
# BRNO UNIVERSITY OF TECHNOLOGY

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

## CENTRAL EUROPEAN INSTITUTE OF TECHNOLOGY BUT

STŘEDOEVROPSKÝ TECHNOLOGICKÝ INSTITUT VUT


# BIOPHYSICAL INTERPRETATION OF QUANTITATIVE PHASE IMAGE

BIOFYZIKÁLNÍ INTERPRETACE KVANTITATIVNÍHO FÁZOVÉHO ZOBRAZENÍ

DOCTORAL THESIS
DIZERTAČNÍ PRÁCE


AUTHOR
AUTOR PRÁCE

Ing. Lenka Štrbková

SUPERVISOR
ŠKOLITEL

prof. RNDr. Radim Chmelík, Ph.D.



BRNO 2017

**SUMMARY**

This work deals with the interpretation of the quantitative phase images gained by coherence-controlled holographic microscopy. Since the datasets of quantitative phase images are of substantial size, the manual analysis would be time-consuming and inefficient. In order to speed up the analysis of images gained by coherence-controlled holographic microscopy, the methodology for automated interpretation of quantitative phase images by means of supervised machine learning is proposed in this work. The quantitative phase images enable extraction of valuable features characterizing the distribution of dry mass within the cell and hence provide important information about the live cell behaviour. The aim of this work is to propose a methodology for automated classification of cells while employing the quantitative information from both the single-time-point and time-lapse quantitative phase images. The proposed methodology was tested in the experiments with live cells, where the performance of the classification was evaluated and the relevance of the features derived from the quantitative phase image was assessed.

**ABSTRAKT**

Práce se zabývá interpretací kvantitativního fázového zobrazení pomocí techniky koherencí řízené holografické mikroskopie. Vzhledem k tomu, že tato technika generuje velké množství kvantitativních fázových obrazů o nezanedbatelné velikosti, manuální analýza by byla časově náročná a neefektivní Za účelem urychlení analýzy obrazů získaných pomocí koherencí řízené holografické mikroskopie je v této práci navržena metodika automatizované interpretace kvantitativních fázových obrazů pomocí strojového učení s učitelem. Kvantitativní fázové obrazy umožňují extrakci parametrů charakterizujících distribuci suché hmoty v buňce a poskytují tak cennou informaci o buněčném chování. Cílem této práce je navrhnout metodologii pro automatizovanou klasifikaci buněk při využití této kvantitativní informace jak ze statických, tak z časosběrných kvantitativních fázových obrazů. Navržená metodika byla testována v experimentech s živými buňkami, jimiž byla vyhodnocena výkonnost klasifikace a významnost parametrů získaných z kvantitativních fázových obrazů.

**KEYWORDS**

digital holographic microscopy, quantitative phase image, supervised machine learning, cell classification

**KLÍČOVÁ SLOVA**

digitální holografická mikroskopie, kvantitativní fázový obraz, strojové učení s učitelem, klasifikace buněk

ŠTRBKOVÁ, L. *Biophysical Interpretation of Quantitative Phase Image*. Brno: Brno University of Technology, Central European Institute of Technology, 2017. 75 p. Thesis supervisor prof. RNDr. Radim Chmelík, Ph.D.

## DECLARATION

I certify that the work presented in this thesis was performed independently, under the supervision of prof. RNDr. Radima Chmelík, Ph.D., and is original with the sole exception of the technical literature and other sources of information that are acknowledged in the text and reference list, and that the material has not been submitted, in whole or in part, for a degree at this or any other university.

Brno . . . . . .

. . . . . .
(author's signature)

**ACKNOWLEDGMENT**

I would like to thank prof. RNDr. Radima Chmelík, Ph.D. for supervision and overall support with encouragement during my postgraduate studies. I would also like to thank MUDr. Pavel Veselý, CSc., for the valuable advice in the field of biology of living cells, and Mgr. Veronika Jůzová, who provided great support during the cell specimen preparation. Finally, I would like to thank all colleagues from Experimental Biophotonics research group for sharing their scientific knowledge and valuable advice.

The research described in this thesis was financially supported by the CEITEC 2020 (LQ1601), by the project Advanced Materials and Nanotechnology (STI-S-14-2523), by the project Automatic detection of cellular apoptosis (FSI/STI-J-15-2752), and by the project Recognition of dynamic cellular processes in quantitative phase images (STI-J-16-3796).

Brno . . . . . .

. . . . . .
(author's signature)

# Table of Contents

# List of Acronyms

The following abbreviations are used in the text:

| | |
|---|---|
| ANN | artificial neural network |
| CCD | charge-coupled device |
| CDF | cumulative distribution function |
| DHM | digital holographic microscopy |
| DWT | discrete wavelet transform |
| EMT | epithelial-mesenchymal transition |
| FFT | fast Fourier transform |
| KNN | *K*-nearest neighbour method |
| PAA | piecewise aggregate approximation |
| PBS | phosphate-buffered saline |
| PCA | principal components analysis |
| QPI | quantitative phase imaging |
| ROI | region of interest |
| SAX | symbolic aggregate approximation |
| SVM | support vector machines |
| TGF-$\beta$ | transforming growth factor beta |

The following symbols are used in the text:

| | |
|---|---|
| $\alpha$ | significance level |
| $\gamma$ | refraction increment (ml.g$^{-1}$) |
| $\lambda$ | illumination wavelength (nm) |
| $\mu_\varphi$ | average phase (rad) |
| $\rho$ | dry mass density of a cell (pg.μm$^{-2}$) |
| $\sigma_\varphi$ | standard deviation of the phase |
| $\varphi$ | phase in the reconstructed image proportional to the optical path difference of the object and reference arm (rad) |
| $\varphi_o$ | phase in the object arm (rad) |
| $\varphi_r$ | phase in the reference arm (rad) |
| $\varphi_{total}$ | total phase of the cell (rad) |
| $\Phi_j$ | approximation coefficients in wavelet transform |
| $\psi$ | mother wavelet in wavelet transform |
| | |
| $A$ | pixel area (μm$^2$) |
| $C$ | concentration of dry protein in the solution (g.ml$^{-1}$) |
| $CA$ | convex area (μm$^2$) |
| $d$ | thickness of the cell (mm) |
| $D$ | directionality of cell motion |
| $d_a$ | accumulated distance travelled by the cell (μm) |
| $d_{Euclid}$ | Euclidean distance travelled by the cell (μm) |
| $EC$ | eccentricity |

| | |
|---|---|
| $EX$ | extent |
| $FA$ | footprint area ($\mu m^2$) |
| $fn_i$ | number of examples that were not recognized as class examples |
| $fp_i$ | number of examples that were incorrectly assigned to a class |
| $h$ | thickness of the medium (mm) |
| $I$ | indentation |
| $Kurt_\varphi$ | kurtosis of the phase |
| $N$ | number of classes |
| $\Delta n$ | difference between the refractive indices of the cellular material and the medium |
| $n_c$ | axially averaged refractive index of the cellular material |
| $n_m$ | refractive index of the surrounding medium |
| $P_{CA}$ | perimeter of the convex area (μm) |
| $P_{FA}$ | perimeter of the footprint area (μm) |
| $R$ | roundness |
| $S$ | solidity |
| $Skew_\varphi$ | skewness of the phase |
| $t$ | time (s) |
| $tn_i$ | number of correctly recognized examples that do not belong to a class |
| $tp_i$ | number of correctly recognized class examples |
| $v$ | velocity of cell motion ($\mu m.s^{-1}$) |
| $Var_\varphi$ | variance of the phase |

# 1. Introduction

Nowadays, the increasing prevalence of automated image acquisition systems is enabling microscopy experiments that generate large image datasets. However, the manual image analysis on large datasets has certain limitations. It requires an expert in the field who would perform inspection for every image, which needs considerable effort and concentration. Moreover, the analysis provided by one person has a tendency to be biased by subjective observation. The analysis result therefore largely depends on personal skills, decisions, and preferences. Another issue of the manual approach is that it is rather time-consuming. Consequently, these aspects impose significant constraints on the speed of the analysis and reliable interpretation of the microscopic images.

One of the approaches to address these limitations is machine learning. The technique has nowadays wide applications in different areas including fingerprint analysis, face identification, speech recognition, navigation and guidance systems, etc. [1]. Lately, it is increasingly being applied also in microscopy to speed up the analysis of microscopy images. Machine learning applied to image analysis provides an objective and unbiased method of scoring the content of microscopic images in contrast to subjective manual interpretation, thus potentially being more sensitive, consistent, and accurate.

Machine learning being a field within the artificial intelligence, exploits two major approaches. In supervised machine learning, a computer system is trained using a set of labelled pre-defined examples and then used to distinguish groups of objects based on the relevant patterns learned during the training. Supervised machine learning can be seen as a classification process, which attempts to assign each input value to one of a given set of classes. The other approach to machine learning is unsupervised learning. Here, the computer system does not rely on the prior knowledge and is not trained on labelled training examples. Instead, the system finds new patterns and subdivides the data by using a set of pre-defined general rules. An example of unsupervised learning is clustering, where a dataset can be divided into several groups based on prior definitions characterising a cluster, or a desired number of clusters. This work focuses solely on the application of supervised machine learning to automated analysis of microscopy image datasets.

During the recent years, Coherence-controlled holographic microscopy (CCHM) [2,3] has been developed in the laboratory of Experimental Biophotonics group, CEITEC Brno University of Technology. CCHM is a label-free interferometric microscopy technique able to provide quantitative phase images of living cells [4]. CCHM enables to detect not only the amplitude, but also the phase of the wave transmitted through a specimen. This fact is of particular importance while observing the live cells that are considered to be weakly scattering and absorbing specimens (termed phase objects). In case of phase objects, the phase carries considerably more information about the specimen than the amplitude of transmitted light and is, therefore, of great significance. The imaging in CCHM is based on the interference of the object and the reference light beams, which enables to detect the phase delay induced by the specimen [5]. The phase in the image contains quantitative information expressed in radians and is proportional to the optical path difference of the object and the reference arm. It has been demonstrated in several publications that the measured phase corresponds to the dry mass distribution within the cell [6,7]. Since CCHM

enables multidimensional imaging with high acquisition rate, the datasets obtained from the experiments are rather large. Therefore, the automated method for microscopic data analysis and interpretation is in great demand.

For the reason stated above, this work focuses on the supervised machine learning and its application for the interpretation of the quantitative phase images. The goal of this work is to propose a methodology for automated analysis of quantitative phase images by means of supervised machine learning and verify the potential of methodology in the experiments with live cells.

Two main approaches for the automated interpretation of quantitative phase imaging were proposed. Firstly, the work focuses on the analysis of static quantitative phase images, where the methodology for automated classification of cells is proposed. The approach is tested in the experiment and compared with the commonly used methods based on bright-field microscopy images. Furthermore, the methodology for automated analysis of time-lapse quantitative phase images incorporating the temporal information is proposed and its functionality is demonstrated in the experiment. The results and potential of both proposed methodologies are critically discussed and, finally, the proposals for further progress and improvements are made.

# 2. Review

This section gives an overview of the state-of-the-art results in the field of supervised machine learning applied to microscopic image analysis. Currently available literature and techniques related to the topic are mentioned and critically evaluated. The novelty of methods used in the thesis is substantiated and the overall purpose of the work is stated.

Analysis of microscopic images based on machine learning has been attracting considerable attention in the past few years. The increasing rate of publications in this topic serves as evidence (Figure 1).

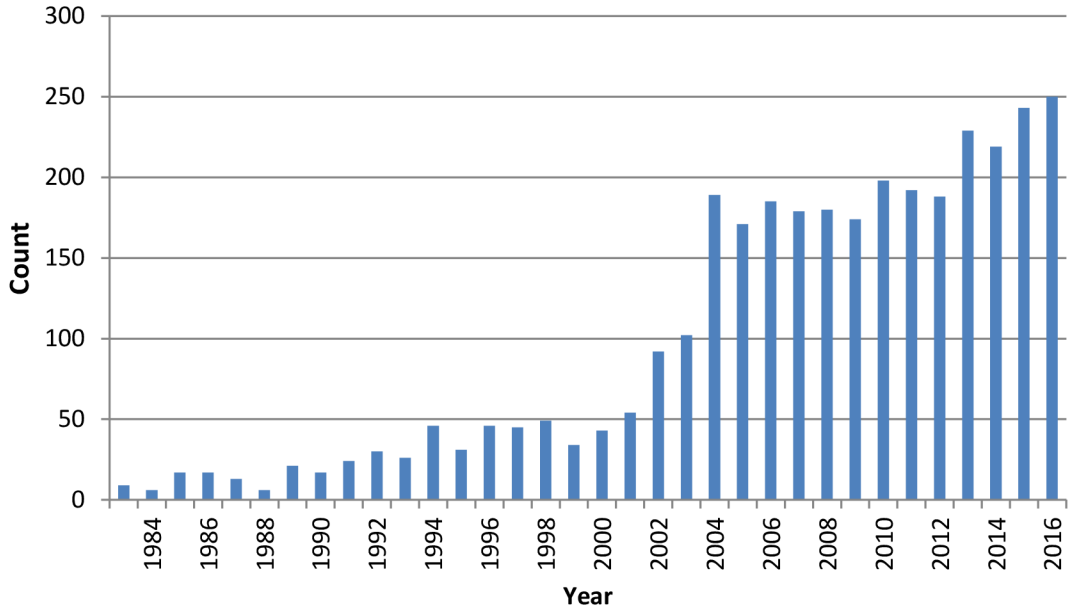


**Figure 1:** Number of publications in recent years regarding "machine learning in microscopy". Source: http://www.scopus.com/

The beginnings of the machine learning in microscopy can be dated back to the year 1951, when a paper [8] was published by Mellors and Silver, who have been focusing on automatic detection of different types of cells. However, it is only recently that digital photography, computer speed, RAM size and secondary storage capacity have made machine learning in microscopic images possible. Since then, many works regarding this topic were published. Among them can be mentioned the work of Comaniciu et al. [9] on image-guided decision support system for pathology, which describes a system designed to assist pathologists to discriminate among malignant lymphomas and chronic lymphocytic leukemia directly from microscopic specimens. Swolin et al. published work [10] describing differential counting of leukocytes in blood samples using bright-field microscopy and a decision support system based on artificial neural networks. Rajpoot wrote paper [11] about hyperspectral colon tissue cell classification, where the algorithm based on supervised support vector machines (SVM) classification between normal and malignant tissue cells of the human colon is presented. Machine learning approach for classification of erythrocytes in anemia based on morphological changes was presented by Das et al. [12]. The automated classification of myeloma cells in microscopic images was proposed by Saeedizadeh et al. [13]. In the menstioned studies, the input images for

classification were gained by bright-field imaging of stained cells. The drawback of this approach is the necessity of the sample preparation by fixing cells before imaging. The cells can undergo different morphological and physiological changes while being fixed, which could possibly affect the measurement.

Another study [14] presents application of machine learning techniques to analysis of cell morphology in phase-contrast microscopy images. However, the images gained by phase-contrast microscopy demonstrate halo artifact, which makes the boundaries of the cells appear brighter and might lead to challenging and inaccurate segmentation results. This may result in a poor accuracy of the machine learning classifier.

Several publications have focused on classification of cells in the images gained by fluorescent microscopy. Automated scoring of diverse cell morphologies by means of machine learning was described in [15]. Several automated image analysis methods for high-content screening of fluorescent images were summarised in [16]. However, the drawback of these techniques is the necessity of sample preparation by fluorescent staining of cells before imaging. Moreover, the fluorescent stain is likely to influence the cell behaviour as well as the cell morphology, which could possibly affect the experiment and classification results.

In the mentioned approaches, the features extracted from the images are mostly representing the cellular shape or the intensity values depending on the stain concentration, but they are not quantitative in terms of cell mass.

