free detection of breast cancer using nonlinear multimodal imaging and the convolutional neural network ResNet50

The Reprinted [N. Ali, E. Quansah, K. Köhler, T. Meyer, M. Schmitt, J. Popp, A. Niendorf, and T. Bocklitz: Automatic label-free detection of breast cancer using nonlinear multimodal imaging and the convolutional neural network ResNet50, Translational Biophotonics, 2019, 1, e201900003] is licensed under a Creative Commons Attribution-NonCommercial NoDerivatives 4.0 International License.

Erklärungen zu den Eigenanteilen des Promovenden sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation.

N. Ali, E. Quansah, K. Köhler, T. Meyer, M. Schmitt, J. Popp, A. Niendorf, and T. Bocklitz, *Automatic label-free detection of breast cancer using nonlinear multimodal imaging and the convolutional neural network ResNet50*, Translational Biophotonics, 2019, 1, e201900003								
Beteiligt an (Zutreffendes ankreuzen)								
	1	2	3	4	5	6	7	8
Konzeption des Forschungsansatzes	×	×		×	×	×	×	×
Planung der Untersuchungen	×	×		×	×	×	×	×
Datenerhebung		×	×	×	×	×	×	×
Datenanalyse und -interpretation	×							
Schreiben des Manuskripts	×	×	×	×	×	×	×	×
Vorschlag Anrechnung Publikationsäquivalente	1.0	1.0						

Received: 14 October 2019 | Revised: 8 November 2019 | Accepted: 8 November 2019
DOI: 10.1002/tbio.201900003

FULL ARTICLE

TRANSLATIONAL BIOPHOTONICS

Automatic label-free detection of breast cancer using nonlinear multimodal imaging and the convolutional neural network ResNet50

Nairveen Ali[1,2] | **Elsie Quansah**[1,2] | **Katarina Köhler**[3] | **Tobias Meyer**[1,2] | **Michael Schmitt**[1,2] | **Jürgen Popp**[1,2,4,5] | **Axel Niendorf**[3] | **Thomas Bocklitz**[1,2]

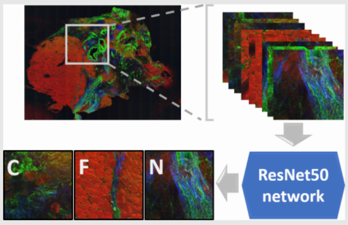
[1]Institute of Physical Chemistry and Abbe Center of Photonics (IPC), Friedrich-Schiller-University, Jena, Germany
[2]Leibniz Institute of Photonic Technology (Leibniz-IPHT), Member of Leibniz Research Alliance 'Health Technologies', Jena, Germany
[3]Institut für Histologie, Zytologie und molekulare Diagnostik, Pathologie Hamburg-West GmbH, Hamburg, Germany
[4]Center for Sepsis Control and Care (CSCC), Jena University Hospital, Jena, Germany
[5]InfectoGnostics, Forschungscampus Jena, Jena, Germany

Correspondence
Thomas 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: FKZ 03ZZ0444J; Deutsche Forschungsgemeinschaft, Grant/Award Numbers: BO 4700/1-1, PO 563/30-1, STA 295/11-1; TMBWK, Grant/Award Numbers: 2015FOR-0001, 2015-0023; Europäischen Fonds für regionale Entwicklung, Grant/Award Numbers: 2017 FGI 0018, 2016 FGI 0011; Free State of Thuringia

Abstract

Breast cancer is the main cause of all female cancer deaths worldwide. Because of the lack of early symptoms, the early detection of breast cancer becomes challenging. This detection is performed by screening techniques in organized preventive examinations. A promising imaging technology that can detect biomolecular alterations and can support the screening technologies by enhancing their low sensitivity, is nonlinear multimodal imaging. To detect these biomolecular alterations, machine-learning algorithms are utilized. Our analysis started by preprocessing the images and comparing them to the pathological diagnosis. We trained two classification models utilizing the deep convolutional neural network ResNet50. This network was either used as feature extractor or to be fine-tuned as a classification model. Beside these two classification approaches, two data validation techniques were investigated: the leave-one-patient-out cross-validation (LOPO-CV) and the training-test validation. The best reported result of breast cancer detection was introduced by the fine-tuned ResNet50 network and LOPO-CV accounting to 86.23% mean-sensitivity.



KEYWORDS
breast cancer imaging, coherent anti-stokes Raman scattering, computer aided diagnosis, convolutional neural network, deep learning, image analysis, nonlinear multimodal imaging, second-harmonic generation, two-photon excited fluorescence

---

Nairveen Ali and Elsie Quansah contributed equally to this study.

Jürgen Popp, Axel Niendorf and Thomas Bocklitz share the senior and corresponding authorship.

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.
© 2019 The Authors. *Translational Biophotonics* published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

1 | INTRODUCTION

Breast cancer is the most commonly diagnosed cancer in women [1, 2]. According to estimates from the World Health Organization (WHO) breast cancer affects 2.1 million women each year and is the leading cause of female cancer deaths [3]. In 2018, breast cancer accounted as the second commonly diagnosed cancer with 11.6% of the total cancer cases and 6.6% of the total female deaths [4]. Survival among breast cancer patients largely depends on early detection. A late detection of this disease is often due to lack of early symptoms which makes the treatment challenging [2]. Fortunately, previous studies have shown that early diagnosis and suitable treatment could reduce significantly the death rate [5]. Consequently, the most critical point for best prognosis of breast cancer patients is an early identification of early cancer stages. Nevertheless, the current gold-standard for definitive diagnosis of breast cancer is visual inspection of histopathological stained tissue sections after a biopsy of tissue material is taken, which is time consuming and invasive.

To find suspicious lesions, which should be biopsied and diagnosed breast cancer screening is necessary. Mammography, which is a low-dose X-ray examination of woman's breast, is the most commonly used method for this purpose. Though popular, it is less effective for imaging small localized and early tissue alterations (small tumors <1 mm, about 100 000 cells). It is also less accurate in patients with dense glandular breasts and for those under 40 years old [6, 7]. Other imaging modalities have emerged to supplement mammography and improve the accuracy of breast cancer diagnosis without the need of a biopsy. Ultrasound imaging has been applied as an additional medical imaging tool for mammography yielding significant cancer detection improvement compared with mammography alone (sensitivity of 92%) [8–10]. However, on its own, ultrasound has a sensitivity of 34% and fails to detect lesions accurately. Magnetic resonance imaging (MRI) has the ability to detect small lesions with high sensitivity of 94.4% [11] and it has a high spatial and temporal resolution. Additionally, MRI exhibits a good signal to noise ratio [12]. It has unfortunately a very low specificity of 26.4%, which can lead to many false positives [13, 14]. It is therefore highly recommended for screening breast cancer in high-risk women, but not for all women [15]. The hybrid technique of Positron emission tomography (PET) and computer tomography (CT) is the most accurate method for visualizing the spread of tumors or detecting the tumor's response to therapy, but the limitation of this method for breast imaging is its poor detection rate for small breast carcinomas [16–19]. Thus, an alternative imaging technology is needed that can measure small breast cancers. This technique would need to provide fast image acquisition without losing molecular contrast and can be applied in vivo to supplement the screening techniques described above. Nonlinear multimodal imaging, the combination of coherent anti-Stokes Raman scattering (CARS), two-photon excited fluorescence (TPEF), and second-harmonic generation (SHG), features these properties and allows for noninvasive and label-free investigation of cells and tissues. Hence, this method might be an appropriate method for in vivo analysis as optical biopsy, especially in combination with fiber-based measurements [20]. Multimodal imaging has been successfully applied in ex vivo tissue investigations of inflammatory bowel disease (IBD) [21], brain tissue [22], larynx carcinoma [23, 24], lung tissue [25], carcinomas in the colon and nonmelanoma skin cancer [26–28].