In the recent years, digital holographic microscopy (DHM) has proven as a very versatile non-invasive tool for the observation of live cells [17–20], while overcoming the limitations of previously mentioned approaches. DHM provides quantitative phase images (QPI) with high intrinsic contrast without labelling and since the images contain quantitative information about cell mass, it may potentially improve the performance of the classification.

Several publications studied cell behavior by monitoring cell features extracted from the QPI. Cell life cycle characterization by monitoring of morphometric and quantitative phase features was proposed in [21]. Assessment of wound healing by monitoring the cellular volume, dry mass and refractive index was presented in [22]. The study of cancer cell growth and drug response by monitoring cell dry mass is described in [23]. In the mentioned publications, the authors extract quantitative phase features and monitor their changes, but do not apply machine learning algorithms for the automated assessment of cell behavior.

Only limited work has been published towards the application of machine learning classification algorithms to QPI. Morphology-based classification of red blood cells using DHM was presented in [24]. Automated detection and classification of living organisms in drinking water resources using DHM was performed in [25]. The automated diagnosis of breast and prostate cancer from tissue biopsies was described in [26] and in [27], respectively. But to my present knowledge, none of the publications studied the potential of QPI for the classification of live adherent eukaryotic cells.

Great progress has already been made since the early beginnings of the machine learning in the field of microscopy. The extent of mentioned publications implies that

machine learning applied to quantitative phase images is a current and rather expanding topic. However, there is still major scope for further investigation in this research area.

None of the above mentioned publications studied the effect of using the features based on quantitative phase images on the performance of classification. Neither have they mentioned analysis of the live adherent eukaryotic cells with the indented boundaries which are difficult to define and segment from the background and which subsequently introduce high variance within the classified groups. To my present knowledge, there is no reference in the literature to the application of machine learning to the time-lapse quantitative phase images with the focus on analysis and interpretation of live cell behaviour.

This thesis focuses on the mentioned issues and is a follow-up to existing results reached in the research area. The goal is to propose a methodology for classification of the live adherent eukaryotic cells based on QPI and to evaluate the potential of features extracted from QPI. In addition, the methodology for cell classification from time-lapse QPI will be proposed in order to gain additional context from the temporal information for the more accurate interpretation of live cell behaviour. Both methodologies will be employed in the experiments with live adherent eukaryotic cells and the results will be discussed.

I expect that the thesis will not only bring the outcomes that are scientifically relevant to the field of machine learning in microscopy, but also will serve to assist and speed up the analysis and interpretation of live cell behaviour by CCHM and therefore promote CCHM as a diagnostic method in biology and medicine.

# 3. Aims of Thesis

The final objective of this work is to propose a methodology, which would serve for the biophysical interpretation of quantitative phase image gained by CCHM in an automated fashion. Therefore, the partial aims of this work are the following:

- acquire datasets of quantitative phase images by CCHM, which are suitable for automated interpretation by means of supervised machine learning,
- propose methodology for classification of cells based on static quantitative phase images,
- propose the appropriate features to be extracted, representing the cell morphology in quantitative phase images,
- apply the methodology in the experiment with live cells,
- evaluate performance and compare the proposed methodology with the current state-of-the-art techniques while estimating the potential of features gained from quantitative phase,
- propose methodology for classification of cells based on time-lapse quantitative phase images,
- propose the appropriate features representing the cell behaviour, exploiting the temporal information from the time-lapse quantitative phase images,
- apply the methodology in the experiment with time-lapse imaging of live cells,
- evaluate performance of the proposed approach and discuss the potential contribution it may have for the interpretation of quantitative phase image gained by CCHM.

# 4. Structure of Thesis

The thesis is organized as follows. The Section 5 explains the basic concept of coherence-controlled holographic microscopy and the quantitative phase image that it provides. In Section 6, machine learning for the interpretation of quantitative phase images is presented. The proposed methodology for classification of cells in the static quantitative phase images is described. Section 7 presents the application of the proposed methodology in the experiment with live cells. In Section 8, the methodology for interpretation of time-lapse quantitative phase images is proposed and demonstrated on the experimental data. Finally, the conclusions and the future outlook are summarized in Section 9 and 10.

# 5. Coherence-Controlled Holographic Microscopy (CCHM)

Coherence-controlled holographic microscopy (CCHM) [2,3] is a label-free interferometric technique developed at Brno University of Technology. The technique is widely used in the laboratory of Experimental Biophotonics (CEITEC BUT) for monitoring of live cell behaviour [17,28,29], but also for technical specimens [30]. The main asset of this technique is the ability to provide quantitative phase image [4]. During the imaging, not only the amplitude, but also the phase of the wave transmitted through the specimen is detected. This fact is of particular importance while observing the live cells that are considered to be weakly scattering and absorbing specimens. In case of such specimens, the phase carries considerable amount of information about the specimen structure and is, for that reason, of great significance. The imaging in CCHM is based on the interference of the object and the reference light beams, which enables to detect the phase delay induced by the specimen.

The optical set-up of the microscope is based on Mach-Zehnder-type interferometer modified for achromatic off-axis holographic microscopy (Figure 2). The illumination system is formed by a low coherence source (halogen lamp), interference filters, collector lens and beamsplitter, which splits the beam into two arms. Microscope therefore consists of two separated nearly identical optical arms – reference and object arm. Both arms contain matching condensers, objectives and tube lenses. The reference arm includes diffraction grating, which spatially separates light of different wavelengths. Only the +1st order of the diffraction grating is separated and interferes with the object arm in the output plane, while creating the interference structure – hologram. The hologram is recorded by the CCD camera and further reconstructed.

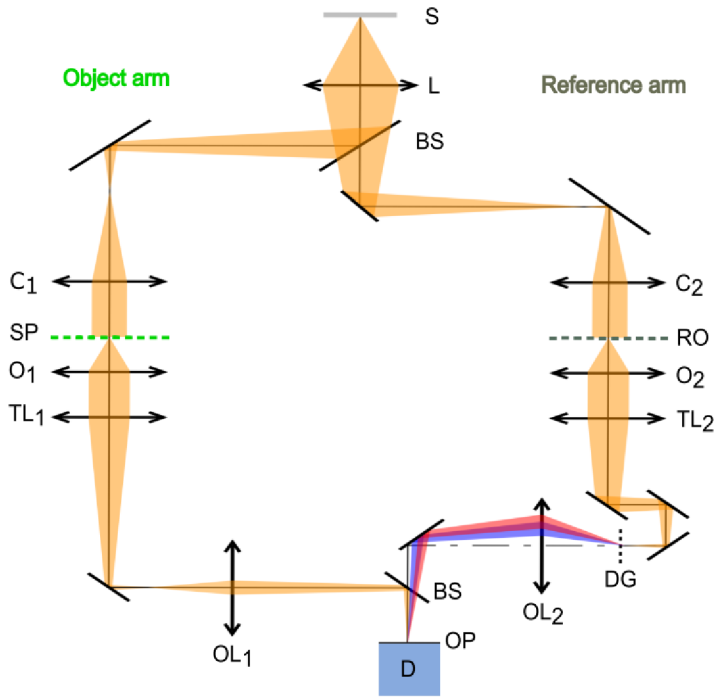


**Figure 2:** Optical setup of coherence-controlled holographic microscope. Light source (S), relay lens (L), beamsplitters (BS), condensers (C), specimen (SP), reference object (RO), microobjectives (O), tube lenses (TL), diffraction grating (DG), output lenses (OL), output plane (OP), detector (D) [2].

The numerical reconstruction of the hologram is performed using the house-built software. Firstly, the 2D Fourier transform [31] is computed from the hologram (Figure 3a) in order to obtain the spatial frequency spectrum (Figure 3b). The object spatial frequency spectrum is selected and the zero spatial frequency is shifted to the centre. By applying the inverse fast Fourier transform algorithm, the complex amplitude of the object wave is obtained. The intensity image (Figure 3c) and the raw phase image (Figure 3d) are then reconstructed from the complex amplitude. Since the values in the raw phase image are wrapped on the interval $(-\pi, \pi)$, the phase unwrapping algorithm [32,33] is applied. After the reconstruction, the image can still be burdened by the optical aberrations of the imaging system, imperfect adjustment of the microscope, or possibly by surrounding temperature changes. This issue is solved by the subtraction of the compensation surface described in detail in [34]. In this way, final unwrapped and compensated phase image is obtained (Figure 3e). Such reconstructed quantitative phase image contains values of phase delays induced by the specimen expressed in radians and can be visualised as 3D surface plot (Figure 3f).

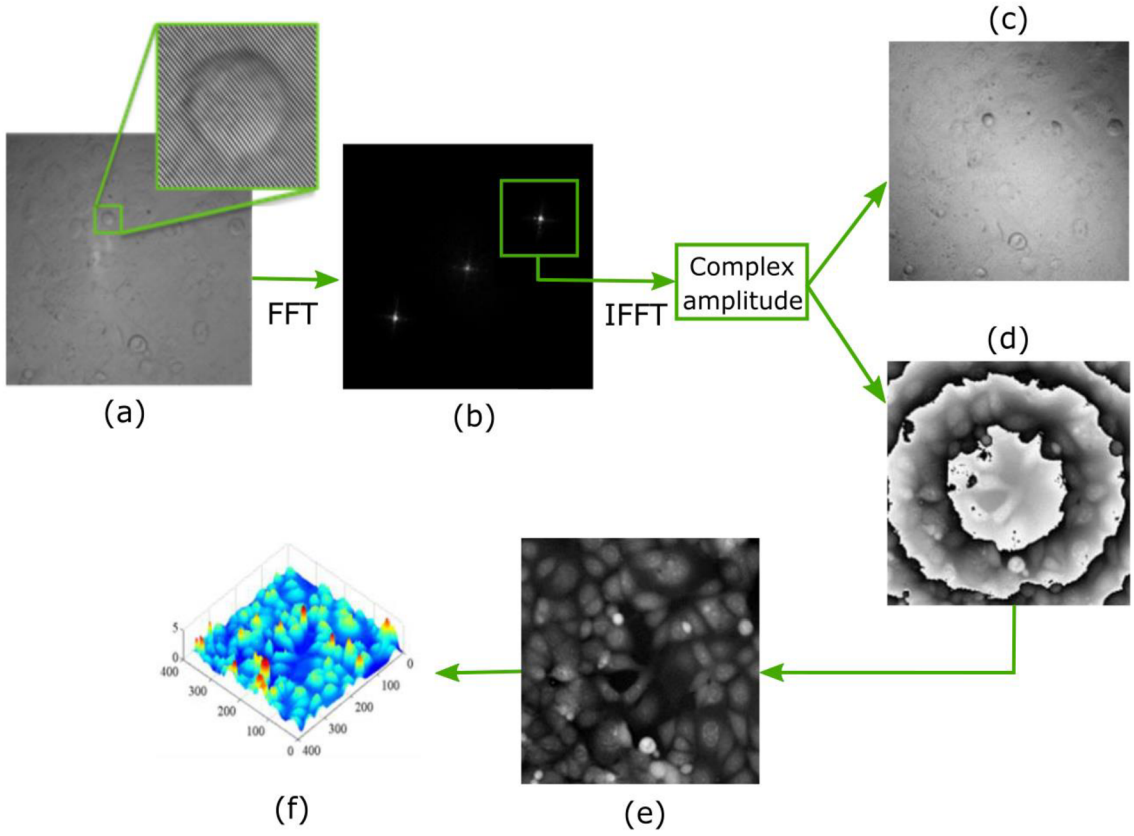


**Figure 3:** Overview of the QPI reconstruction process. Hologram (a), spatial frequency spectrum with indicated windowing operation (b), intensity image (c), raw phase image (d), unwrapped and compensated phase image (e) and 3D surface plot (f).

In contrast to existing DHM techniques [21,22,24,35], the use of incoherent illumination enables high-quality quantitative phase imaging with strong suppression of coherent noise and parasitic interferences while providing high temporal stability and spatial uniformity of the phase measurement [2]. Using the approach described in [36], the temporal and spatial phase sensitivity were determined as 0.0081 rad and 0.0094 rad,

respectively. The lateral resolution is comparable with the lateral resolution of conventional wide-field optical microscopes, thus twice better than in typical DHM techniques with a coherent source of illumination. Moreover, the low illumination power of the incoherent source (0.2 μW.cm$^{-2}$) is not likely to influence the physiological functions of the imaged cells, which is very convenient for live cell imaging.

## 5.1 Quantitative Phase Image

The phase in the reconstructed image contains quantitative information and is proportional to the optical path difference of the object and reference arm according to the following equation [4]:

$$\begin{aligned}\varphi(x,y) &= \varphi_o(x,y) - \varphi_r(x,y) \\ &= \frac{2\pi}{\lambda}\left[n_m\big(h - d(x,y)\big) + n_c\,(x,y)d(x,y)\right] - \frac{2\pi}{\lambda}n_m h \\ &= \frac{2\pi}{\lambda}d(x,y)(n_c\,(x,y) - n_m) = \frac{2\pi}{\lambda}d(x,y)\Delta n(x,y)\,,\end{aligned} \tag{1}$$

where $\varphi_o$ is phase in the object arm, $\varphi_r$ is phase in the reference arm, $\lambda$ is the illumination wavelength, $n_m$ is the refractive index of the surrounding medium, $h$ is the thickness of the medium, $d$ is the thickness of the cell, $n_c$ is the axially averaged refractive index of the cellular material and $\Delta n$ is the difference between the refractive indices of the cellular material and the medium (Figure 4).

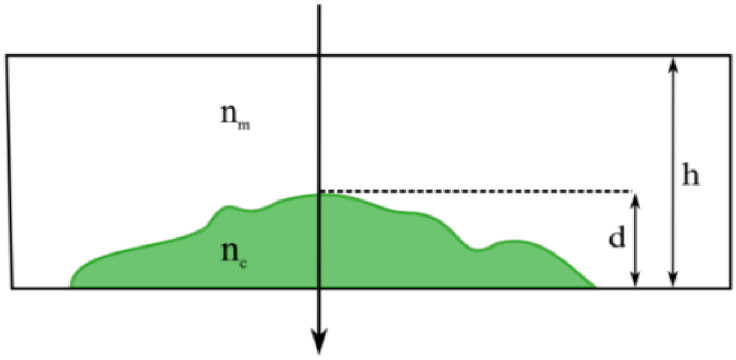


**Figure 4:** Model of an adhered cell surrounded by the medium observed by CCHM.

The phase can be also interpreted in terms of cell dry mass. The dry mass of the cell consists of its non-aqueous content and was defined as the mass of the cell after evaporation of the water. The value of the cell dry mass is dependent mainly on the protein concentration within the cell [37]. It has been shown that the refractive index of the cellular material is directly proportional to the dry mass of the cell with the proportionality constant $\gamma$ referred to as the refraction increment (units of ml.g$^{-1}$) according to the following equation [7]:

$$n_c(x,y) = n_m + \gamma C(x,y)\,, \tag{2}$$

where $C$ is the concentration of dry protein in the solution (in g.ml$^{-1}$). The refraction increment $\gamma$ indicates how much the refractive index of the aqueous solution increases for each increase in the dry mass concentration in the solution. Various cell components have very similar values of the refraction increment (0.0017 – 0.0019 ml.g$^{-1}$). It has been

published years ago that the measured phase corresponds to the dry mass distribution within the cells [6,7,37,38]. The dry mass density of the cell (units of pg.μm$^{-2}$) can be obtained from the measured phase as follows:

$$\rho(x,y) = \frac{\lambda}{2\pi\gamma}\varphi(x,y)\ . \tag{3}$$

# 6. Machine Learning in QPI

## 6.1 Introduction to Machine Learning

Machine learning as a field of artificial intelligence, explores the algorithms that have the ability to automatically learn and improve from the data and subsequently make predictions on the unknown data [1]. In order to gain the general overview, the basic machine learning methods are introduced in this part. Machine learning algorithms are often categorized as supervised or unsupervised (Figure 5).