The main advantage of the previous mentioned multimodal nonlinear imaging is the direct visualization not only of tissue morphology but also of the molecular composition of the tissue. Particularly, CARS visualizes mainly the lipid distribution of tissue, TPEF visualizes autofluorophores including proteins like elastin and keratin, pigments like melanin and enzymes like NADH and flavines [29, 30] and SHG visualizes specific proteins organized in quasi crystalline structures such as collagen, the most frequent protein and a main constituent of the extracellular matrix, actin-myosin, the motor proteins of the muscle cells and tubulin [31]. A visualization of the three nonlinear processes is depicted in Figure 1. However, it is required that the optical data obtained from multimodal imaging is translated into diagnostic relevant information. Recently, computer aided diagnosis using machine learning (ML) algorithms is widely utilized for this task [32].

The essential aim of computer aided diagnosis using ML algorithms is to extract diagnostic relevant information based on automatic analysis of the obtained images in biomedical studies. This automatic analysis becomes a necessity due to the fast-growing amounts of acquired

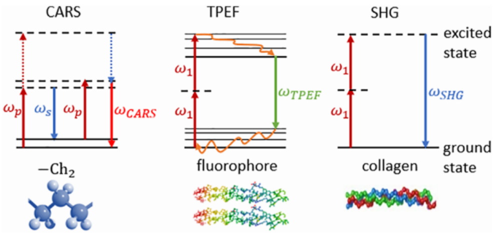


FIGURE 1 Overview of nonlinear multimodal techniques. CARS visualizes lipids and proteins, TPEF visualizes endogenous fluorophores, such as elastin, keratin and SHG visualizes collagen

spectroscopic and microscopic imaging data in biomedical studies [33, 34]. Thereby, the computational models are integrated within the computer system in order to understand and analyze the data by exploring patterns within the collected datasets [35]. These patterns can be unraveled by ML, which is considered as an important discipline in artificial intelligence. The ML algorithms can be categorized roughly into supervised ML and unsupervised ML techniques [36]. In our work, we focus on supervised learning, which aims to find a mathematical relation between the image data and the image labels. Until few years ago, the efficiency of ML algorithms and other pattern recognition applications relied on extracting manually designed image features. After extraction, these extracted features are mapped to the decision variable with easy classification algorithms [32]. The previous type of ML is mostly known as classical ML and it was successfully implemented, for example, in the investigation of inflammatory bowel disease [21], the detection of head and neck carcinoma [23], and in the analysis of basal cell cancer of skin tissue [37]. Beside the classical ML approach, deep learning (DL) is often used for biomedical image analysis. This approach showed great advantages in reducing the human effort by automating most data learning phases. DL algorithms are ideally suitable for biomedical image analysis tasks including cell detection and cell counting [38–40], image segmentation [41, 42] and tissue classification [43, 44].

In this contribution, we present an automatic detection of breast cancer based on 21 multimodal images (from 21 patients). This cancer investigation was performed based on different combinations of ML algorithms and data validation techniques. The main analysis part, e.g, the deep learning method, was constructed using the fully connected convolutional neural network ResNet50 [45]. Hereby, the ResNet50 network was employed either as a feature extractor for classical ML or as a direct classification model. In the latter case, the ResNet50 network was fine-tuned on our image dataset.

2 | PATIENT SAMPLES AND METHODS

2.1 | Patient sample preparation

For this study, breast cancer samples were obtained from 21 patients undergoing routine biopsy at University hospital, Hamburg. Immediately after excision, the sample integrity was preserved by fast freezing in liquid nitrogen. For sectioning, the samples were mounted with a drop of water on the cryostat sample holder and cut into 20 μm thick sections. The sections were then deposited on CaF_2 slides to avoid nonresonant background from CARS. No fixatives or embedding medium were used therefore preserving the lipid distribution.

Nonlinear multimodal images were obtained from these breast cancer tissue samples. After performing multimodal imaging, the tissue samples were stored at $-20^{\circ}C$ and later stained with hematoxylin and eosin (H&E) stain and analyzed with transmission light microscopy. Through 21 samples were measured, only 15 multimodal images with their corresponding H&E stain images were used for training and validating our image analysis approach. This was due to tissue disruption after the staining process. Another challenge with the dataset was the hand-annotated tissue images. The tissue was classified into three classes; tumor, fat and normal, but this annotation process remains difficult due to registration challenges and introduced tissue alterations due to the staining process.

2.2 | Nonlinear multimodal microscopy

The schematic representation of the experimental setup depicted in Figure 2, has previously been described in detail [46]. Briefly, a continuous-wave Neodymium-Vanadate laser at a wavelength of 532 nm with an average power of 18 W is used to pump a Coherent Mira HP Titanium-Sapphire (Ti:Sa) laser (Coherent, Santa Clara, California). The Ti:Sa-laser generates 2-3 ps pulses (FWHM) with an average output power of 3.5 W and operating at a repetition rate of 76 MHz. The Ti:Sa-laser output of 830 nm is split into two parts with a beam splitter. The first fraction is directly used as the Stokes beam while the second fraction is coupled into an optical parametric oscillator (OPO, APE, Berlin, Germany). The OPO allows tunability of the pump wavelength from 500 to 800 nm (SHG of signal wavelength), 1000-1600 nm (signal wavelength) and 1600-3200 nm (idler). Here, the OPO is tuned to 671 nm to match the CH_2 symmetrical stretching vibration at 2850 cm^{-1} for the CARS measurements. Both the pump and Stokes beams are spatially combined by a dichroic filter and temporally overlapped by adjustment of a mechanic delay stage. The combined laser beams are then coupled into a laser scanning microscope (LSM 510 Meta; Zeiss, Jena, Germany) and focused onto the sample with a 20× (NA 0.8) apochromatic objective (Zeiss). The nonlinear optical response of the sample is filtered from residual laser light by means of various dielectric filters and detected by photomultiplier tubes (PMT, Hamamatsu Photonics, Hamamatsu, Japan) in forward (CARS, SHG) and backward direction (TPEF). Large area scans of the samples of up to 15×15 tile-scans, each having a size of 450 μm × 450 μm, were recorded. For the tile-scan, a resolution of 1024×1024

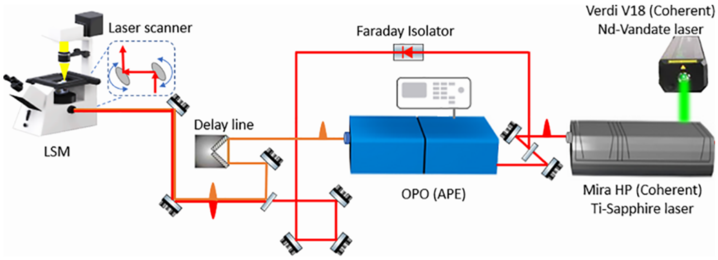


FIGURE 2 Schematic of experimental setup with the combination of three modalities; coherent anti-Stokes Raman scattering (CARS), two-photon excited fluorescence (TPEF) and second-harmonic generation (SHG) measurements

TABLE 1 Overview of the experimental parameters

Configuration	Excitation source	Central wavelength; FWHM, nm	Average power at sample, mW	Peak irradiance at sample, W/cm^2
CARS @ 2850 cm^{-1}	OPO (Pump) + Ti-Sapphire (Stokes)	671; 0:6 and 830; 0:5	~50 + 50	≈1.9x10^{10} ≈2.5x10^{10}
TPEF @ 435-485 nm	OPO (Pump) + Ti-Sapphire (Stokes)	671; 0:6 and 830; 0:5	~50 + 50	≈1.9x10^{10} ≈2.5x10^{10}
SHG @ 415 nm	Ti-Sapphire	830; 0:5	~50	≈2.5x10^{10}

Abbreviations: CARS, coherent anti-Stokes Raman scattering; FWHM, full width at half maximum; OPO, optical parametric oscillator; SHG, second-harmonic generation; TPEF, two-photon excited fluorescence.

pixels and a pixel dwell time of 1.6 μs were set. With an average of 4, the time per single tile does not exceed 16 seconds for CARS/TPEF/SHG. Therefore, the acquisition time for an image of 15 × 15 squares corresponding to a size of 6.75 mm × 6.75 mm is about 1 hour. The average power at the sample was adjusted to 50 and 50 mW for the pump and Stokes beam, respectively, in order to avoid photodamage [47]. The experimentally relevant parameters used in this study (see Table 1), are considered to cause negligible photodamage [46, 48].