In supervised machine learning [39], the algorithm relies on the prior knowledge and is trained from labelled training data. The training data consist of a set of training examples, while each example is a pair consisting of an input object (typically a vector) and the desired output label (pre-defined class). Based on the training, the supervised learning algorithm produces a function that maps the input objects to the output classes. After the training phase, the algorithm should correctly determine the class labels for unseen objects.

In contrast, the unsupervised learning [40] does not rely on any prior knowledge. This approach is used, when the training data are neither classified nor labelled. Unsupervised learning algorithms (e.g. *k*-means clustering, fuzzy c-means clustering, hierarchical clustering, Gaussian mixture models, neural networks, hidden Markov models) infer a function to describe a hidden structure from the unlabelled data, based on which the data are divided into clusters.

Other approaches to machine learning include semi-supervised learning, which uses both labelled and unlabelled data for training, and reinforcement learning, where decisions are improved in an iterative process based on feedback and specified scoring.

However, only supervised machine learning will be employed in this work. Supervised learning problems can be further divided into regression and classification problems. In classification algorithms (e.g. support vector machines, discriminant analysis, *k*-nearest neighbour, ensemble methods, decision trees, neural networks, etc.), the output variable is a category, while in regression algorithms (e.g. linear/nonlinear regression, Gaussian process regression - GPR, support vector regression - SVR, ensemble methods, decision trees, neural networks) it is a real value. This work focuses solely on the classification algorithms.

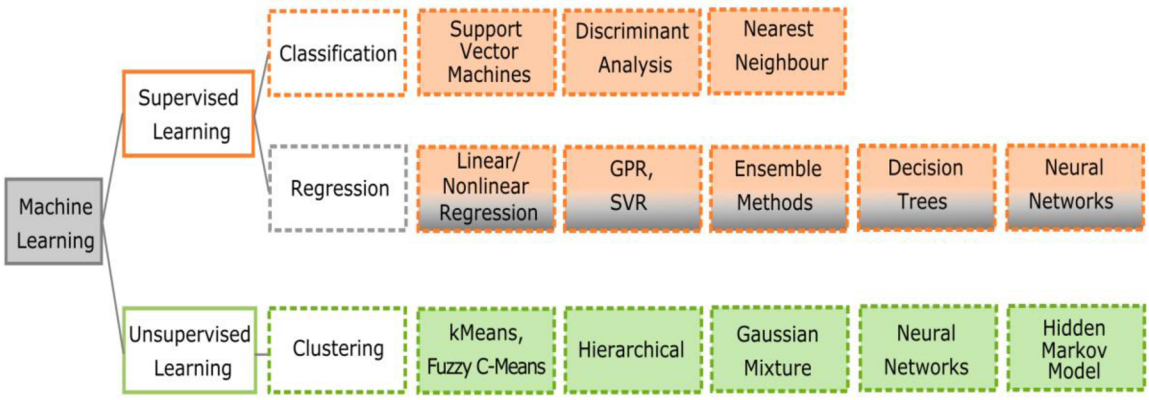


**Figure 5**: The overview of the two main types of machine learning with the corresponding categories of algorithms.

## 6.2 Classification in QPI

In this part of thesis, the approach for automated classification of cells in the quantitative phase images obtained by CCHM is proposed. In cell classification, the algorithm identifies patterns in the input images of cells and trains a model based on class labels which were assigned to the cells in the images by expert. Such trained model is able to classify cells in new so far unseen images. The essential precondition for the successful classification is a sufficiently large database of labelled cell images on which the classifier is trained.

The classification process starts with image pre-processing of quantitative phase images from the database, where the cells are segmented from the background and each cell is identified as a separate region of interest (ROI). From each ROI, features representing the cell are extracted. There are several types of features generally used, characterising the texture, geometry and morphology of cells. Thanks to the quantitative information contained in the images obtained by CCHM, it is possible to extract also the features related to the dry mass distribution within the cell. These features carry valuable information characterizing the cell behaviour. The best features are then selected and the data are split into the training and testing set in order to avoid overfitting. The training data are labelled by expert biologist and serve as an input for the classification algorithm. After the training of the classification algorithm on the labelled data, the testing unlabelled data can be fed into the classifier. The overview of the proposed classification process based on QPI is shown in Figure 6. In the following chapters, the steps of the classification process will be introduced in detail.

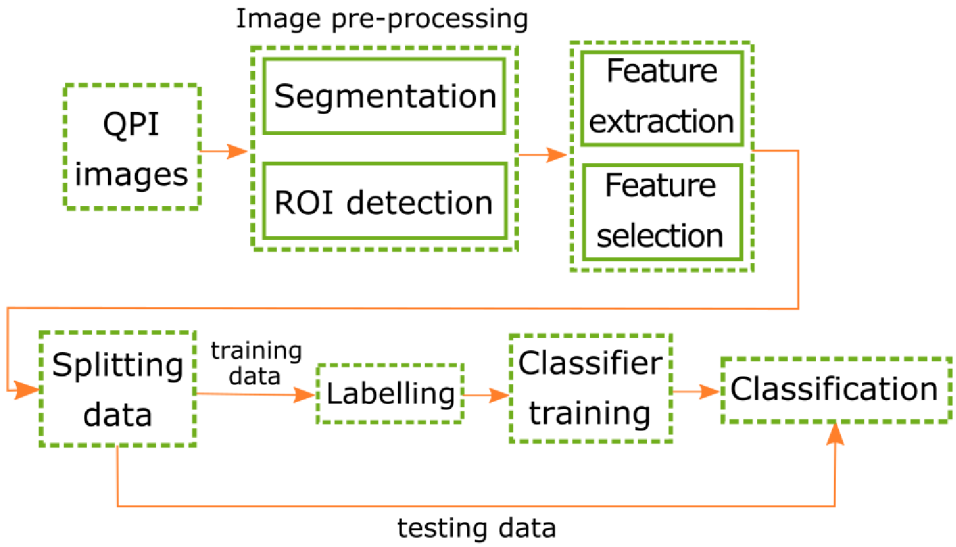


**Figure 6:** Overview of the proposed classification process based on QPI. Firstly, image pre-processing is carried out. The cells in the image are segmented from the background and identified as regions of interest (ROIs). Cell features are extracted for every ROI and the best features are selected. The data are split into training and testing set. The training data are labelled by expert biologist and form an input for the classifier. The classifier is trained on labelled data and prepared to perform the classification on testing unlabelled data.

### 6.2.1 Image Pre-processing

In this stage, the cells in the quantitative phase images are firstly segmented from the background. Several methods for the segmentation exist [41], in this work the marker-controlled watershed segmentation approach [42], implemented in Q-Phase software (TESCAN ORSAY HOLDING a.s., Brno, Czech Republic), is applied. The segmented cells are then identified as separate ROIs, while each of them is labelled by a unique integer number as shown in Figure 7.

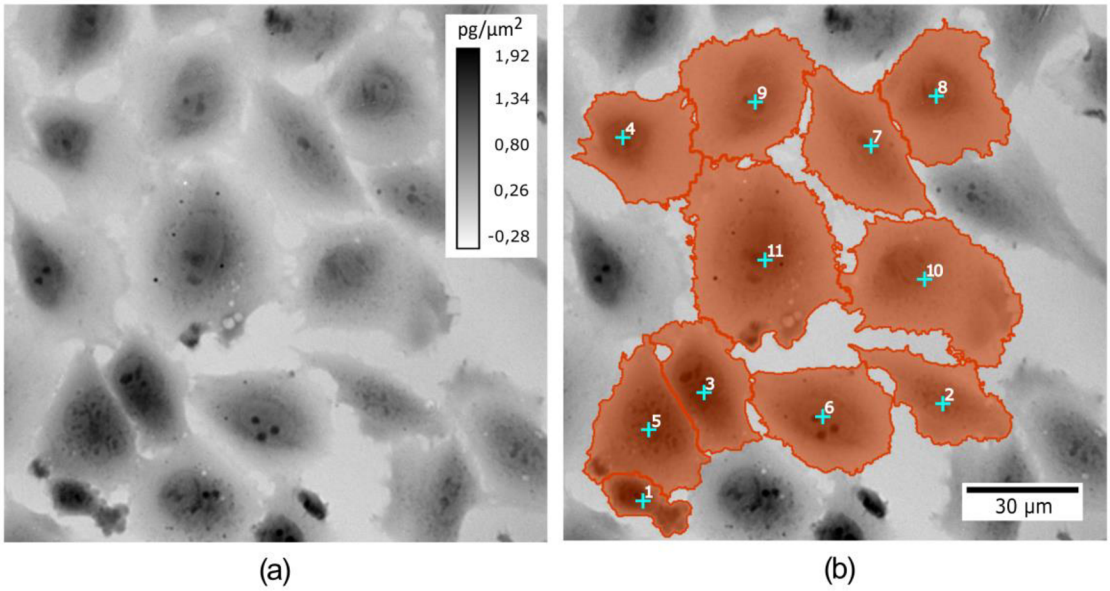


**Figure 7:** Results of the segmentation of nearly confluent LW13K2 cells by marker-controlled watershed approach. (a) Original quantitative phase image. (b) Segmented image. Quantitative phase images are shown in grayscale in units of pg.µm$^{-2}$ recalculated from phase (in radians) according to Davies [7].

### 6.2.2 Feature Extraction

After the image pre-processing, each ROI is represented by a set of cell features. Representation of the cell by cell features is used in machine learning to overcome the problem of high dimensionality of the input image. The representation must be in a form suitable for the classification algorithm, mostly a numeric feature vector. The process, in which the input data are transformed into a reduced representation by feature vector, is often termed feature extraction [43]. In this work, two types of cell features were extracted: morphometric and QPI features.

*(a) Morphometric (MO) cell features*. The features mostly reflect the shape of the cell and are explained as follows.

*(i) Footprint area (FA)* is calculated as the sum of the pixels of the projected cell area. Pixels belonging to the cell region have the value $m = 1$, otherwise $m = 0$. When multiplied by the pixel area according to following equation, the resulting value of *FA* is obtained in units of area:

$$FA = \sum_{i=1}^{n} m_i A\,, \tag{4}$$

where $n$ is the number of pixels in the image and $A$ is the pixel area.

*(ii) Perimeter of the footprint area ($P_{FA}$)* is defined as the sum of pixels in the inner boundary of the region. When multiplied by the pixel size, the resulting value of $P_{FA}$ is in units of length.

*(iii) Convex area (CA)* is calculated as the sum of pixels of the convex cell region, multiplied by the pixel area. The boundaries of the convex cell region are defined by the smallest convex polygon that contains the region of the cell.

*(iv) Perimeter of the convex area ($P_{CA}$)* is calculated as a sum of pixels in the inner boundary of the region, and multiplied by the pixel size.

*(v) Solidity (S)* specifies the proportion of the pixels belonging to the cell footprint area to those which are contained in the convex area.

*(vi) Roundness (R)* determines the deviation of the cell region from the circular shape. Roundness depends on the footprint area and its perimeter according to the following relationship:

$$R = \frac{4\pi\, FA}{(P_{FA})^2}. \tag{5}$$

*(vii) Indentation (I)* evaluates the level of cell boundary indentation. Indentation can be calculated as the ratio of perimeter of the convex area and perimeter of the footprint area as follows:

$$I = \frac{P_{CA}}{P_{FA}}. \tag{6}$$

*(viii) Eccentricity (EC)* specifies the eccentricity of the ellipse that has the same second-moments as the cell region. The eccentricity is calculated as the ratio of major axis and minor axis length. The value of eccentricity is between zero and one, while values close to zero describe circular shape and values close to one elongated shape of the region.

*(ix) Extent (EX)* is given by the ratio of pixels in the cell region to pixels in the total bounding box. Bounding box is the smallest rectangle containing the region. The extent is computed as the footprint area divided by the area of the bounding box.

*(b) QPI cell features.* The features are extracted from the phase values of the cell in quantitative phase image and therefore contain quantitative information about the dry mass density distribution within the cell.

*(i) Total phase of the cell ($\varphi_{total}$)* is calculated as the sum of phase values (in radians) in the pixels belonging to the region of the cell. $\varphi_{total}$ is calculated as follows:

$$\varphi_{total} = \sum_{i=1}^{k} \varphi_i, \tag{7}$$

where $k$ is the number of pixels of the cell region and $\varphi_i$ is the phase value in the $i$th pixel belonging to the region of the cell.

*(ii) Average phase ($\mu_\varphi$)* specifies the average phase value in the cell region. The average phase value is defined as the total phase over the footprint area of the cell.

*(iii) Variance ($Var_\varphi$) and standard deviation of the phase ($\sigma_\varphi$)* determine the variation of the phase values and therefore also of dry mass distribution within the cell. The variance and standard deviation of the phase are calculated as follows:

$$Var_{\varphi} = \frac{1}{k-1}\sum_{i=1}^{k}(\varphi_i - \mu_{\varphi})^2, \tag{8}$$

and

$$\sigma_{\varphi} = \sqrt{Var_{\varphi}}. \tag{9}$$

*(iv) Skewness ($Skew_{\varphi}$)* is calculated from the histogram of the phase values and describes its shape. Skewness measures the symmetry of distribution of the phase values from the mean value. The parameter is determined by the following equation:

$$Skew_{\varphi} = \frac{\sum_{i=1}^{k}\left(\varphi_i - \varphi_{avg}\right)^3}{(k-1)\sigma_{\varphi}^3}. \tag{10}$$

The values of skewness close to zero report about symmetrical distribution of phase values, which is characteristic for spread and well-adhered cells.

*(v) Kurtosis ($Kurt_{\varphi}$)* is also derived from the histogram of the phase values and quantifies the extent to what shape of the data distribution matches the normal distribution. Kurtosis is described as follows:

$$Kurt_{\varphi} = \frac{\sum_{i=1}^{k}(\varphi_i - \varphi_{avg})^4}{(k-1)\sigma_{\varphi}^4}. \tag{11}$$

The distribution matching to normal results in values of kurtosis close to zero. A flatter distribution and a more peaked distribution have negative and positive kurtosis value, respectively.

All extracted features are summarized into feature vectors, each feature vector representing one cell. Each cell feature vector is then assigned one of the class labels determined by the expert biologist. The class labels are later used for training of the classification algorithm.

Additional step before the classification is the feature scaling, which is commonly used in case that the values of features are not of similar scale. In this process, the feature values are scaled to a fixed range from 0 to 1. The main advantage of scaling is to avoid features in greater numeric ranges dominating those in smaller numeric ranges [43]. Moreover, feature scaling speeds up the training of the classifier, prevents the classifier from getting stuck in local optima and is an essential step in the classification for some algorithms. The scaling of the features is done via the following equation:

$$Y_{scaled} = \frac{Y - Y_{min}}{Y_{max} - Y_{min}}, \tag{12}$$

where $Y_{scaled}$ is the scaled value of the feature, $Y$ is the original value of the feature, $Y_{min}$ is the minimum value of the feature, $Y_{max}$ is the maximum value of the feature.

### 6.2.3 Feature Selection

In general, the variance of features within the class should be small, which means that features derived from different samples of the same class should have similar values. Also, the interclass separation should be large, i.e. feature values extracted from samples of

different classes should differ significantly. In order to evaluate the ability of the extracted features to discriminate between the classes, further analysis is performed. Parametric *t*-test for samples with different variances also known as Welch's *t*-test [44] was chosen for that purpose. Statistical parametric tests (including *t*-test) are based on the assumption that the data follows a normal distribution [45]. Therefore, before performing any parametric analysis, it should be proved that the data meet this requirement.