2.3 | Software and computational analysis

All computations were carried out based on in-house written functions in the programing language Python version 3.6.5 and the statistical programing language R version 3.4.2.

2.4 | Image preprocessing and image patch extraction

The image analysis pipeline starts with image preprocessing. The aim of image preprocessing is to improve the image quality by suppressing unwanted distortions and enhance image features that improve the statistical analysis. In our work, a common image preprocessing pipeline was implemented for all parts of the multimodal images (see Figure 3A). It starts by down-sampling the image by a factor of two, followed by a median smoothing. The obtained smoothed images were afterwards corrected for the mosaicking artifacts produced by the uneven illumination of the image tiles [49]. Then the image contrast was adjusted based on the contrast limited adaptive histogram equalization algorithm (CLAHE) [50].

After the image preprocessing, the obtained multimodal images were compared with the corresponding annotated H&E images that describe different tissue regions within the patient tissues. The tissues considered were cancerous tissue, fat and normal breast tissue. Thereafter, each multimodal image was sliced into patches of size 512 × 512 pixels, and the patches that have only one specific tissue label were selected for further analysis. This procedure of patch extraction and selection is presented in Figure 3B, and it led to 1053 selected patches that were distributed as follows: 470 patches represented cancerous tissue, 143 patches showed fat tissue and 440 patches represented normal tissue.

2.5 | Patch classification

The selected patches from the multimodal images were utilized to check the quality of two ML algorithms for

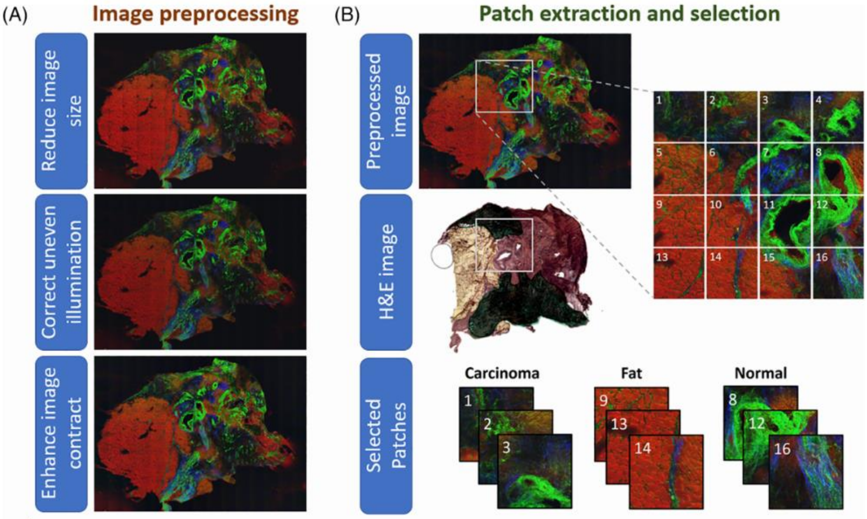


FIGURE 3 Multimodal image preprocessing and patch selection. A, The considered preprocessing pipeline starts by reducing the image size of a factor of 2. Then the mosaic artifacts caused by the uneven illumination are corrected. Finally, the contrast of the corrected images is adjusted using the CLAHE algorithm. B, The enhanced multimodal images are compared with the H&E annotated images and the suspicious regions are determined. Thereafter, patches of the size 512 × 512 pixels are extracted from each image, and patches with one label, that is, carcinoma, fat or normal, are selected to be used in detection of the breast carcinoma

detecting breast cancer tissue. These algorithms represent examples of the main two ML approaches: the classical machine learning approach and the deep learning approach. For the classical machine learning, the patch classification was accomplished using a linear discriminant analysis (LDA) model after extracting patch features while the deep convolutional neural network ResNet50 was utilized as a representative of the deep learning approach. The ResNet50 network is a publicly available fully connected convolutional neural network (CNN) that was trained on the ImageNet dataset [45]. In our contribution, the pretrained ResNet50 network was fine-tuned to suit the mentioned multiclass classification task based on the multimodal images. This tuning was accomplished on the basis of additional layers on the top of the ResNet50 network (see Figure 4) [51]. These layers were batch normalization layers, a global average pooling layer, dropout layers with the probability of 50%, a dense layer with 32 neurons and a SoftMax activation layer, which provides a label probability for each patch. The main advantages of using batch normalization and dropout layers are to prevent model over fitting and to let each layer of the network learn more independently by itself [52, 53] while adding an additional dense layer allows the network to find the combinations of the already learnt features that improve objects recognition in our new dataset [54].

In our work, the ResNet50 network was implemented twice: as a feature extractor for the PCA-LDA model and as a classification model. In the first case the pertained model was kept stable and the shelf features were extracted, which were used by the PCA-LDA model for classification. In order to use the ResNet50 as classification model we added some layers and fine-tuned the model using the multimodal images (see Figure 4 for details). If the ResNet50 model was only used as feature extractor we call the model pretrained ResNet50 network and characterize it as classical machine learning while it was termed fine-tuned ResNet50 network and characterized as deep learning model, if the ResNet50 model was fine-tuned and used as a classification model. As the feature extraction and learning are time and memory consuming for the large image size, we decided to resize the selected patches again using down-sampling of a factor of four. The ResNet50 network was fed with the obtained resized patch for both implementations of the ResNet50 network.

3 | RESULTS

In this section, we present the results of an automatic detection of breast cancer based on multimodal images of

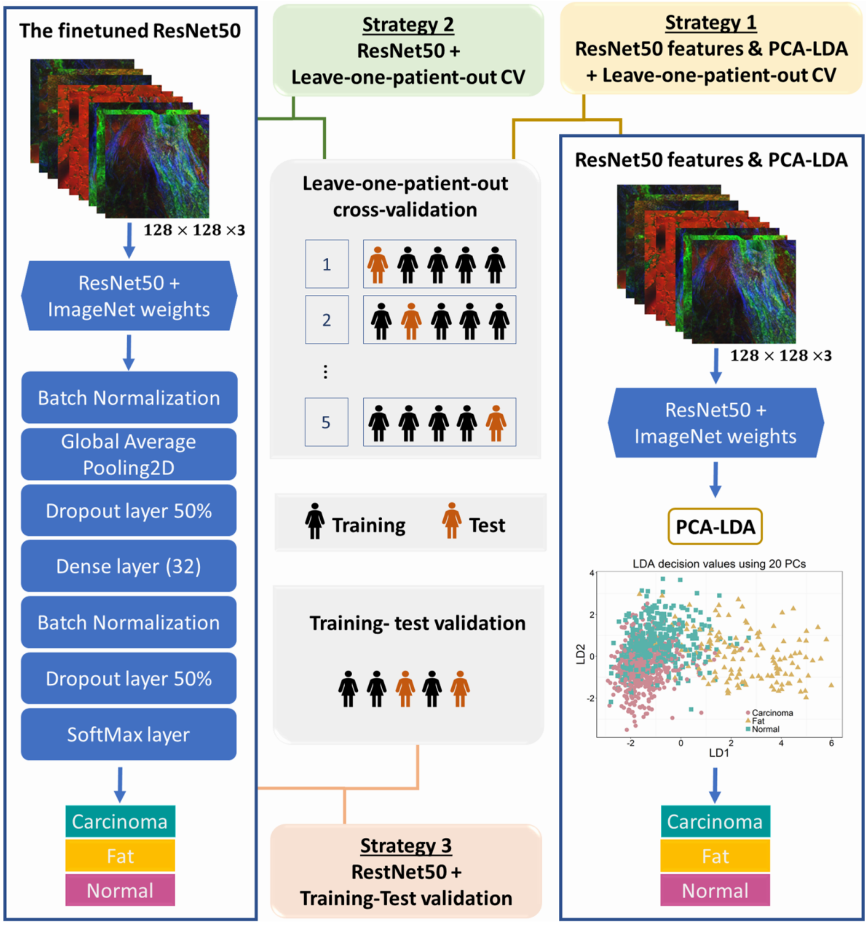

The finetuned ResNet50
128 × 128 ×3
ResNet50 + ImageNet weights
Batch Normalization
Global Average Pooling2D
Dropout layer 50%
Dense layer (32)
Batch Normalization
Dropout layer 50%
SoftMax layer
Carcinoma
Fat
Normal
Strategy 2
ResNet50 + Leave-one-patient-out CV
Strategy 1
ResNet50 features & PCA-LDA + Leave-one-patient-out CV
Leave-one-patient-out cross-validation
1
2
⋮
5
Training
Test
Training- test validation
Strategy 3
RestNet50 + Training-Test validation
ResNet50 features & PCA-LDA
128 × 128 ×3
ResNet50 + ImageNet weights
PCA-LDA
LDA decision values using 20 PCs
LD2
LD1
Carcinoma
Fat
Normal
Carcinoma
Fat
Normal

FIGURE 4 A schematic diagram of the utilized classification and validation strategies. The pretrained ResNet50 network is either utilized as a feature extractor or finetuned to be utilized directly as a classification model. In the first implementation of the ResNet50, the extracted features are projected using a PCA model, then an LDA model is utilized to classify the image patches. The validation of this PCA-LDA model is accomplished by LOPO-CV which performs the first classification strategy. In the second implementation of the ResNet50 network, different layers are added on top of this network. Thereafter, the LOPO-CV and train-test validation are combined with this finetuned network to perform the other two classification strategies

breast tissue. The studied dataset consists of 21 breast tissue biopsies collected from 21 different patients. Due to the mentioned issues in section 2.1 regarding the staining and annotation, only 15 multimodal images with their corresponding H&E stain images were involved to learn and validate the utilized machine learning algorithms. Although, the ground truth of the remaining five multimodal images is unknown, we utilized the best classification model to predict the patches of these six images as test set. Based on the 15 well annotated H&E images, the selected patches of the corresponding multimodal images were utilized to construct a classification model for breast cancer detection. To do so, we checked different classification and validation strategies. Figure 4 shows a schematic overview of the considered classification and validation techniques. For all presented strategies, the statistical independence between the training, validation and test sets was secured based on the following rule:

The patient patches that are utilized to train a classifier are completely different of the patients that are utilized to validate or test the learned classifier. Nevertheless, we compared two different validation methods in this work. These validation methods are the leave-one-patient-out cross-validation (LOPO-CV) and training-test validation. The LOPO-CV can be described simply by fixing patches of one patient as a test set, then building and validating the classifier using the patches of the remaining patients from the studied dataset. This procedure is repeated for all patients to be a test set once, and the patch labels of this test set are predicted within each iteration using the learned classification model. In the second validation method namely training-test validation, the dataset is divided into three subsets: training set, validation set and test set. Using the training set, we build the classification model, then we optimize its parameters using the validation set. After that, we examine the constructed classifier

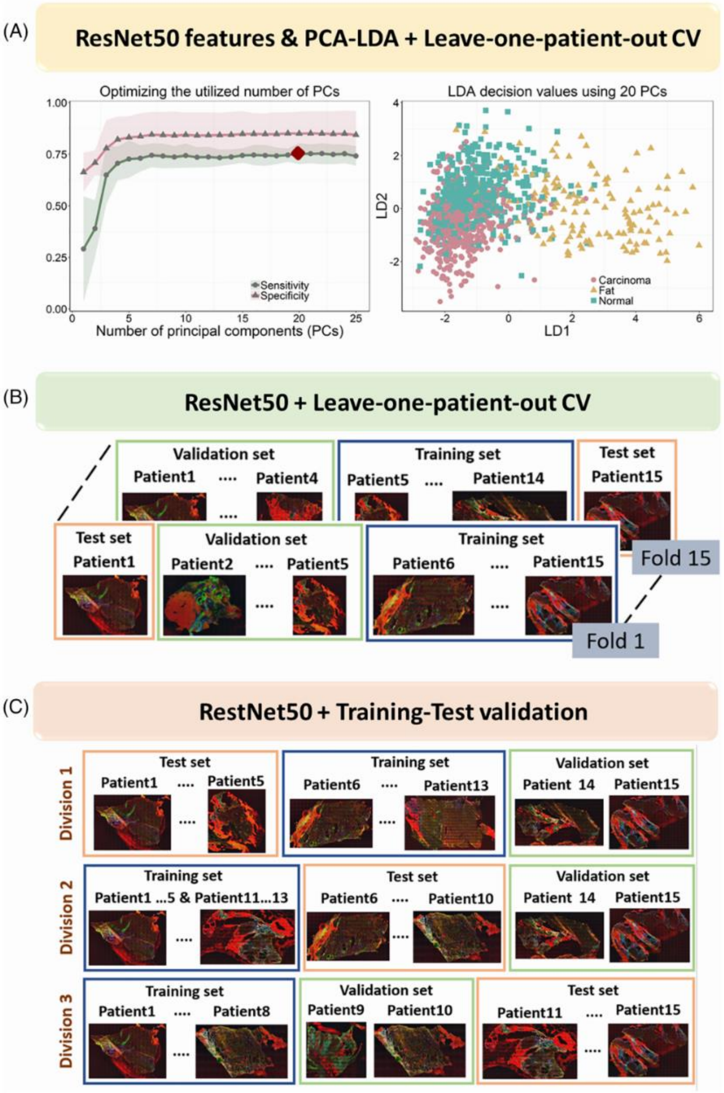


FIGURE 5 Overview of the implemented classification strategies. A, The pretrained ResNet50 network is utilized to extract the patch features, then a PCA-LDA model is utilized within the cross-validation loop for dimension reduction and patch classification. B, The fine-tuned ReNet50 network in combination with LOPO-CV are utilized to predict the labels of patient patches. Here, patches of 10 patients are always utilized to train the network and patches from other four patients are used to validate the results. Then the labels of the remaining patient patches are predicted. C, Three different data divisions are checked using the ResNet50 network. In all these data divisions, the fine-tuned ResNet50 network is trained on patches extracted from eight patients and validated on patches from other two patients. After that, the labels of the remaining five patients are predicted

on the test set. Also in this validation method, the utilized patient patches for training, validating, or even testing a classifier are completely independent. In Figure 5B, C, the applied data division techniques using LOPO-CV and training-test validation are presented.

In addition to the validation methods, we compared the classification results using an LDA model with the results of a fine-tuned convolutional neural network ResNet50. For the first classification model, the ResNet50 network was implemented as an image feature extractor only. Then a combination of a principal component analysis model and a linear discriminant analysis (PCA-LDA) was utilized to reduce the high dimensionality of the obtained ResNet50 feature matrix and to differentiate between the tissue patches. Moving to the second classification model, the fully connected network ResNet50 was fine-tuned as it was explained in subsection 2.5. Like described earlier all models were evaluated using training-test validation and LOPO-CV. For both validation methods, the ResNet50 network was trained for the same hyperparameters while the network parameters were optimized using a backpropagation algorithm. The fine-tuned network was trained for 20 epochs using the Adam optimizer with a learning rate of 0.001 and categorical cross entropy as loss function.

In the following, we present the results of the patch classification by the previous introduced combinations of classification and validation techniques, which are sketched in Figure 4. In Subsection 3.1 we compared the results of both classification models (pretrained ResNet50 in combination with PCA-LDA and fine-tuned ResNet50) using the same evaluation method; namely a LOPO-CV. In the subsection 3.2 the difference between the estimated performance of a fine-tuned ResNet50 utilizing the LOPO-CV and training-test validation was investigated. The results of these classification and validation strategies were compared based on the classification mean sensitivity or the sensitivity of a cancer diagnostic model.