**Testing of Data Normality**

It is possible to confirm the normality visually or by significance tests. As a visual method for the normality check, box-whisker plot is used in this work (Figure 8). The box-whisker plot shows the median as a horizontal line inside the box and the interquartile range (between the 25th to 75th percentiles) as the length of the box. The whiskers symbolize the minimum and maximum values within 1.5 times the interquartile range. Samples outside this range are out of the box-whisker plot and are considered as outliers. A box-whisker plot that is symmetric with the median line at approximately the centre of the box and with symmetric whiskers suggests that the data may be normally distributed [46]. The notches of the box-whisker plot also provide a rough measure of the significance of differences between the medians of samples. If the notches do not overlap, then there is evidence that the medians are significantly different at the 5% significance level.

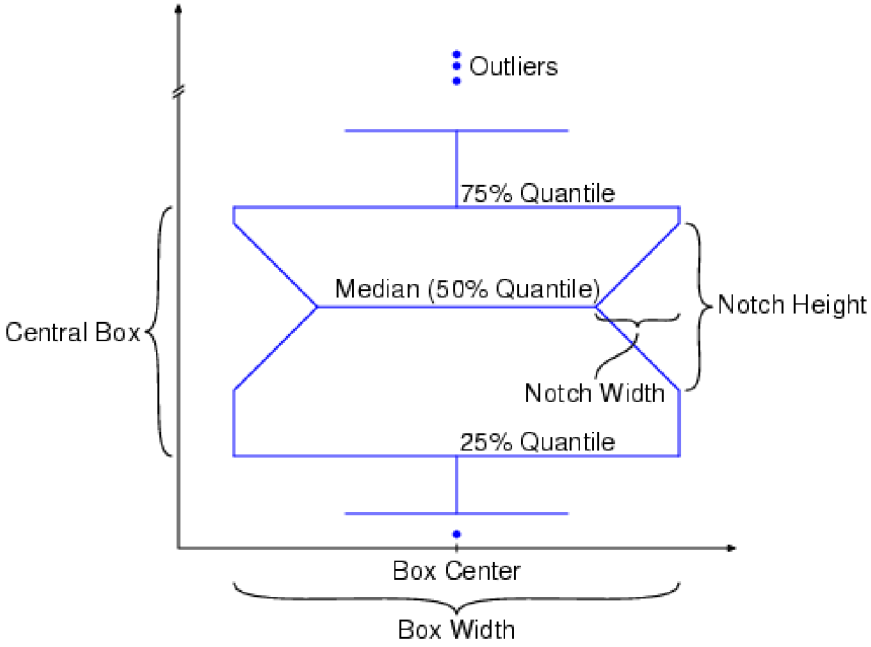


**Figure 8:** Box-whisker plot representation.

As a significance test for the normality, the Shapiro-Wilk test [47] is used. The test belongs to the correlation tests, which are based on the ratio of two weighted least-squares estimates of scale obtained from order statistics. Given an ordered random sample $x_1 < x_2 < \ldots < x_n$, the test statistic is defined as follows:

$$SW = \frac{\left(\sum_{i=1}^{n} \omega_i x_i\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{13}$$

where $x_i$ is the $i$th order of the statistics, $\bar{x}$ is the sample mean, $\omega_i$ are the weights which are derived from the expected values and covariance matrix of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution.

The null hypothesis of the test assumes that the data came from the normal distribution. Thus if the $p$-value is less than the chosen significance level, the null hypothesis is rejected and there is an evidence that the data tested are not from a normally distributed population. Otherwise the null hypothesis that the data are normally distributed cannot be rejected.

**T-test**

As mentioned earlier, the $t$-test belongs to the parametric tests and hence its precondition is the normal distribution of the analysed data. The independent two-sample $t$-test for samples with different variances also known as Welch's $t$-test was chosen to assess whether there have been significant differences between the means of the features of the analysed classes. The statistic to test the difference of the samples means is expressed as follows:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \tag{14}$$

where $\bar{x}$ and $\bar{y}$ are the sample means, $s_x$ and $s_y$ are the sample standard deviations, and $n_x$ and $n_y$ are the sample sizes.

The formula for the degrees of freedom is known as the Welch-Sattertwaite-equation [48] and is calculated as

$$df = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y - 1}} \tag{15}$$

The null hypothesis assumes that the samples come from populations with equal means and equal but unknown variances. The alternative hypothesis is that the samples come from populations with unequal means. If the $p$-value is less than the chosen significance level, the null hypothesis is rejected, otherwise it holds. In this specific case, the outcome of the $t$-test indicates the potential of individual features to differentiate between the classes of cells.

Based on the analysis, only the features with the potential to discriminate between the classes are kept in the feature vector. The resulting feature vector is then used as input for the classifier.

### 6.2.4 Supervised Classification Algorithms

After the feature extraction and selection, the final step is application of the classification algorithm. It is well known that the performance of the classification is highly dependent on the selection of the classification algorithm [49] and thus we employ several supervised machine learning algorithms in this work to correctly assess the performance of the

classification. Moreover, each algorithm can be adjusted by setting its parameters and, therefore, most of the algorithms are tested in several possible variations. The classification was performed in Matlab 2016b (MathWorks, Inc.). A short description of the used algorithms is presented below.

*(a) Decision trees.* In the decision tree classifier [50], a tree structure is built with root node and leaf nodes. The leaf nodes represent the class labels, while the branches represent conjunctions of features that lead to those class labels. Every interior node in the tree consists of a decision criterion. The features are partitioned based on homogeneity until a leaf node is assigned to a particular class label. Three types of decision tree classifiers were used in this work: complex, medium and simple tree, with defined maximum number of splits: 100, 20 and 4, respectively.

*(b) Discriminant analysis.* Discriminant analysis [51] assumes that different classes generate data based on different Gaussian distributions. To train a classifier, the fitting function estimates the parameters of a Gaussian distribution for each class. We used both linear and quadratic discriminant analysis.

*(c) Support vector machines (SVM).* SVM [52] classifies data by finding the best discriminating hyperplane that separates objects with different class membership as shown in Figure 9. Depending on a given problem, larger number of hyperplanes may exist. The distance from the hyperplane to the closest data point is called the margin of separation. The aim of a support vector machine is to find the particular hyperplane, for which the margin of separation is maximized. The hyperplane fulfilling this condition is referred to as the optimal hyperplane. The closest data points to the margin of separation are called support vectors. The support vectors thus specify the discrimination function.

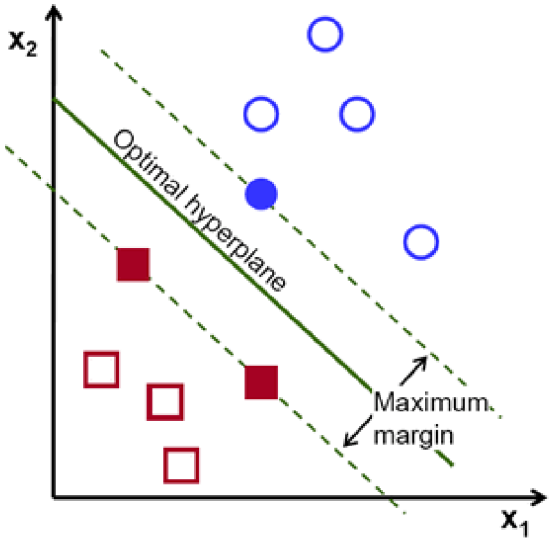


**Figure 9:** Example of a SVM classifier for the case of two linearly separable classes. Dotted lines mark the margin and full points represent the support vectors.

The SVM can handle both linearly separable data and non-linearly separable data using kernel functions. The kernel function transforms the training examples of input space into a higher dimensional feature space. Here we used linear, quadratic, cubic and Gaussian kernel.

In this work, the multi-class SVM is applied. This approach exploits one-against-all method. If the number of classes is *N*, the *N*-class classification by SVM is accomplished by combining *N* two-class classifiers, each discriminating between a specific class and the rest of the training set. During the classification stage, a pattern is assigned to the class with the largest positive distance between the classified pattern and the individual separating hyperplane for the *N* binary classifiers.

*(d) K-nearest neighbour (KNN) classifier.* The principle behind *k* nearest neighbour method [53] is to find a predefined number of training samples closest in distance to the object, and predict the class label of the object from these. The distance based on which the *k* samples are chosen can, in general, be any metric measure. Euclidean distance is the most common choice. A class label is finally assigned to an object based on the majority vote of its *k* neighbours. The number of samples (*k*) can be a user-defined constant. Here we used fine KNN ($k$ = 1, Euclidean distance), medium KNN ($k$ = 10, Euclidean distance), cosine KNN ($k$ = 10, cosine distance), cubic KNN ($k$ = 10, cubic distance) and weighted KNN ($k$ = 10, weighted by the inverse square of the Euclidean distance).

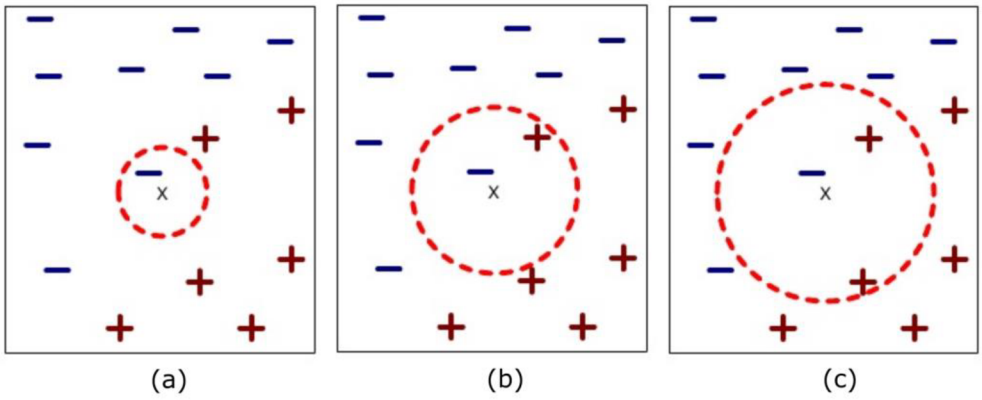


**Figure 10**: *K*-nearest neighbours of an object *x* are *k* data points that have the smallest distance to *x*. (a) 1-nearest neighbour, (b) 2-nearest neighbour and (c) 3-nearest neighbour.

*(e) Ensemble classifiers.* Ensemble methods [54] combine multiple learning algorithms in order to improve generalization and robustness over a single learning algorithm. Common types of ensembles are bagging, boosting and the random subspace ensembles.

In bagging, abbreviated from the bootstrap aggregating, the driving principle is to build several classifiers independently and then to average their results. On average, the combined classifier is usually better than any of the single classifier because its variance is reduced.

Boosting ensemble algorithms create a sequence of models that attempt to correct the mistakes of the models before them in the sequence. Once created, the models make predictions which may be weighted by their demonstrated accuracy and the results are combined to create a final output prediction.

Random subspace ensembles attempt to reduce the correlation between classifiers in an ensemble by training them on random samples of features instead of the entire feature set. Subspace ensembles also have the advantage of using less memory than ensembles with all predictors.

In this work, the following ensemble classification algorithms were used: bagged trees, boosted trees, subspace discriminant and subspace KNN.

*(f) Artificial Neural Network (ANN).* The ANN [55] was inspired by the human learning process and is based on combinations of elementary processors (neurons), each of which takes a number of inputs and generates an output (Figure 11). Each neuron is associated with adaptive weight, i.e. numerical parameter that is tuned by a learning algorithm. The output of the neuron is a function of the weighted sum of inputs. Moreover, neuron is associated with activation function, which defines the output of that neuron given an input or set of inputs.

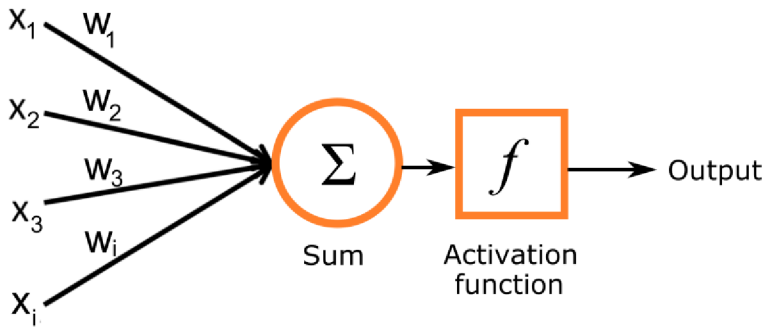


**Figure 11:** Example of neuron with inputs ($x_1$, $x_2$, $x_3$, …, $x_i$) and corresponding weights ($w_l$, $w_2$, $w_3$, …, $w_i$).

The neural network is formed by the collection of interconnected neurons, usually organized in layers, where the output of one neuron becomes the input of other neurons (Figure 12). This architecture resembles the high-level interconnections of elementary neurons in brain. There are many types of network architectures, the common type feed-forward neural network was used in this work.

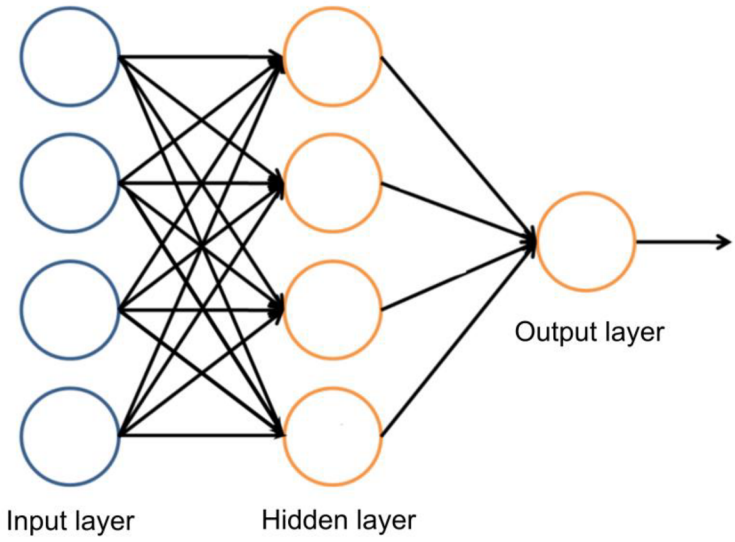


**Figure 12:** Example of three-layered feed-forward neural network structure with one output neuron.

The training process requires a set of labelled examples similarly as in other classification algorithms. In the first step of training, the random numbers are assigned to the weights of the individual neurons. The outputs are computed for the labelled examples serving as a training examples. The outputs are evaluated and compared to given labels based on the selected performance function. Commonly used performance function for neural networks is mean square error [55]. Therefore, the average squared error between

the network outputs and the desired labels is calculated. Afterwards, the weights of neurons are modified properly to obtain better classification results by the backpropagation algorithm. Once the neural network is trained, it can be used for the recognition of unknown input.

Here we used feed-forward backpropagation neural network with one hidden layer containing 10 hidden neurons. The network was trained with scaled conjugate gradient backpropagation algorithm and mean square error was applied as a performance function.

### 6.2.5 Classifier Performance Evaluation

For indication of the performance of a classification algorithm, confusion matrices are a widely used tool [56]. The confusion matrix compares training class labels with output class labels determined by the classification algorithm. Figure 13 shows the example of a confusion matrix for a three-class classifier.


**Figure 13:** Example of confusion matrix for a three-class classifier, correctly classified (green) and incorrectly classified (red) examples.

Several performance parameters can be calculated from the confusion matrix for a classification algorithm: accuracy, precision, recall and F-score.

The classification accuracy of a multi-class classifier is calculated as the ratio of the sum of the principal diagonal values to the sum of all values in the confusion matrix. The accuracy expresses the ratio of correctly classified examples by the classifier and is calculated as follows:

$$Accuracy = \frac{\sum_{i=1}^{N} \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + fn_i}}{N},$$

where $N$ is the number of classes, $tp_i$ is the number of correctly recognized class examples, $tn_i$ is the number of correctly recognized examples that do not belong to the class, $fp_i$ is the number of examples that were incorrectly assigned to the class and $fn_i$ is the number of examples that were not recognized as class examples.