3.1 | A comparison between the classification techniques using LOPO-CV

In this subsection we compare the classification results of two machine learning approaches, the LDA model and the finetuned ResNet50 network, based on the LOPO-CV. Using the first classification method, that is, PCA-LDA, the ResNet50 features were extracted for all reduced image patches. This extraction produced in total 16 384 features per patch. Since the number of these extracted features is much larger than the number of patients, a principle component analysis model was combined with an LDA model to reduce the high dimensionality and to optimize the classification results. In Figure 5A, the selection method for the optimal number of principal components (PCs) is shown. Beside this, the scattering plot of the LDA decision values using 20 PCs is plotted. The optimization of the utilized number of PCs was done based on the highest mean sensitivity of the PCA-LDA model and LOPO-CV. It turned out that using 20 PCs for constructing the LDA model provided the highest mean sensitivity. Moving to the scattering plot in Figure 5A, it is observed that the fat tissue group was well separated from the normal and cancerous tissues while this tissue separation decreased significantly for the differentiation between normal and cancer tissue patches. These results can be also seen in the obtained confusion table of the PCA-LDA model (Table 2). In this table, the PCA-LDA classification results in addition to the sensitivity and specificity for each class are presented. The highest sensitivity and specificity were observed for the fat tissue followed by the cancerous tissue. Nevertheless, the mean sensitivity of the PCA-LDA model based on LOPO-CV is around 75.31%, and the mean specificity is 85.05%.

After evaluating the classification results using the PCA-LDA, we checked the results using the fine-tuned ResNet50 network when it was inserted within the LOPO-CV loop. For each iteration, patches of one patient were defined as test set while patches of four patients and patches of 10 patients were utilized to train and validate the considered network, respectively. Because the number of training set is quite small, the training patches were augmented by using random rotations by steps of 10° degrees within the range of 0° to 180° (see Figure 6). This strategy is a common technique to increase artificially the training dataset size and to prevent over fitting caused by the huge number of parameters of the utilized convolutional neural network. The next step after data

TABLE 2 The confusion table of the PCA-LDA model and the leave-one-patient-out cross-validation[a]

	True		
Prediction	**Carcinoma**	**Fat**	**Normal**
Carcinoma	342	5	105
Fat	9	116	18
Normal	119	22	317
Sensitivity	0.7277	0.8112	0.7205
Specificity	0.8113	0.9703	0.7700

[a] The PCA-LDA model was build using 20 PCs. The best separation was provided for fat tissue followed by the cancer tissue.

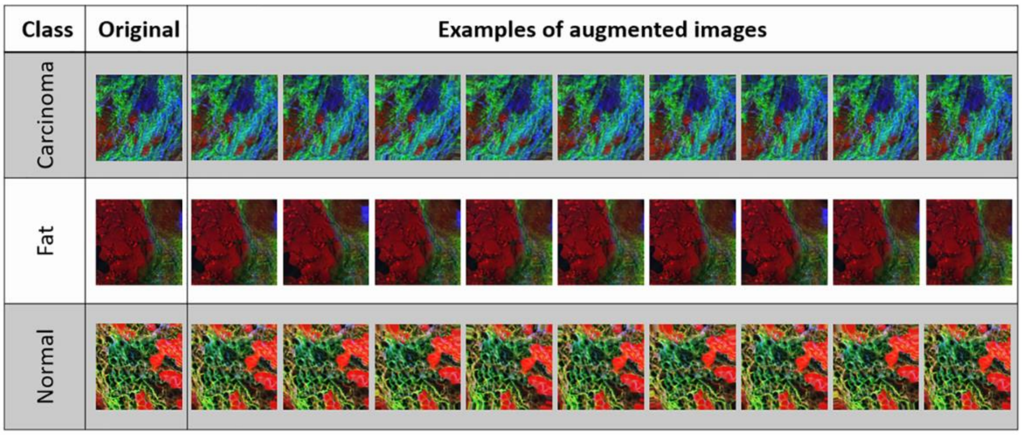


FIGURE 6 Image augmentation. All the patches of the training set are rotated in 10° steps within the range of 0° to 180°

augmentation was to train the ResNet50 network and to predict the patch labels of the considered test set. The obtained classification results by the fine-tuned ResNet50 network and LOPO-CV are presented in Table 3. It is observed that the ResNet50 network was more sensitive to the differences between the tissue labels within our dataset. This can be seen in the obtained confusion table of the ResNet50 network. From Table 3, the mean sensitivity and mean specificity of the fine-tuned ResNet50 network and LOPO-CV is 86.23% and 91.31%, respectively.

To compare the previous utilized machine learning algorithms, we evaluated the model quality based on the identification of cancerous tissues. This evaluation was done after combining the patches of normal and fat tissues in one category representing the noncancer tissues, then calculating the mean sensitivity of the PCA-LDA model and the ResNet50 network to be considered as cancer diagnostic model. The obtained results of both models with respect to patient patches are summarized in Table 4. The second and the third columns of this table represent the number of extracted patches from non-cancerous and cancerous tissue per patient while the forth and the fifth columns shows the mean sensitivity of the binary classification results using the PCA-LDA and the ResNet50 network, respectively. Form most individual patient results, the cancer identification based on the finetuned ResNet50 network was much better compared to the obtained results using the PCA-LDA model. This improvement was detected as an increase in the mean sensitivity of the ResNet50 network for most patients. In this context, the mean sensitivity of the cancer diagnostic model was larger than 11% if the ResNet50 network was utilized as a classification model. Nevertheless, the obtained mean sensitivity of the cancer diagnostic model is 73.33% using the PCA-LDA model, and it is 84.50% for the ResNet50 network.

TABLE 3 The confusion table of a ResNet50 model and the leave-one-patient-out cross-validation[a]

	True		
Prediction	**Carcinoma**	**Fat**	**Normal**
Carcinoma	356	3	29
Fat	6	130	6
Normal	108	10	405
Sensitivity	0.7574	0.9091	0.9205
Specificity	0.9451	0.9868	0.8075

[a] The ResNet50 networks could predict the normal and fat tissue quite well while this detection decreased for the cancerous tissue.

3.2 | Studying the influence of the validation method on the results of the pretrained ResNet50

The aim of this part is to study the influence of the data division and validation methods on the classification results of the ResNet50 network. Therefore, we compared the classification sensitivity using the two presented validation methods: The LOPO-CV and the training-test validation. By training-test validation, the studied dataset was partitioned on the patient level with the ratio of 8:2:5 into training set, validation set and test set, respectively.

TABLE 4 The cancer diagnostic model[a]

Patient ID	Cancer	Noncancer	Mean Sens. PCA-LDA	Mean Sens. ResNet50
1	58	14	0.7771	0.8485
2	16	38	0.7336	0.7237
3	22	48	0.6878	0.4737
4	6	34	0.9265	0.8578
5	58	20	0.7991	0.6578
6	44	28	0.5260	0.9432
7	43	36	0.7442	0.7513
8	26	63	0.8608	1.0000
9	14	62	0.7604	1.0000
10	14	12	0.9226	0.8929
11	15	25	0.5267	0.8267
12	21	34	0.5581	1.0000
13	71	71	0.7254	0.8803
14	33	54	0.8510	0.9907
15	29	54	0.6009	0.8292
Total	470	143	Mean = 0.7333	Mean = 0.8450

[a] For each patient, the mean sensitivity of the binary classification (cancer vs noncancer) is calculated using the PCA-LDA model and the ResNet50 network. For most patients, the ResNet50 network introduced better diagnostic results than the PCA-LDA model.

TABLE 5 The confusion matrices of the ResNet50 network trained by different patient subsets using the training-test data validation[a]

		True				
Division	Prediction	Carcinoma	Fat	Normal	Sensitivity	Specificity
1	Carcinoma	88	8	15	0.5500	0.8403
	Fat	0	32	4	0.5517	0.9837
	Normal	72	18	67	0.7791	0.5872
2	Carcinoma	40	0	23	0.2837	0.8856
	Fat	2	14	8	0.2456	0.9649
	Normal	99	43	113	0.7847	0.2828
3	Carcinoma	128	0	106	0.7574	0.5546
	Fat	0	24	7	0.8571	0.9815
	Normal	41	4	97	0.4619	0.7716

[a] The identification of the normal tissue based on the data division 1 and 2 was good while only the third data division could detect cancer and fat tissue in good manner.