Precision is the ratio of correctly classified positive examples to the total number of positive examples. The precision for multiclass classification task is determined according to the equation:

$$Precision = \frac{\sum_{i=1}^{N} \frac{tp_i}{tp_i + fp_i}}{N},$$

Recall is the ratio of correctly classified positive examples to the all examples in actual class. The recall for multiclass classification is determined as follows:

$$Recall = \frac{\sum_{i=1}^{N} \frac{tp_i}{tp_i + fn_i}}{N},$$

F-score can be interpreted as a harmonic mean of precision and recall, calculated as follows:

$$F_{score} = \frac{2 \times Precision \times Recall}{Precision + Recall},$$

*K*-fold cross-validation was used to evaluate the performance of the classification algorithms. The data were partitioned into *k* randomly chosen subsets of roughly equal size. One subset (testing set) was used to validate the classifier, which had been trained on the remaining subsets (training set). This process was repeated *k* times, such that each subset was used for the validation (we used $k = 5$). Since cross-validation does not use all of the data for training, it is a commonly used method to avoid overfitting.

# 7. Application of Machine Learning to Classification of Cells in QPI

In the last few years, classification of cells by the supervised machine learning became frequently used in biology. However, most of the approaches are based purely on morphometric features, which are not quantitative in terms of cell mass. This may result in poor classification accuracy. The proposed methodology exploiting the quantitative information about the dry mass density distribution within the cell will be applied in the experiment. The obtained results will be compared with the commonly used approach based on the morphometric features.

## 7.1 Experiment Design

Both mentioned classification approaches are tested in the experiment with live adherent eukaryotic cells, which are nutritionally deprived in order to manifest different morphologies for the classification. Since the dry mass density distribution within the viable and nutritionally deprived cells differs markedly, the features extracted from the quantitative phase images play an important role in the classification. The cells are classified using several supervised machine learning algorithms. There is an assumption that most of the classifiers could provide higher performance when quantitative phase features are employed. In such case, the methodology could be a valuable help in refining the monitoring of live cells in an automated fashion.

In the following chapters, methodology for classification of cells based on QPI is demonstrated on the experimental data. The contribution of features extracted from QPI for the classification of cells is evaluated. The approach is compared with a commonly used methods based on morphometric features and the results are discussed.

## 7.2 Cell Culture Techniques

In the experiment, LW13K2 cells (spontaneously transformed rat embryonic fibroblasts) were exposed to conditions that induce nutritional deprivation. The cells were firstly grown attached to a solid surface and maintained in Eagle's minimal essential medium (Sigma-Aldrich, Czech Republic) supplemented with 10% fetal bovine serum (Sigma-Aldrich, Czech Republic) and gentamicin (Sigma-Aldrich, Czech Republic) in an incubator at 37°C and humid 3.5% $CO_2$ atmosphere. The cells were harvested by trypsinization and transferred into 5 sterilised observation chambers μ-Slide I (Ibidi GmbH, Germany). The seeding densities were 20 cells/mm$^2$ in order to achieve sparse coverage for the purposes of segmentation of individual cells. The observation chambers were kept in the incubator under the same conditions.

The culture medium was replaced by phosphate-buffered saline (PBS) after two days. For the experiment, standard PBS (NaCl 8 g/l, KCl 0.2 g/l, $KH_2PO_4$ 0.24 g/l, $Na_2HPO_4$ 1.44 g/l, pH 7.4) was used. PBS deprives cells of nutrients and causes changes in cell morphology. The cells were imaged immediately after PBS application. The same procedure was repeated for all 5 observation chambers.

## 7.3 Image Acquisition

The cells were imaged by CCHM. During the experiment, the samples were illuminated with halogen lamp through the interference filter ($\lambda = 650$ nm, 10 nm FWHM). Microscope objectives (Nikon Plan Fluor 20×/0.5) were utilised for the imaging. At least 100 images were acquired from each sample in pursuit of collecting enough data for the classification. Images were obtained by scanning in a random manner across each sample. All images were gathered in the database, which was used for the classification.

Morphological changes of cells appeared in the order of minutes after the application of PBS. Most of the cells became slightly deprived after 5 minutes. The majority of cells were seriously deprived after 20 minutes. The images of cells were divided by the expert biologist into three categories based on their morphology: viable, semi-deprived and deprived cells. Viable cells did not exhibit any changes in morphology, cells in semi-deprived category were influenced by PBS and started to shrink while their boundaries became indented. The deprived cells, which were influenced the most, adopted a rounded morphology. All images of cells were gathered in the database consisting of 1400 cells. According to the labels assigned by the expert biologist, the database contained the following distribution of class labels based on their morphology: viable (540), semi-deprived (470) and deprived cells (390). The cells with uncertain class membership were excluded from the database. Three distinct types of cell morphologies are shown in Figure 14.

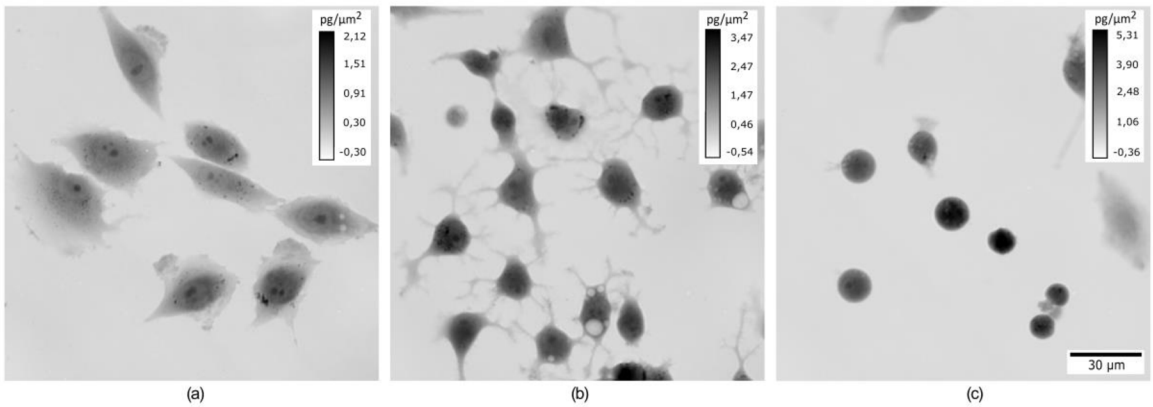


**Figure 14:** Morphological changes of LW13K2 cells induced by PBS. (a) Viable cells, (b) semi-deprived cells and (c) deprived cells. Quantitative phase images are shown in grayscale in units of pg/µm$^2$ recalculated from phase (in radians) according to Davies [7].

## 7.4 Image Pre-processing and Feature Extraction

The cells in the quantitative phase images were segmented from the background by marker-controlled watershed segmentation approach provided by the house-built software. The segmentation has proven to be a crucial step, which affects the performance of the classification. Therefore, we did not consider highly overlapping cells where the segmentation was not clear. The cells located on the border of the image were excluded as well. For the purpose of more accurate cell segmentation, we used sparse seeding densities to obtain subconfluently grown cells. The segmentation results in case of more confluent cell layer were satisfying as well. However, in case of confluent cell layers, there is a higher

chance of less accurate segmentation results, which may lead to poorer performance of the classification. Subsequently, the cells were identified as separate ROIs (cells). From each cell, two sets of cell features were extracted: morphometric and QPI cell features. Two types of feature vectors were composed for each cell, while the first one included only morphometric feature set and the second one both sets. Each feature vector was then assigned one of the class labels determined by the expert biologist. Prior to the classification, the feature values are scaled to a fixed range from 0 to 1 according to Equation (12). The image processing was performed in Matlab.

## 7.5 Feature Selection

Before feeding the feature vectors into the classification algorithms, the potential of extracted features to discriminate between given classes of cells was evaluated by the statistical analysis.

The independent two-sample *t*-test for data with different variances was used to assess whether there are significant differences between the means of parameters of the three cell classes (viable - V, semi-deprived - SD, deprived - D). Since the precondition for parametric analysis (including *t*-test) is the normal distribution of the analysed data, the Shapiro-Wilk test was performed to confirm that the values of cell features are normally distributed. For a visual confirmation, the box-whisker plots of the individual feature values distributions can be seen in Figure 15.

Afterwards, the *t*-test was performed for each feature and between all possible pairs (viable vs. semi-deprived, viable vs. deprived and semi-deprived vs. deprived) in order to investigate whether the defined features are reliable for the distinguishing between the cell classes. The results of the independent two-sample *t*-test for data with different variances are shown in Table 1. The feature values of different cell classes with significant differences between the means indicate the potential for reliable discrimination between the classes.

When tested for the pair V vs. SD class, the statistics rejects the null hypothesis about the equal means at $\alpha = 0.001$ significance level in case of most features. In case of feature *eccentricity (EC)*, the null hypothesis is rejected at $\alpha = 0.05$ significance level, which means that this parameter is less discriminative between these two classes than the rest of the parameters. The exceptions make the features *convex area (CA), perimeter of the convex area ($P_{CA}$), total phase of the cell ($\varphi_{total}$)* and *kurtosis ($Kurt_{\varphi}$)*, for which the differences between the means are not significant and the null hypothesis was not rejected. The situation is visible also from the box-whisker plots (Figure 15). This result is not surprising, since the similarity of mentioned feature values between these two classes is obvious also from the quantitative phase images (Figure 14). The cells in the SD class are more indented, but *CA* is similar to the cells in the V class, and so is $P_{CA}$. The value of $\varphi_{total}$ should not change after the PBS treatment, so it is only expected that the means of this feature will be equal between the classes. When tested for the pair V vs. D class, the null hypothesis about the equal means is rejected at $\alpha = 0.001$ significance level in case of all features except $\varphi_{total}$, which was justified earlier. The same results of the *t*-test statistics can be observed for the pair of classes SD vs. D. Therefore, all the features except $\varphi_{total}$ have the potential for the discrimination between these two classes. From the overall *t*-test analysis it is possible to

assume, that features provide better discrimination between the classes V vs. D and SD vs. D than between the classes V vs. SD. This assumption is in correspondence with the box-whisker plots as well (Figure 15).

**Table 1:** Results of the independent two-sample *t*-test for data with different variances (V - viable, SD - semi-deprived, D – deprived cell class). QPI cell features are highlighted in bold font.

| | V vs. SD | V vs. D | SD vs. D |
|---|---|---|---|
| *FA* | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ |
| $P_{FA}$ | $p < 0.05$ | $p < 0.001$ | $p < 0.001$ |
| *CA* | $p > 0.05$ | $p < 0.001$ | $p < 0.001$ |
| $P_{CA}$ | $p > 0.05$ | $p < 0.001$ | $p < 0.001$ |
| *S* | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ |
| *R* | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ |
| *I* | $p < 0.001$ | $p < 0.05$ | $p < 0.001$ |
| *EC* | $p > 0.05$ | $p < 0.001$ | $p < 0.001$ |
| *EX* | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ |
| **$\mu_\varphi$** | **$p < 0.001$** | **$p < 0.001$** | **$p < 0.001$** |
| **$\varphi_{total}$** | **$p > 0.05$** | **$p > 0.05$** | **$p > 0.05$** |
| **$Var_\varphi$** | **$p < 0.001$** | **$p < 0.001$** | **$p < 0.001$** |
| **$\sigma_\varphi$** | **$p < 0.001$** | **$p < 0.001$** | **$p < 0.001$** |
| **$Skew_\varphi$** | **$p < 0.001$** | **$p < 0.001$** | **$p < 0.001$** |
| **$Kurt_\varphi$** | **$p > 0.05$** | **$p < 0.001$** | **$p < 0.001$** |

It should be noted that QPI cell features derived from the phase (except for $\varphi_{total}$ in case of all pairs and $Kurt_\varphi$ in case of pair V and SD) showed much lower *p*-value (mostly one order of magnitude lower) in the *t*-test than morphometric cell features, which indicates that they have a higher discrimination power. This leads to assumption, that the QPI features could enhance the performance of the classification in comparison to using only morphometric features.

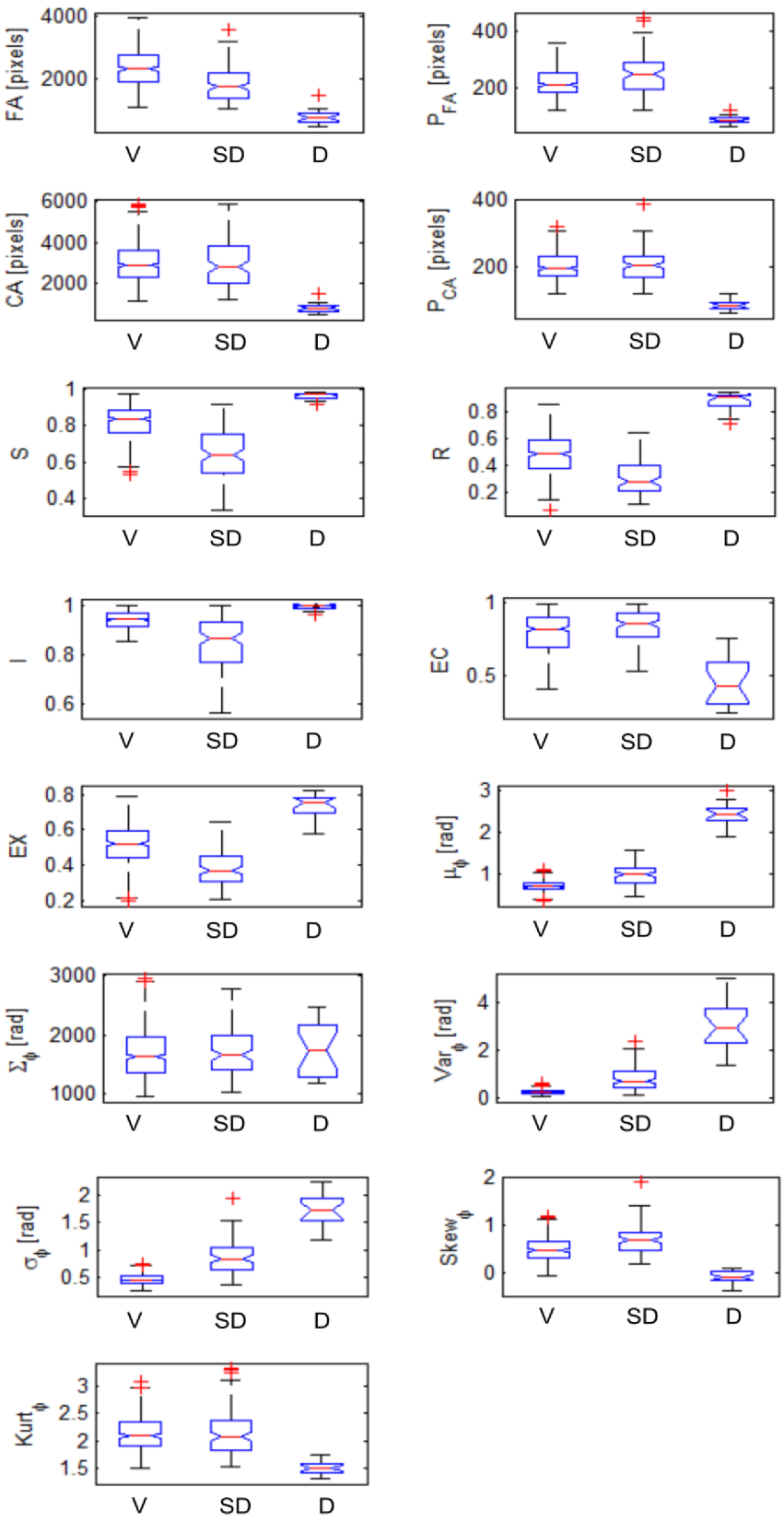


**Figure 15:** Box-whisker plots of the values of features (from the top to the right): footprint area (*FA*), perimeter of the footprint area ($P_{FA}$), convex area (*CA*), perimeter of the convex area ($P_{CA}$), solidity (*S*), roundness (*R*), indentation (*I*), eccentricity (*EC*), extent (*EX*), average phase value ($\mu_\varphi$), total phase value ($\varphi_{total}$), variance of the phase ($Var_\varphi$), standard deviation of the phase ($\sigma_\varphi$), skewness ($Skew_\varphi$) and kurtosis ($Kurt_\varphi$).