This data division was accomplished based on different selections of the data subsets. Thus, we checked three different cases of the data division as described in Figure 5C. Herein, the finetuned ResNet50 network was trained on patches of eight patients and validated on patches of the other two patients. The tissue patches of the remaining five patients were utilized as test set. Similar to the previous validation method, that is, LOPO-CV, the training patches were rotated randomly using multiples of 10° rotation angles within the range from 0° to 180° (see Figure 6). After training and validating the network, the obtained classification results of the test sets were summarized in Table 5. For normal tissue, the ResNet50 network based on the first and the second data divisions could predict the tissues labels in a good manner with a sensitivity of 77.91% and 78.47%, respectively. But the prediction quality decreased into the sensitivity of 46.19% when the considered network was trained and validated using the third case of data division. Moving to the prediction of fat tissue, we can see different results if the fine-tuned network was trained on different training sets. The ResNet50 network based on third data division provided the best identification results with a sensitivity of 85.71%. For the diagnosis of cancerous tissue, when the fine-tuned ResNet50 network was trained on the first and second data division, poor cancer detection results were introduced. However, using the third case of data division to train and validate the ResNet50 network provided much better identification of breast cancer patches. This improvement in breast cancer identification was characterized by at least 20% increase in the classification sensitivity if the ResNet50 network was trained and validated on third data division. Nonetheless, the classification mean sensitivity of the pretrained ResNet50 network using the first, the second and the third data division was around 62.69%, 43.80% and 69.21%, respectively.

To summarize this part, the classification performance of the fine-tuned ResNet50 network was influenced strongly by the method of data division and validation. Moreover, completely different results were obtained for each case of the data division and the data validation. Reasons that may produce this large variation in the obtained results of the ResNet50 network are as follows. One of these reasons refer to the problem of small training set size which represents the number of patients included in our study. In this training-test validation, patches of 15 patients were included, and only the patches from eight patients were used to train the ResNet50 network. Beside the small training size, we expect that the biological variation of the training patches is not always consistent with the variation within the validation or test sets which produces challenges to detect new cases. Another reason can arise from the huge number of hyperparameters utilized by the network. A solution, for the previous mentioned issues caused by the training-test validation and small datasets size, can be the use of by cross-validation, where more patients can be utilized for training the classifier. In the Subsection 3.1, patches of 10 patients were always utilized to train the ResNet50 network. Therein, the obtained classification mean sensitivity based on the LOPO-CV is 86.23% which dropped down, at least, 15% when we applied training-test validation. Nevertheless, it was observed that the results of training-test validation are fluctuating a lot in comparison to LOPO-CV. This is due to the utilized mean sensitivity in the LOPO-CV which introduced more robust classification model to be utilized in the prediction of new cases.

As we could see, the classification models and validation methods were studied only for the multimodal images which have corresponding annotated H&E images. However, to introduce the presented nonlinear imaging technology into clinical daily routine, the efficiency of cancer diagnosis based on the multimodal images needs to be proven for new patients without using the H&E annotation. In our study, the ground truth annotation of six multimodal images was missing. Therefore, we decided to test the efficiency of classification models for predicting the patch labels using the six images even though the annotated H&E stain images are not available yet. This prediction was done using the best classification model constructed based on the 15 well annotated images. Among all presented classification and validation strategies, the best classification performance was

TABLE 6 The predication results of multimodal images that do not have corresponding H&E annotated images[a]

	The ResNet50 model		
Prediction	**8**	**9**	**12**
Cancer	123	103	153
Noncancer	691	711	661

[a]The best three ResNet50 models were utilized to detect patch labels of unknown tissue regions.

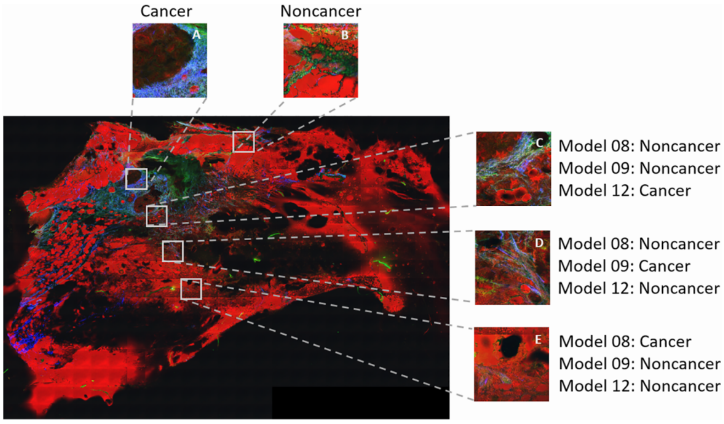


FIGURE 7 An example of patch prediction for a not annotated image. The best ResNet50 models based on the LOPO-CV are utilized to detect breast cancer patches. The patches A and B are predicted the same using the ResNet50 models 8, 9 and 12 while different prediction results are obtained for patches C, E and D

achieved by the fine-tuned ResNet50 network and the LOPO-CV. For each iteration of the CV loop, the best ResNet50 model was saved and the patch labels of test patients were predicted. Table 4 shows that cancer detection results of the unannotated images. Here, the best cancer identification was introduced by the ResNet50 network model that was tested on patient 8, 9 and 12. The sensitivity of cancer diagnostic model for all these test patients was 100%. Thus, we utilized these ResNet50 models to predict the patch labels of the six test patients. To do so, around 814 patches were selected then their labels were predicted. The obtained results of cancer patch prediction are presented in Table 6. The results differed between the ResNet50 models and these differences are mostly caused according to main two expected reasons. The first one is produced by the patches which have more than one label; for instance, some patches may have cancer and noncancer tissue together which confuses the classifier. In addition to the double patch labeling, each ResNet50 model (the model 8, 9 and 12) was trained and validated on 10 patients and validated on four patients. This means, the model which learnt from 10 patients, might have different features in comparison to the ResNet50 model that was trained on other 10 patients. Nevertheless, with respect to the small training data size and the two other mentioned issues, the results of patch prediction using the three ReNet50 models are still close to each other. Figure 7 displays an example of this patch prediction using the three ResNet50 models. Therein, all ResNet50 models provided the same tissue prediction for patches A and B, which was not the case for the patches C, D and E. However, these results need to be validated again by the annotated H&E stain images which are still not available for this study.

4 | SUMMARY AND CONCLUSION

We presented in this paper the results of a label-free breast cancer detection based on a small dataset size consisting of 21 images collected from 21 patients. These images were obtained using a combination of three nonlinear imaging modalities CARS, TPEF and SHG. This imaging combination provides diagnostic relevant information from breast cancer tissues. The main challenge was to translate this biomolecular information into a ML model that can be used in further studies. Thereby, an image preprocessing pipeline was designed and two classification models were utilized to detect breast cancer regions. Our preprocessing pipeline started with resizing the multimodal images into the half size followed by correcting the mosaic artifacts arising from the uneven illumination. The last step was to enhance the contrast which was accomplished using the contrast limited adaptive histogram equalization algorithm. Thereafter, the preprocessed images were compared with the corresponding annotated H&E stain images. Using this comparison and with respect to the 15 annotated H&E stain images, patches from tissue regions that have only one label were extracted. In total, 1053 patches of normal, fat and cancerous tissues and 15 multimodal images were included to train and validate the studied classifiers.

After the image preprocessing and patch selection, the detection of the breast cancerous tissue was accomplished using three strategies. These strategies represent different combinations between two ML algorithms a fine-tuned ResNet50 network and PCA-LDA model) and two data validation methods (LOPO-CV and the training-test validation). For all implemented strategies, the deep convolutional neural network ResNet50 was utilized either to extract image features from the patches or to detect directly the cancerous regions. The utilized ResNet50 network is a publicly available fully connected convolutional neural network that was trained on the ImageNet dataset. The results of the presented classification and validation strategies were evaluated based on the classification mean sensitivity for a three-class model and the binary cancer diagnostic model. This cancer diagnostic model characterizes the model quality of cancerous tissues. It turned out that the best detection of cancerous tissues was achieved by the fine-tuned ResNet50 network and the LOPO-CV. Thereby, the mean sensitivity of LOPO-CV using the fine-tuned ResNet50 network is 86.23% which decreased to 75.31% if a PCA-LDA model was implemented. Using the training-test data validation, the mean sensitivity was strongly influenced by the chosen data subsets, that is, training set, validation set and test set. In this case, the classification mean sensitivity varied between 43.80% and 69.21%.

In the last part of our study, the best classification model was tested for detecting cancerous and noncancerous tissue of the remaining six multimodal images. The challenge of these images is that the H&E stain images were not available. However, using the best classification models obtained by the fine-tuned ResNet50 network and the LOPO-CV, around 814 patches were selected then their labels were predicted. In most cases the classification models provide the same predictions of the patches.

To conclude, the combination of the three nonlinear imaging modalities; namely CARS, TEPF and SHG, provided a label-free cancer detection tool which showed its efficiency in diagnosing breast cancer tissue based on a small sample size of patients. This efficiency was demonstrated via a computer aided diagnosis using machine learning, specifically the deep convolutional neural network ResNet50. Nevertheless, the noninvasive nature of imaging technique enables for further in vivo measurements which offers a low-risk diagnostic approach to supplement mammography in an optical biopsy approach.

ACKNOWLEDGMENTS

The funding from the DFG for the projects BO 4700/1-1, PO 563/30-1 and STA 295/11-1 is highly appreciated. Financial support by the BMBF for the project Uro-MDD (FKZ 03ZZ0444J) is highly acknowledged. The infrastructure used for the research was funded by the ministry of science and education of Thuringia (TMBWK) grant numbers (2015FOR-0001 and 2015-0023). Funding from the Free State of Thuringia and cofounding of the Europäischen Fonds für regionale Entwicklung (EFRE) for the projects In-SpekT (2016 FGI 0011) und IR-Photonik (2017 FGI 0018) are gratefully acknowledged.

ORCID

Michael Schmitt https://orcid.org/0000-0002-3807-3630
Thomas Bocklitz https://orcid.org/0000-0003-2778-6624

REFERENCES

[1] M. R. Ataollahi, J. Sharifi, M. R. Paknahad, A. Paknahad, *J. Med. Life* **2015**, *8*(Spec Iss 4), 6.

[2] M. Milosevic, D. Jankovic, A. Milenkovic, D. Stojanov, *Technol. Health Care* **2018**, *26*(4), 729.

[3] World Health Organization (WHO). *Breast Cancer*, https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed: 30 July 2019).

[4] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, A. Jemal, *CA Cancer J. Clin.* **2018**, *68*(6), 394.

[5] A. Migowski, *Cien. Saude Colet.* **2015**, *20*(4), 1309.

[6] B. N. Hellquist, K. Czene, A. Hjalm, L. Nystrom, H. Jonsson, *Cancer* **2015**, *121*(2), 251.

[7] T. Onega, L. E. Goldman, R. L. Walker, D. L. Miglioretti, D. S. Buist, S. Taplin, B. M. Geller, D. A. Hill, R. Smith-Bindman, *J. Med. Screen.* **2016**, *23*(1), 31.

[8] L. E. Duijm, G. L. Guit, J. O. Zaat, A. R. Koomen, D. Willebrand, *Br. J. Cancer* **1997**, *76*(3), 377.

[9] K. M. Kelly, J. Dean, W. S. Comulada, S. J. Lee, *Eur. Radiol.* **2010**, *20*(3), 734.

[10] N. Ozmen, R. Dapp, M. Zapf, H. Gemmeke, N. V. Ruiter, K. W. van Dongen, *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **2015**, *62*(4), 637.

[11] L. Wang, *Sensors (Basel)* **2017**, *17*(7), 1572.

[12] C. D. Lehman, M. D. Schnall, *Breast Cancer Res.* **2005**, *7* (5), 215.

[13] A. M. Hassan, M. El-Shenawee, *IEEE Rev. Biomed. Eng.* **2011**, *4*, 103.

[14] D. Roganovic, D. Djilas, S. Vujnovic, D. Pavic, D. Stojanov, *Bosn. J. Basic Med. Sci.* **2015**, *15*(4), 64.

[15] J. M. Lee, E. F. Halpern, E. A. Rafferty, G. S. Gazelle, *Acad. Radiol.* **2009**, *16*(11), 1323.

[16] X. He, L. Sun, Y. Huo, M. Shao, C. Ma, *Q. J. Nucl. Med. Mol. Imaging* **2017**, *61*(4), 429.

[17] S. Qi, S. Hoppmann, Y. Xu, Z. Chen, *Mol. Imaging Biol.* **2019**, 21(5), 907-916

[18] E. L. Rosen, W. B. Eubank, D. A. Mankoff, *Radiographics* **2007**, *27*(Suppl 1), S215.

[19] S. K. Yang, N. Cho, W. K. Moon, *Korean J. Radiol.* **2007**, *8*(5), 429.

[20] A. Lukic, S. Dochow, H. Bae, G. Matz, I. Latka, B. Messerschmidt, M. Schmitt, J. Popp, *Optica* **2017**, *4*(5), 496.

[21] O. Chernavskaia, S. Heuke, M. Vieth, O. Friedrich, S. Schürmann, R. Atreya, A. Stallmach, M. F. Neurath, M. Waldner, I. Petersen, M. Schmitt, T. Bocklitz, J. Popp, *Sci. Rep.* **2016**, *6*, 29239.

[22] T. Meyer, N. Bergner, C. Bielecki, C. Krafft, D. Akimov, B. F. Romeike, R. Reichart, R. Kalff, B. Dietzek, J. Popp, *J. Biomed. Opt.* **2011**, *16*(2), 021113.

[23] 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, J. Popp, *Head Neck* **2016**, *38*(10), 1545.

[24] T. Meyer, O. Guntinas-Lichius, F. von Eggeling, G. Ernst, D. Akimov, M. Schmitt, B. Dietzek, J. Popp, *Head Neck* **2013**, *35* (9), E280.

[25] X. Xu, J. Cheng, M. J. Thrall, Z. Liu, X. Wang, S. T. Wong, *Biomed. Opt. Express* **2013**, *4*(12), 2855.

[26] V. DeGiorgi, D. Massi, S. Sestini, R. Cicchi, F. S. Pavone, T. Lotti, *J. Eur. Acad. Dermatol. Venereol.* **2009**, *23*(3), 314.

[27] S. Heuke, N. Vogler, T. Meyer, D. Akimov, F. Kluschke, H. J. Rowert-Huber, J. Lademann, B. Dietzek, J. Popp, *Healthcare (Basel)* **2013**, *1*(1), 64.

[28] N. Vogler, T. Meyer, D. Akimov, I. Latka, C. Krafft, N. Bendsoe, K. Svanberg, B. Dietzek, J. Popp, *J. Biophotonics* **2010**, *3* (10–11), 728.

[29] A. Diaspro, G. Chirico, M. Collini, *Q. Rev. Biophys.* **2005**, *38* (2), 97.

[30] B. G. Wang, K. Konig, K. J. Halbhuber, *J. Microsc.* **2010**, *238* (1), 1.

[31] W. Mohler, A. C. Millard, P. J. Campagnola, *Methods* **2003**, *29* (1), 97.

[32] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, N. Karssemeijer, *Med. Image Anal.* **2017**, *35*, 303.

[33] I. Arganda-Carreras, V. Kaynig, C. Rueden, K. W. Eliceiri, J. Schindelin, A. Cardona, H. Sebastian Seung, *Bioinformatics* **2017**, *33*(15), 2424.

[34] M. A. Marchetti, N. C. F. Codella, S. W. Dusza, D. A. Gutman, B. Helba, A. Kalloo, N. Mishra, C. Carrera, M. E. Celebi, J. L. DeFazio, N. Jaimes, A. A. Marghoob, E. Quigley, A. Scope, O. Yelamos, A. C. Halpern, *J. Am. Acad. Dermatol.* **2018**, *78* (2), 270.