## 7.6 Classification Results

The feature vectors representing the cells in the quantitative phase images form an input for the classification algorithms. Before the training of the classifiers, the feature vectors are filtered based on the feature selection in order to contain only discriminative features. Therefore, the feature *total phase value* ($\varphi_{total}$) is eliminated.

Several supervised machine learning algorithms were employed in this work to correctly compare the performance of the classification with two different sets of features. To verify the assumption that QPI features improve the classification performance over the commonly used morphometric features, two types of feature vectors have been used for the classification. In the first case, the feature vector consisted of morphometric features only. In the latter case, also QPI features were added.

Performance measures (accuracy, precision, recall and F-score) of each classification algorithm were determined as a mean of the values obtained by 5-fold cross-validation and can be found in Table 2. The overall performance of the classification for the two types of feature vectors was determined as the mean of performance measure values reached by all classification algorithms. The whole classification procedure was performed in Matlab.

**Table 2:** Performance of the classification by different supervised machine learning algorithms for two types of feature vectors.

| | Accuracy | Precision | Recall | F-score | Accuracy | Precision | Recall | F-score |
|---|---|---|---|---|---|---|---|---|
| | MO features | | | | MO + QPI features | | | |
| Decision trees (complex) | 0.896 | 0.865 | 0.861 | 0.863 | 0.961 | 0.948 | 0.945 | 0.946 |
| Decision trees (medium) | 0.902 | 0.887 | 0.879 | 0.883 | 0.949 | 0.935 | 0.931 | 0.933 |
| Decision trees (simple) | 0.851 | 0.824 | 0.743 | 0.781 | 0.931 | 0.918 | 0.913 | 0.915 |
| Linear discriminant | 0.876 | 0.837 | 0.743 | 0.787 | 0.962 | 0.949 | 0.936 | 0.942 |
| Quadratic discriminant | 0.898 | 0.872 | 0.843 | 0.857 | 0.953 | 0.932 | 0.935 | 0.933 |
| SVM (linear) | 0.892 | 0.879 | 0.870 | 0.874 | 0.948 | 0.928 | 0.931 | 0.929 |
| SVM (quadratic) | 0.885 | 0.840 | 0.830 | 0.835 | 0.958 | 0.944 | 0.943 | 0.944 |
| SVM (cubic) | 0.878 | 0.842 | 0.762 | 0.800 | 0.965 | 0.953 | 0.946 | 0.949 |
| SVM (Gaussian) | 0.893 | 0.889 | 0.855 | 0.872 | 0.963 | 0.957 | 0.947 | 0.952 |
| KNN (fine) | 0.878 | 0.857 | 0.720 | 0.783 | 0.971 | 0.954 | 0.953 | 0.954 |
| KNN (medium) | 0.878 | 0.849 | 0.695 | 0.764 | 0.948 | 0.931 | 0.918 | 0.924 |
| KNN (cosine) | 0.884 | 0.831 | 0.810 | 0.820 | 0.979 | 0.959 | 0.956 | 0.957 |
| KNN (cubic) | 0.913 | 0.894 | 0.890 | 0.892 | 0.953 | 0.941 | 0.939 | 0.940 |
| KNN (weighted) | 0.905 | 0.885 | 0.870 | 0.877 | 0.965 | 0.951 | 0.949 | 0.950 |
| Bagged trees | 0.872 | 0.835 | 0.790 | 0.812 | 0.955 | 0.950 | 0.941 | 0.946 |
| Subspace discriminant | 0.905 | 0.865 | 0.860 | 0.862 | 0.953 | 0.942 | 0.935 | 0.938 |
| Subspace KNN | 0.913 | 0.884 | 0.870 | 0.877 | 0.938 | 0.925 | 0.913 | 0.919 |
| Boosted trees | 0.873 | 0.846 | 0.786 | 0.815 | 0.955 | 0.939 | 0.925 | 0.932 |
| Neural networks | 0.884 | 0.845 | 0.815 | 0.830 | 0.962 | 0.949 | 0.939 | 0.944 |
| **MEAN ± SD** | 0.888 ± 0.015 | 0.859 ± 0.022 | 0.815 ± 0.058 | 0.836 ± 0.039 | 0.956 ± 0.011 | 0.942 ± 0.011 | 0.937 ± 0.012 | 0.939 ± 0.011 |

The overall accuracy of the classification using only morphometric features was 0.888 ± 0.015, which is comparable to values mentioned in the previous studies on cell morphology classification [12,14,57]. The overall precision, recall and F-score were 0.859 ± 0.022, 0.815 ± 0.058 and 0.836 ± 0.039, respectively. The classification using both sets of features led to higher performance of the classifier, with the overall accuracy of the classification reaching 0.956 ± 0.011. In this case, the overall precision, recall and F-score were 0.942 ± 0.011, 0.937 ± 0.012 and 0.939 ± 0.011, respectively.

For comparison of the two classification approaches, the performance results were evaluated by statistical hypothesis testing. The Wilcoxon signed rank test [58] was used as a paired nonparametric statistical hypothesis test which can reveal the existence of significant differences between two distributions. The null hypothesis is that the median difference between pairs of observations is zero. *P*-value 0.05 was considered to be statistically significant. The test revealed significant differences between the two classification approaches ($p < 0.001$) in terms of all performance parameters (accuracy, precision, recall and F-score). The performance results of both approaches are shown in the form of box-whisker plots in Figure 16. The results indicate that QPI cell features enhance the performance of the classification. It should be also noted that in case of employing both QPI and morphometric cell features, the classification performance of all used algorithms has much lower variance than in case of using solely morphometric cell features.

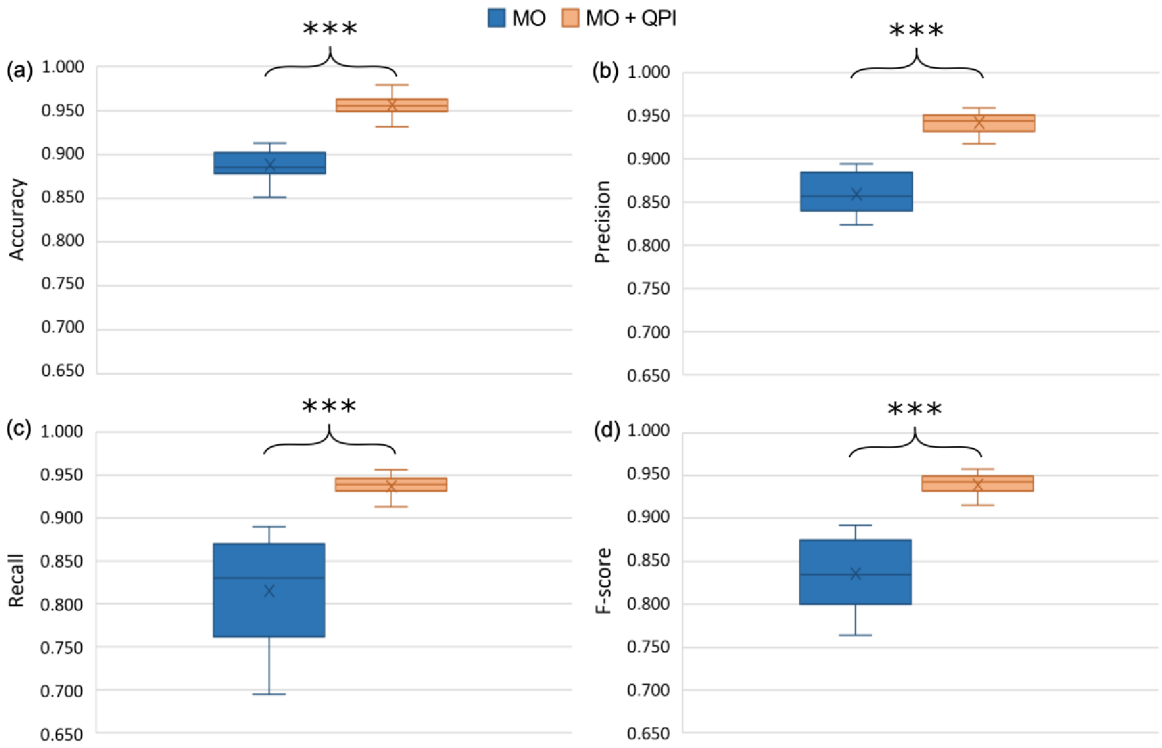


**Figure 16:** Box-whisker plots of overall classification performance for two types of feature vectors: (a) accuracy, (b) precision, (c) recall and (d) F-score. Wilcoxon signed rank test was used for the statistical analysis. Symbols indicating significance are placed above (***: $p < 0.001$).

In order to study the impact of cell sample preparation and other experimental conditions on classification performance, the approach was tested on the data gained from another independent experiment. The experiment was identically designed, however, the

cell preparation was performed by a different person and the classification algorithms were trained on the images of cells from the first experiment. The performance of the classification is summarized in Table 3 together with the results from the first experiment. The performance of the classification on data obtained in two independent experiments was compared by Wilcoxon rank sum test [58], which revealed no significant differences between the classification performance in the two experiments. According to the results, we assume that cell sample preparation and other experimental conditions do not significantly influence the performance of the classification.

**Table 3:** Performance of the classification on data obtained in two independent experiments.

| | Accuracy | Precision | Recall | F-score | Accuracy | Precision | Recall | F-score |
|---|---|---|---|---|---|---|---|---|
| | MO features (mean ± SD ) | | | | MO + QPI features (mean ± SD ) | | | |
| 1st experiment | 0.888 ± 0.015 | 0.859 ± 0.022 | 0.815 ± 0.058 | 0.836 ± 0.039 | 0.956 ± 0.011 | 0.942 ± 0.011 | 0.937 ± 0.012 | 0.939 ± 0.011 |
| 2nd experiment | 0.872 ± 0.022 | 0.846 ± 0.026 | 0.809 ± 0.056 | 0.827 ± 0.041 | 0.949 ± 0.014 | 0.933 ± 0.016 | 0.929 ± 0.014 | 0.931 ± 0.015 |
| Wilcoxon rank sum test | $p>0.05$ | $p>0.05$ | $p>0.05$ | $p>0.05$ | $p>0.05$ | $p>0.05$ | $p>0.05$ | $p>0.05$ |

Based on the overall results, it can be concluded that the quantitative phase information gained by CCHM increases the performance of the classification of cell morphologies in contrast to commonly used methods based on morphometric features. The study shows that CCHM offers preconditions for an accurate classification of cell morphologies, while the main asset of the technique lies in the accurate cell segmentation and the quantitative nature of the images it provides.

Although the performance of the classification in the experiment was rather high, there are several options for the further improvement. One of them is enlargement of the training set, which would enable the classifier to improve the performance by training based on more extensive data. This could however lead to overtraining and worse generalisation for the new examples. The other option is to tune the parameters of the classification algorithms, however, the algorithm tuning is individual for each application. Another options are the extraction of additional features or obtaining extra information from time-lapse QPI. Implementation of these two options will be the main objective of the following chapter.

# 8. Application of Machine Learning to Time-lapse QPI

In the previous chapter, the cells were classified by supervised machine learning based on single-time-point quantitative phase images gained by CCHM. However, some complex dynamic processes demand time-resolved live-cell imaging in order to correctly interpret the cell states. For that reason, in this chapter the methodology of classification will be adjusted in order to gain more information about cell behaviour from the time-lapse images. The time-lapse quantitative phase images of cells will be obtained and additional features, which represent the dynamic cell behaviour in time, will be extracted. Incorporation of time information into the classification process might help overcoming the confusion between different cell states with similar morphology and, therefore, could improve the performance of the classification of cells. However, it may also allow for classification of dynamic cellular processes, or even detection of stages within a process. The applicability of the proposed methodology will be demonstrated in the experiment with time-lapse quantitative phase images of live cells.

## 8.1 Experiment Design

The proposed approach was tested in the experiment with live adherent eukaryotic cells undergoing epithelial-mesenchymal transition (EMT) [59]. The EMT plays important role in cancer research. The cells undergoing EMT lose epithelial characteristics and gain invasive potential with the increased ability to migrate. Two morphologically distinct phenotypes can be observed during EMT: epithelial and mesenchymal. These were the two classes discriminated in the classification.

Most stages of the classification process are similar to those in the classification based on static QPI as shown in Figure 17. However, as the input the time-lapse images gained by CCHM are used in order to take into account also the temporal context of the cell behaviour. In the image pre-processing, the cells are segmented from the background and identified as ROIs. Both morphometric and QPI features are extracted for each ROI (cell). Since the cells were recorded in time, the feature values in several time-instants provide a time series.

There are two possible ways for the representation of time series. In the first one, the values of time series itself represent the input for the classification, which will be referred to as value-based approach. On the other hand, in the feature-based approach, the time series is further represented by the newly defined time-lapse features, which subsequently form time-lapse feature vector. The time-lapse feature vector therefore represents a unique behavioural pattern of a cell and creates an input for the classification. Even though there is an assumption that the feature-based approach is more robust and less sensitive to the amount of noise in the time series than the value-based approach, both options are examined in this work.

In both approaches, the data are further split into training and testing set, while the training data are labelled by expert biologist. The features with the highest potential to distinguish between the given classes are selected and form an input for the classifier. The same set of supervised machine learning algorithms was used for the classification as in the previous case with static QPI.

To compare the perspective of value-based and feature-based approach, both approaches were applied on the data from the experiment with cells undergoing EMT and their performance was evaluated. In order to correctly evaluate the benefit of incorporating the temporal information over the classification based solely on the static QPI, the classification was performed also on the static quantitative phase images from the same experiment. The image processing, feature extraction and classification was performed in Matlab.

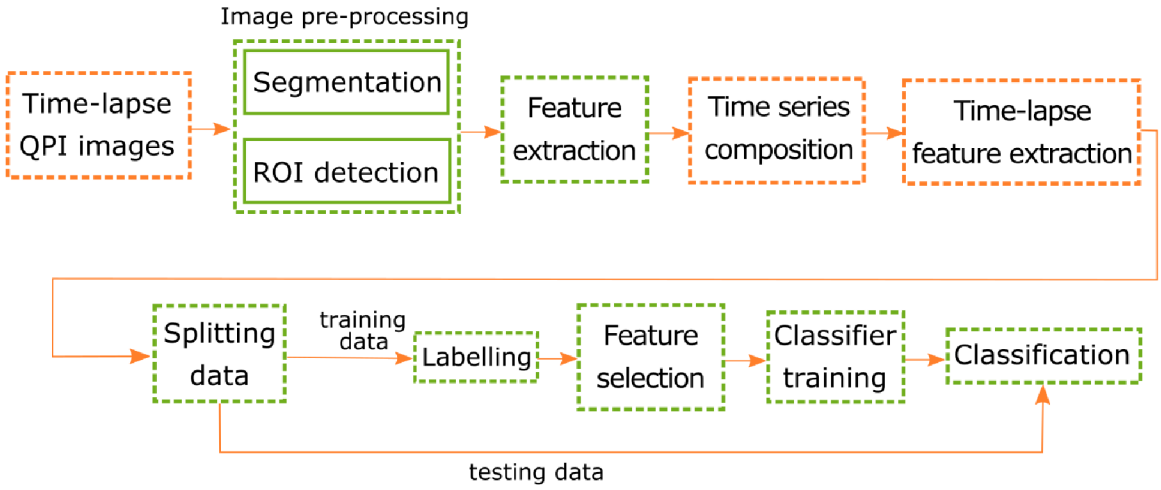


**Figure 17:** Overview of the proposed feature-based classification process based on time-lapse QPI. Firstly, image pre-processing is carried out. The cells in the image are segmented from the background and identified as regions of interest (ROIs). Cell features are extracted for every ROI. Feature values in several time-instants create a time series. Dynamic features are extracted from the time series, while creating the feature vectors representing behaviour of cells. The data are split into training and testing set. The training data are labelled by expert biologist and after the feature selection form an input for the classifier. The classifier is trained on labelled data and prepared to perform the classification on testing unlabelled data.

## 8.2 Epithelial–Mesenchymal Transition

The epithelium is one or more layers of cells with different functions (e.g., cover, respiratory, etc.). The epithelial cells are closely adjacent and take on polyhedral shapes, while being connected by different types of intercellular connections. Epithelial tissue rests on thin extracellular film of fibrils called a basement membrane, which acts as a scaffolding on which epithelium can grow. The basement membrane acts as a selectively permeable membrane that determines which substances will be able to enter the epithelium. Epithelial cells have apical-basal polarity. Such arrangement ensures epithelial integrity and does not allow cells to migrate. On the other hand, the mesenchymal cells are characterized by increased migration capacity, invasiveness and increased production of extracellular matrix components.

In the process of organism development, epithelial–mesenchymal transition (EMT) plays an important role. During EMT, the cells lose their epithelial features and acquire mesenchymal, fibroblast-like properties. Mesenchymal cells show reduced intercellular adhesion and increased motility, which allows them to move away from their epithelial cell community and to infiltrate into surrounding tissue, even at remote locations.

While being an essential process during development, EMT is also occurring under pathological conditions, particularly in fibrosis, wound healing and in invasion and metastasis of carcinomas. For that reason, EMT is considered as an important step in tumour progression and metastatic cascade.

Normal epithelium lined by a basement membrane can proliferate locally to give rise to an adenoma as shown in Figure 18. Additional transformation by epigenetic changes and genetic modifications leads to a carcinoma in situ, still outlined by an intact basement membrane. Subsequently, the carcinoma cells can be locally disseminated after undergoing EMT. After the EMT, the cells weaken their intercellular adhesion and gain mobility which results in increased cellular migration and tissue changes. After the basement membrane becomes fragmented, the cells can penetrate into the bloodstream (intravasation) allowing them the transport to distant organs (extravasation). At secondary locations, the carcinoma cells that retain the ability to survive and divide can form a new carcinoma by means of a complementary process called mesenchymal-epithelial transition (MET) [60].

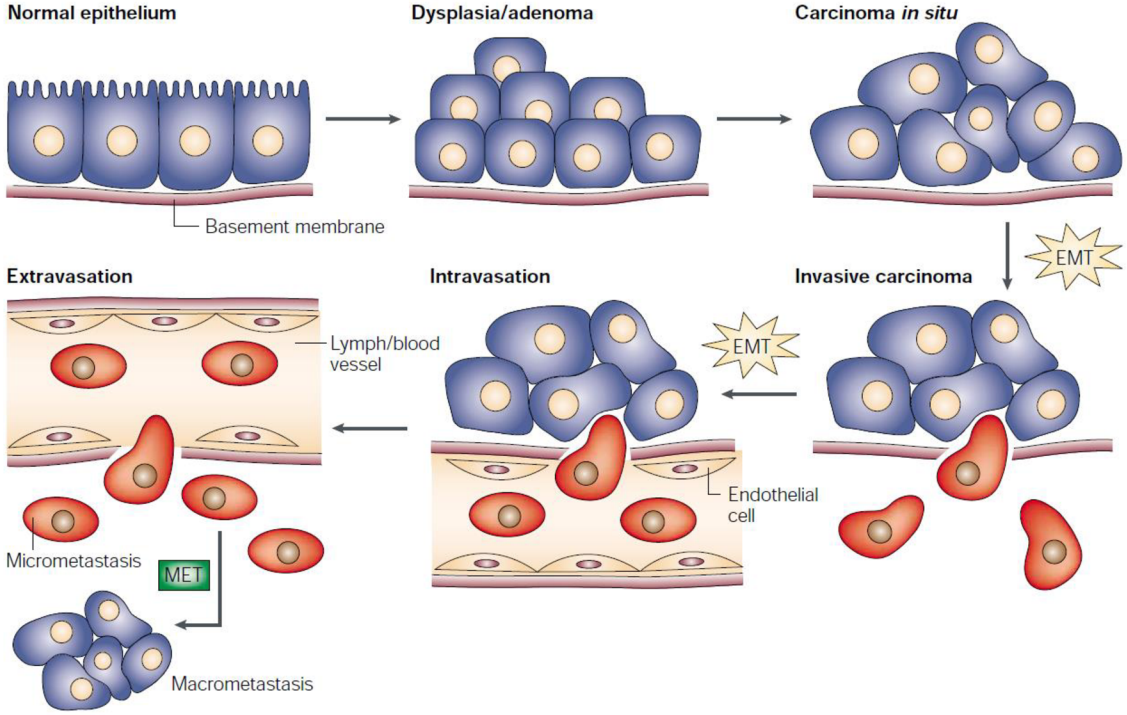


**Figure 18:** The role of EMT and MET in the progression of carcinoma. Normal epithelia on the basement membrane can proliferate to give rise to an adenoma. Further transformation may lead to a carcinoma. The EMT can induce local dissemination of carcinoma cells, while the basement membrane becomes fragmented. The cells penetrate into lymph or blood vessels and can transport to distant organs. At secondary sites, carcinoma cells can extravasate and form a new carcinoma through MET [60].

The transformation of epithelial cells into mesenchymal (Figure 19) is regulated by a sequence of strictly controlled molecular processes. At first, the intercellular junctions break down, then the apical–basal polarity changes to front–rear, the cytoskeleton is reorganized and the cell shape changes. The epithelial gene program is attenuated and the genes determining the mesenchymal phenotype are activated. The cells have increased cell mobility and invasiveness including the ability to produce extracellular matrix (ECM). Cells

that have undergone EMT have increased resistance to cell aging and apoptosis (programmed cell death).

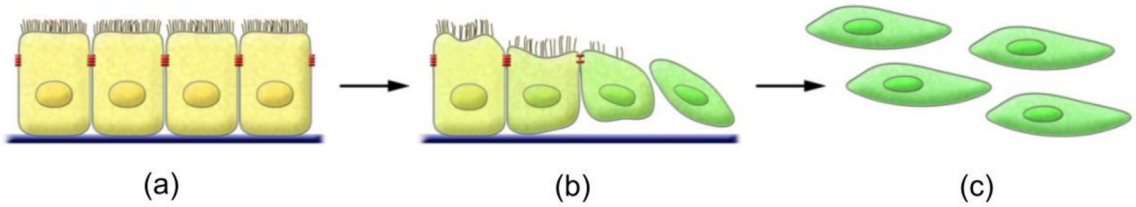


**Figure 19:** The steps of epithelial-mesenchymal transition (EMT). Polarized epithelial cells (a) lose their epithelial characteristics and reduce intercellular junctions and polarity (b). The cells acquire mesenchymal phenotype (c). The change is accompanied by degradation of the basal membrane [59].

EMT in cells may be induced by different physical, chemical or biological factors. EMT-inducing molecules include growth factors, cytokines, hormones, and ECM. The most well-known and most explored growth factor inducing EMT is transforming growth factor beta (TGF-β). The EMT in various epithelial cells can be induced by adding TGF-β to epithelial cells in culture [61].

The process of EMT is still not well understood and remains a subject for further research. The automated analysis of cells undergoing EMT based on QPI could have a significant meaning for the its study.

## 8.3 Cell Culture Techniques

The experiment was performed in cooperation with the research group "Molecular cancer and stem cell therapeutics" at Karolinska Institutet. For the experiment, NMuMG cells (normal murine mammary gland epithelial cells) provided by Karolinska Institutet were used. The cells were firstly grown attached to a solid surface and maintained in Dulbecco's modified Eagle's medium (Sigma-Aldrich, Czech Republic) supplemented with GlutaMAX$^{TM}$ (Life Technologies, Czech Republic), 10% fetal bovine serum (Sigma-Aldrich, Czech Republic), 100 U/ml penicillin and 0.1 mg/ml streptomycin (Life Technologies, Czech Republic). The cells were grown in an incubator at 37°C and humid 3.5% $CO_2$ atmosphere. The cells were harvested by trypsinization and transferred into 6 sterilised observation chambers μ-Slide I (Ibidi GmbH, Germany). The seeding densities were 50 cells/mm$^2$ in order to achieve sparse coverage for the purposes of segmentation of individual cells. The observation chambers were kept in the incubator under the same conditions. The chambers were imaged the next day after. The first three chambers were directly imaged, while in the other three chambers, TGF-β with the concentration 5ng/ml was added prior to the imaging. The cells were imaged immediately after TGF-β application.

## 8.4 Image Acquisition

The NMuMG cells were imaged by CCHM. During the experiment, the samples were illuminated with halogen lamp through the interference filter (λ = 650 nm, 10 nm FWHM). Microscope objectives (Nikon Plan Fluor 20×/0.5) were utilised for the imaging. For the purpose of classification, it was essential to acquire reasonably large number of cells

undergoing EMT, therefore, six fields of view were imaged with the interval 5 minutes. Each chamber was imaged for 48 hours to obtain the time-lapse QPI for the classification. The cells in three chambers (control) were imaged in the cultivation media without any intervention. In the other three chambers, the cells were exposed to the TGF-β during imaging.

The cells in the control chamber preserved characteristic epithelial morphology for the whole duration of the experiment. The cells in the chamber with added TGF-β started to change the morphology approximately 17 hours after the application of TGF-β. The cells became elongated and adopted mesenchymal morphology. These two morphologies later represented the categories for the classification.

All time-lapse images of cells were gathered in the database. The database consisted of six 48 hour-long records. Since none of the cells remained in the field of view for the whole imaging, 150 minutes (30 time-lapse images with interval 5 minutes) were determined as an optimal length of the time-lapse record for one cell. 100 cells were chosen for the monitoring. Based on their morphology, the cells were labelled by the expert biologist as either epithelial (48 cells) or mesenchymal (52 cells). The cells with uncertain class membership were not considered and were excluded from the database. The two types of classified cell morphologies are shown in Figure 20.

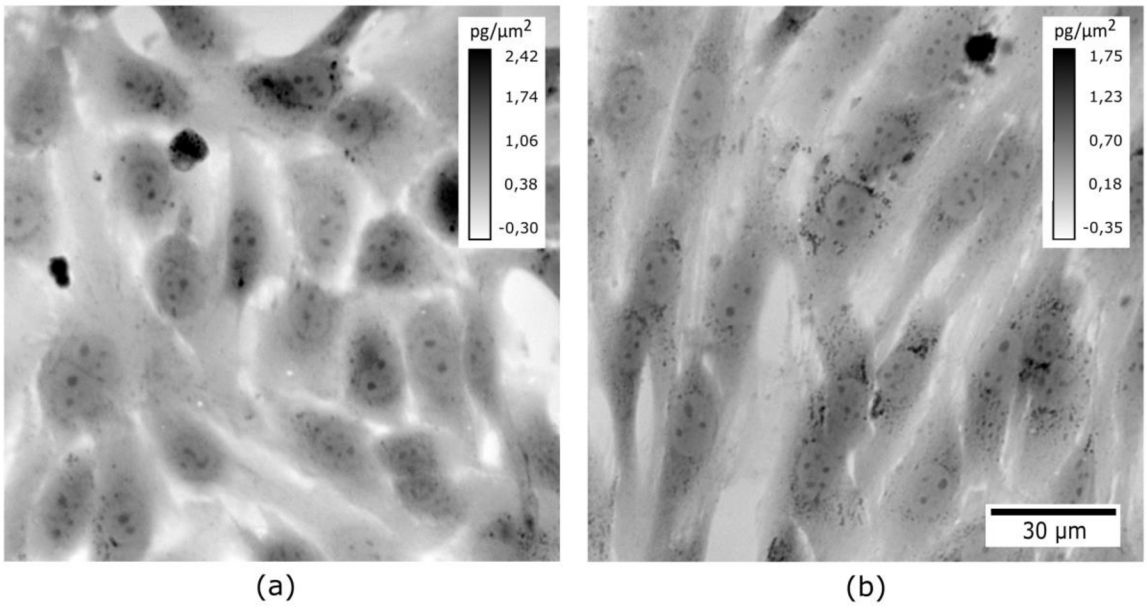


**Figure 20:** Examples of quantitative phase images of epithelial (a) and mesenchymal (b) phenotype gained by CCHM.

## 8.5 Image Pre-processing and Feature Extraction

The cells in the time-lapse quantitative phase images were segmented from the background by marker-controlled watershed segmentation in the same way as in the previous case with static images. The individual cells were tracked using the cell tracking algorithm scripted in Matlab. The algorithm performs cell tracking by linking every segmented cell in the given frame to the nearest cell in the next frame. Again, we did not consider highly overlapping cells where the segmentation was not clear. The cells located on the border of the image were excluded as well. We also considered only cells staying in the field of view

for the whole time determined by the experiment design. Subsequently, the cells were identified as separate ROIs (cells).

Two types of cell features were extracted from each ROI: morphometric and QPI features. Each cell in one time instant is therefore represented by a feature vector composed of these cell features. Since every cell was recorded in time, each cell feature provides a univariate time series composed of the values of cell features over time. Considering all cell features therefore gives rise to a multivariate time series. The example of the multivariate time series composed of 11 time-lapse images of one cell can be seen in Figure 21.

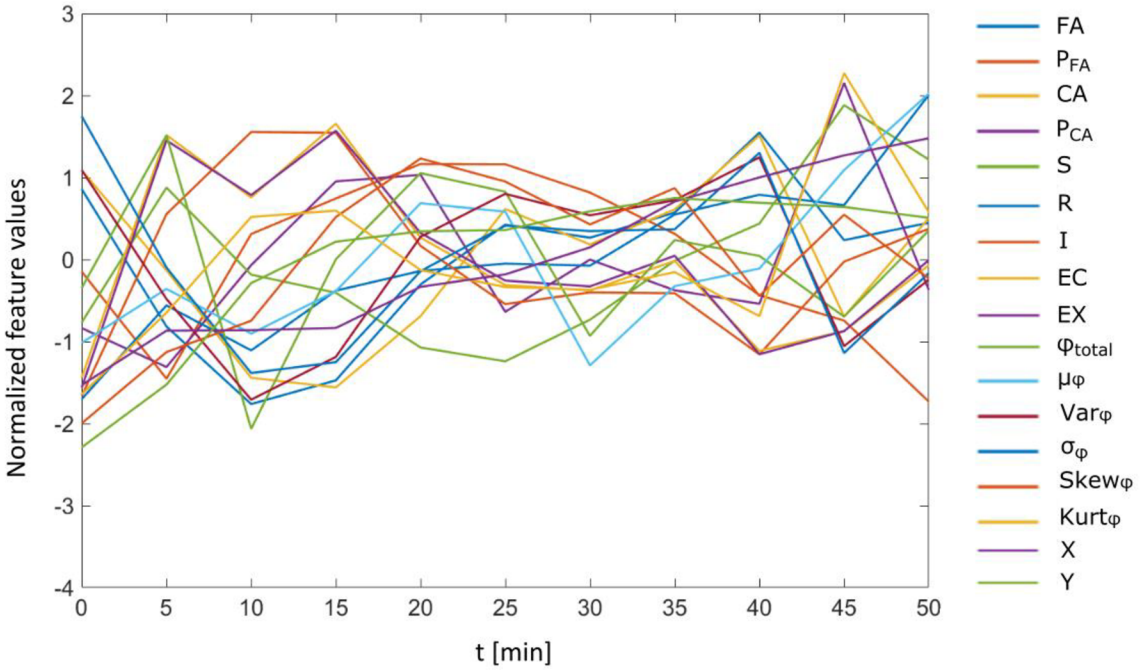


**Figure 21:** Example of the multivariate time series segment. Each univariate time series is composed of the feature values obtained within 50 min with 5 min interval. Footprint area (*FA*), perimeter of the footprint area ($P_{FA}$), convex area (*CA*), perimeter of the convex area ($P_{CA}$), solidity (*S*), roundness (*R*), indentation (*I*), eccentricity (*EC*), extent (*EX*), total phase of the cell ($\varphi_{total}$), average phase ($\mu_\varphi$), variance ($Var_\varphi$) and standard deviation of the phase ($\sigma_\varphi$), skewness ($Skew_\varphi$), kurtosis ($Kurt_\varphi$), centroid X (*X*), centroid Y (*Y*).