[35] M. Motwani, D. Dey, D. S. Berman, G. Germano, S. Achenbach, M. H. Al-Mallah, … P. J. Slomka, *Eur. Heart J.* **2016**, *38*(7), 500.

[36] H. O. Alanazi, A. H. Abdullah, K. N. Qureshi, *J. Med. Syst.* **2017**, *41*(4), 69.

[37] K. Kong, C. J. Rowlands, S. Varma, W. Perkins, I. H. Leach, A. A. Koloydenko, H. C. Williams, I. Notingher, *Proc. Natl. Acad. Sci.* **2013**, *110*(38), 15189.

[38] C. L. Chen, A. Mahjoubfar, L.-C. Tai, I. K. Blaby, A. Huang, K. R. Niazi, B. Jalali, *Sci. Rep.* **2016**, *6*, 21471.

[39] T. Falk, D. Mai, R. Bensch, O. Cicek, A. Abdulkadir, Y. Marrakchi, A. Bohm, J. Deubner, Z. Jackel, K. Seiwald, A. Dovzhenko, O. Tietz, C. Dal Bosco, S. Walsh, D. Saltukoglu, T. L. Tay, M. Prinz, K. Palme, M. Simons, I. Diester, T. Brox, O. Ronneberger, *Nat. Methods* **2019**, *16*(1), 67.

[40] M. Habibzadeh, M. Jannesari, Z. Rezaei, H. Baharvand, M. Totonchi, in *Tenth Int. Conf. on Machine Vision, SPIE*, Vol. 10696, **2018**.

[41] P. Pradhan, T. Meyer, M. Vieth, A. Stallmach, M. Waldner, M. Schmitt, J. Popp, T. Bocklitz, in Semantic Segmentation of Non-linear Multimodal Images for Disease Grading of Inflammatory Bowel Disease: A SegNet-based Application, 2019, pp. 396–405.

[42] O. Ronneberger, P. Fischer, T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, Springer International Publishing., Cham **2015**.

[43] A. Kumar, J. Kim, D. Lyndon, M. Fulham, D. Feng, *IEEE J. Biomed. Health Inform.* **2017**, *21*(1), 31.

[44] E. Rodner, T. Bocklitz, F. von Eggeling, G. Ernst, O. Chernavskaia, J. Popp, J. Denzler, O. Guntinas-Lichius, *Head Neck* **2019**, *41*(1), 116.

[45] K. He, X. Zhang, S. Ren, J. Sun, **2016**. 770–778.

[46] S. Heuke, N. Vogler, T. Meyer, D. Akimov, F. Kluschke, H. J. Rowert-Huber, J. Lademann, B. Dietzek, J. Popp, *Br. J. Dermatol.* **2013**, *169*(4), 794.

[47] F. Fischer, B. Volkmer, S. Puschmann, R. Greinert, W. Breitbart, J. Kiefer, R. Wepf, *J. Biomed. Opt.* **2008**, *13*(4), 041320.

[48] H. G. Breunig, R. Buckle, M. Kellner-Hofer, M. Weinigel, J. Lademann, W. Sterry, K. Konig, *Microsc. Res. Tech.* **2012**, *75* (4), 492.

[49] F. B. Legesse, O. Chernavskaia, S. Heuke, T. Bocklitz, T. Meyer, J. Popp, R. Heintzmann, *J. Microsc.* **2015**, *258*(3), 223.

[50] A. M. Reza, *J. VLSI Signal Process. Syst. Signal Image Video Technol.* **2004**, *38*(1), 35.

[51] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv 1409.1556, **2014**.

[52] S. Ioffe, C. Szegedy, in *Proc. 32nd Int. Conf. on Int. Conf. on Machine Learning, Volume 37*, JMLR.org, Lille, France, 2015, pp. 448–456.

[53] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, *J. Mach. Learn. Res.* **2014**, *15*(1), 1929.

[54] G. Huang, Z. Liu, L. van der Maaten, K. Weinberger, Connected Convolutional Networks, 2017.

How to cite this article: Ali N, Quansah E, Köhler K, et al. Automatic label-free detection of breast cancer using nonlinear multimodal imaging and the convolutional neural network ResNet50. *Translational Biophotonics*. 2019;1:e201900003. https://doi.org/10.1002/tbio.201900003

Conferences & Workshops

I. Conferences

Conference Name	**Date**
DOKDOK 2016	25 – 29 Sep. 2016
7th International Chemometrics Research Meeting ICRM 2017	10 – 14 Sep. 2017
DOKDOK 2017	18 – 22 Sep. 2017
Life meets light conference 2017	18 Oct. 2017
ESMI Imaging Technology Summer Workshop	9 – 14 July 2018
Life meets light conference 2018	05 – 06 Sep. 2018
The International Symposium Image-based Systems Biology 2018	06 – 07 Sep. 2018
DOKDOK 2018	17 – 21 Sep. 2018
49. Kongress der Deutschen Gesellschaft für Endoskopie und Bildgebende Verfahren	28. - 30. March 2019
Bunsentagung 2019	28 – 31 May 2019

II. Workshops

Workshop title	**Date**
Effective Presentation, self-management, and creativity	25 – 29 Sep. 2016
Good scientific practice	10 – 11 July 2017

Acknowledgements

I would like to express my sincere gratitude to all people who have contributed to let this work be completed. I would like to thank Prof. Dr. Jürgen Popp for giving me the opportunity to be part of his great research group. My deepest thanks go to PD. Dr. Thomas Bocklitz for all his invaluable support and direct supervision during the Ph.D. time.

I would like to acknowledge Dr. Petra Rösch, Dr. Tobias Meyer-Zedler, Dr. Joren Jansen, Prof. Dr. Maximilian C. Kriegmair, Prof. Dr. Ute Neugebauer, and Prof. Dr. Michael Schmitt for their contributions to this thesis regarding imparting their knowledge and experience.

My special gratitude goes to my colleagues in the department of photonics data science Rola, Darina, Pranita, Pegah and Oleg for all joyful time and conducive working environment they have created. Also, I would like to thank my friends Rahaf, Lama, and Maha and all other colleagues and friends form IPHT & IPC for all their valuable support during my Ph.D. time.

Lastly, I want to express great thanks to my parents and my siblings Nabil, Nawar, Esraa, and Ahmad for their helping hands and engorgements they are always providing me.

Erklärungen

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und unter Verwendung der angegebenen Hilfsmittel, persönlichen Mitteilungen und Quellen angefertigt habe.

Name der Verfasserin	Datum	Ort	Unterschrift

Erklärung zu den Eigenanteilen der Promovendin sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an den Publikationen und Zweitpublikationsrechten bei einer kumulativen Dissertation (in die kumulative Dissertation aufzunehmen**).**

Für alle in dieser kumulativen Dissertation verwendeten Manuskripte liegen die notwendigen Genehmigungen der Verlage ("Reprint permissions") für die Zweitpublikation vor.

Die Co-Autoren der in dieser kumulativen Dissertation verwendeten Manuskripte sind sowohl über die Nutzung, als auch über die oben angegebenen Eigenanteile der weiteren Doktoranden/ Doktorandinnen als Koautoren an den Publikationen und Zweitpublikationsrechten bei einer kumulativen Dissertation informiert und stimmen dem zu.

Die Anteile der Promovendin sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an den Publikationen und Zweitpublikationsrechten bei einer kumulativen Dissertation sind in der Anlage aufgeführt (Musterbeispiel).

Name der Promovendin	Datum	Ort	Unterschrift

Ich bin mit der Abfassung der Dissertation als publikationsbasierte, d. h. kumulative, einverstanden und bestätige die vorstehenden Angaben. Eine entsprechend begründete Befürwortung mit Angabe des wissenschaftlichen Anteils der Doktorandin an den verwendeten Publikationen werde ich parallel an den Rat der Fakultät der Chemisch-Geowissenschaftlichen Fakultät richten.

Name Erstbetreuer(in)	Datum	Ort	Unterschrift
Name Zweitbetreuer(in)	Datum	Ort	Unterschrift