### 8.5.1 Time-lapse Feature Extraction

In order to explain the formation of the final time-lapse feature vector in the feature-based approach, the brief notation will be introduced. Let $\mathbf{X} = \{X_1, X_2, \ldots, X_Q\}$ represent a collection of $Q$ multivariate time series, where $Q$ is the number of cells in the experiment. Each multivariate time series $X_i$ is formed by $n$ observations ($n$ is the number of time points) and $d$-dimensional variable ($d$ is the number of cell features) as shown in Figure 22. The multivariate time series $X_i$ can be written as

$$X_i = \{X_{ijt}\}, \qquad \text{for } j = 1, \ldots, d;\ t = 1, \ldots, n, \tag{16}$$

with the total number of observations $dnQ$.

We will consider the $j$-th component of the $i$-th time series $X_{ij} = \{X_{ij1}, \ldots, X_{ijn}\}$ to be a univariate time series. Therefore, the univariate time series will be composed of the values of one cell feature recorded in time. For each univariate time series $X_{ij,}$ a partial time-lapse

feature vector $M = (m_1, m_2, \ldots, m_L)$ is formed, where each $m$ is a time-lapse feature extracted from the time series and $L$ is the number of time-lapse features. In this way, each time series $X_{ij}$ is transformed into a partial time-lapse feature vector $M_{ij}$.
Each multivariate time series is therefore transformed into $d$ $M$-vectors. The vectors are then concatenated into a final time-lapse feature vector of $dL$ dimensions. Such feature vector therefore represents a unique behavioural pattern of a cell.

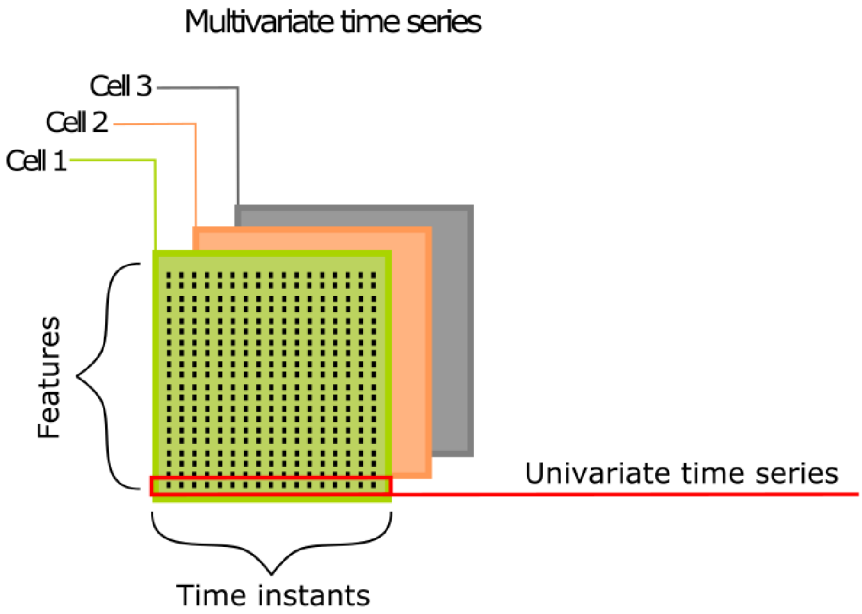


**Figure 22:** Illustrative demonstration of the multivariate time series representing cell behaviour. Each cell is represented by the multivariate time series composed of univariate time series (formed by cell feature values obtained within a defined time period).

There are several possible methods used for dealing with feature-based representation of the time series. The employed feature extraction techniques are briefly described in the following paragraphs.

*(a) Statistical features*

The statistical features carry the information about the time series in terms of global picture. The following metrics were chosen in order to statistically represent the structure of the time series: mean value, median value, standard deviation, minimum value, maximum value, skewness and kurtosis.

*(b) Fourier transform features*

The basic idea of spectral decomposition is that any time series can be represented by the superposition of a finite number of sine (and/or cosine) waves, where each wave is represented by a single complex number known as a Fourier coefficient as illustrated in the Figure 23. The Fourier transform [62] therefore generates an approximation to a time series using as a basis cosine and sine functions with frequency $\omega_j$. The Fourier transform approximates a time series as follows:

$$\hat{X}(t) = \sum_{j \in B} \left( a_j \cos \omega_j t + b_j \sin \omega_j t \right), \tag{17}$$

where $\hat{X}(t)$ is a time series which creates an approximation to $X(t)$, $B$ is a subset of the frequencies within the basis, $a_j$ is the coefficient related to the cosine basis function with frequency $\omega_j$, and $b_j$ is the coefficient related to the sine basis function with frequency $\omega_j$. Fast Fourier transform (FFT) algorithm was employed for the time series representation. The features extracted by the Fourier transform for the purpose of classification contain the coefficient pairs $a_j$ and $b_j$ for each frequency $\omega_j$ in $B$. The representation of the time series is therefore in the frequency domain. There are many advantages to that, one of them is data compression. A signal of length $n$ can be decomposed into $n$ sine/cosine waves that approximate the original time series. However, many of the Fourier coefficients have very low amplitude and thus contribute little to approximation of the time series. Only the largest coefficients are chosen and are stored as the time-lapse features, thereby producing compression.

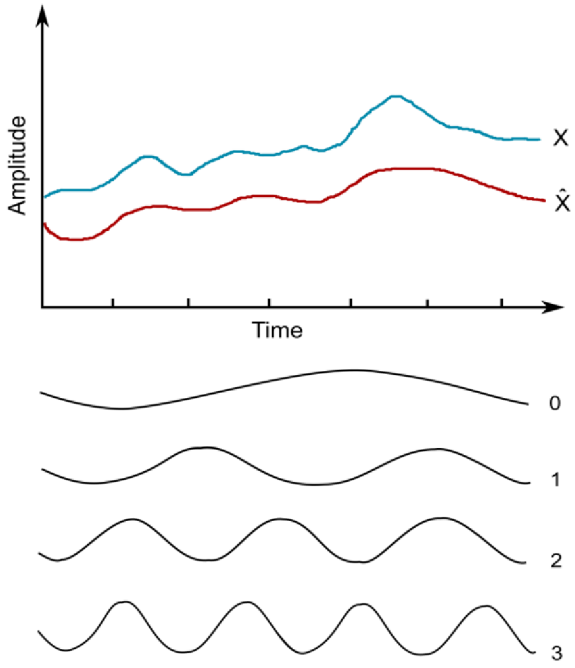


**Figure 23:** The illustration of application of the Fourier transform on the time series. The first four Fourier bases can be combined in a linear combination to produce $\widehat{\boldsymbol{X}}$, an approximation of the time series *X*.

*(c) Wavelet transformation features*

The wavelet transform [63] uses a basis containing waveforms that are localized in space and, therefore, is more suitable for approximating time series including local structures than Fourier transformation. The wavelet transform uses a basis including $n$ (length of the time series) waveforms. The basis waveforms are derived from scaling and translations of a mother wavelet $\psi$. The wavelet transform can be thought as a cross-correlation of a signal with a set of wavelets of various scales at different time positions. The transforms are ordered based on the degree of localization in the resulting basis waveform such that $j$ is the $j^{th}$ transform. With the increasing $j$, the basis waveforms become more localized (detailed). The wavelet transformation approximates a time series as follows:

$$\hat{X}(t) = \sum_{j \in B} \phi_j \psi_j(t), \tag{18}$$

where $\hat{X}(t)$ is a time series which creates an approximation to $X(t)$, $B$ is a subset of the transforms of $\psi$ within the basis, and $\phi_j$ is the coefficient related to the basis waveform $\psi_j$. Algorithm computing discrete wavelet transform (DWT) was employed for the time series representation. The features extracted by the wavelet transform contain the approximation coefficients $\phi_j$ for each transform $j$ in $B$. While the first few coefficients contain an overall, coarse approximation of the data, the additional coefficients represent the details in the original time series. The approximation better represents the data as the number of transforms in $B$ increases. The largest coefficients are chosen and saved as the time-lapse features.

*(d) Trend*

The trend is represented by the coefficients obtained by the linear least squares fitting of the time series. The trend characterizes a long-term change in the mean value of the cell feature.

*(e) Entropy*

Approximate entropy is a method for estimating the complexity of time series data. It quantifies the unpredictability of fluctuations in the time series. The presence of repetitive patterns of fluctuation in a time series renders it more predictable than a time series in which such patterns are absent. A time series containing many repetitive patterns has a relatively small approximate entropy while a less predictable process has a higher value.

*(f) Symbolic aggregate approximation features*

The symbolic aggregate approximation (SAX) method [64] has been developed to reduce the dimensionality of a time series into a short chain of symbols. SAX is composed of two steps: piecewise aggregate approximation (PAA) [65] and the conversion of a PAA sequence into a string composed of letters. PAA divides the original time series of length $n$ into $w$ equally spaced segments and computes the mean values for each segment. The sequence assembled from the mean values is the PAA representation of the original time series where the number of dimensions was reduced from $n$ to $w$. Each segment is subsequently mapped into a symbol (letter) corresponding to the region in which it resides. The length of segments and alphabet size (number of symbols used) are two parameters to be specified. As such, the original time series is converted to a symbol string.

All so far mentioned time-lapse features were extracted from each of the univariate time series and created a partial time-lapse feature vector as shown in Figure 24.


**Figure 24:** Time-lapse feature extraction from the univariate time series. Extracted time-lapse features are assembled into a partial time-lapse feature vector. Individual segments represent the group of time-lapse features obtained by the particular extraction technique. The length of the segments indicates the approximate number of extracted time-lapse features for the particular group.

Subsequently, the partial time-lapse feature vectors obtained from each univariate time series were concatenated into a final time-lapse feature vector, while other extracted time-

lapse features (principal components analysis and motion features) were added on the tail as shown in Figure 25. The principal components analysis and motion features were extracted in a different way, which will be described in the next sections.

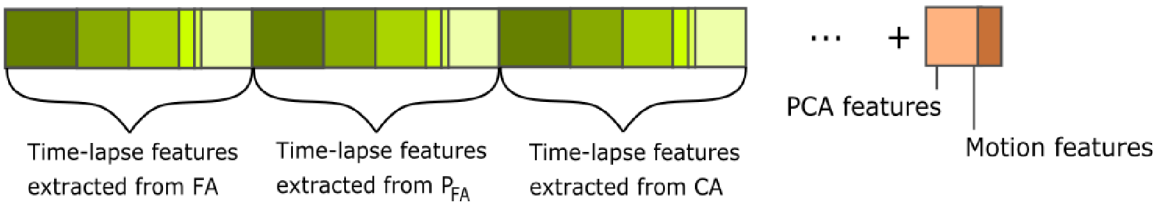


**Figure 25:** Final time-lapse feature vector construction. The final feature vector representing a single cell is formed by concatenation of partial time-lapse feature vectors obtained from univariate time series of QPI and morphological cell features. In addition, the motion and PCA features are added.

*(g) Principal components analysis features*

Principal components analysis (PCA) [66] is a statistical technique used to eliminate the less significant components (features) and reduce the data representation only to the most significant ones. While the other mentioned time-lapse feature extraction techniques were applied on the univariate time series formed by the cell feature values recorded in time, PCA was applied on the whole multivariate time series. PCA maps the multivariate data into a lower dimensional space. Given $n$ observations of $d$ features in the multivariate time series, the goal of PCA is to reduce the dimensionality of the data matrix by finding $r$ new variables, where $r$ is less than $d$. Termed principal components, these $r$ new variables together account for as much of the variance in the original $n$ variables as possible while remaining mutually uncorrelated and orthogonal. After the PCA, the values in the columns are coefficients of the principal components that are related to each of the $n$ time points. Only the first $k$ principal components are kept stored in the final time-lapse feature vector, since they contain most of variance in the data.

*(h) Motion features*

The motion characteristics such as accumulated distance, Euclidean distance, motion speed or directionality of the cell movement were calculated from the cell centroids.

*Accumulated distance* is the overall distance travelled by the cell between the initial and the end point as shown in Figure 26 and is calculated as

$$d_a = \sum_{i=1}^{n} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2},$$

where $n$ is the number of time points in which the $x$ and $y$ coordinates were recorded.

*Euclidean distance* is defined as the length of the straight line between the cell starting and end point and is calculated as

$$d_{Euclid} = \sum_{i=1}^{n} \sqrt{(x_{end} - x_{ini})^2 + (y_{end} - y_{ini})^2}.$$

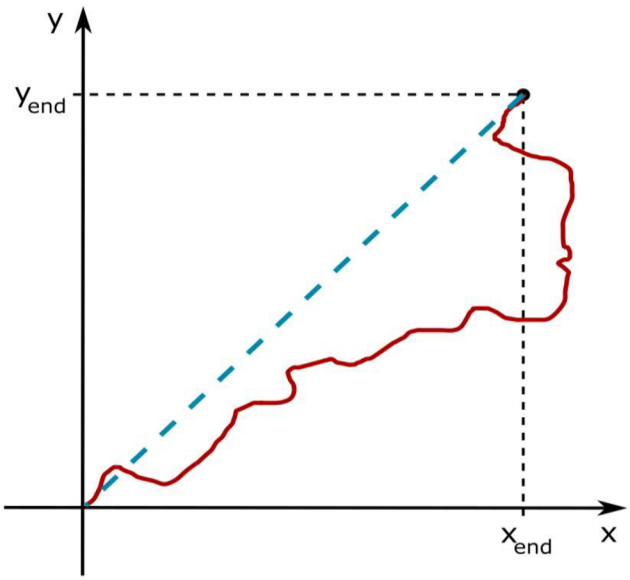


**Figure 26:** Representation of the Euclidean distance and accumulated distance. The red line depicts the accumulated distance and the blue line Euclidean distance.

*Velocity* of the cell motion is calculated as the overall distance travelled by the cell over the elapsed time:

$$v = \frac{d_a}{t}.$$

*Directionality* of the cell motion is calculated by comparing the *Euclidian distance* to the *accumulated distance* as follows:

$$D = \frac{d_{Euclid}}{d_a}.$$

The values of directionality closer to zero report about indirect motion, while he values close to one indicate straight motion. The described motion characteristics were added into the final time-lapse feature vector.

In the value-based approach, the extraction of time-lapse features is omitted, since the final time-lapse feature vector is composed of the raw data (values in each time point) contained in the multivariate time series. The final time-lapse feature vector is created by concatenating the univariate time series behind each other.

In both approaches, the final time-lapse feature vector represents a unique behavioural pattern of a cell. Before passing the vectors to the classification algorithms, the time-lapse feature values are scaled to a fixed range from 0 to 1 according to Equation (12). The example of a set of final time-lapse feature vectors gained by feature-based approach can be seen in Figure 27, where the first 32 rows represent feature vectors extracted from epithelial cells and the other 35 rows from mesenchymal cells with the columns representing individual time-lapse feature values. The data are further split into training and testing set, while the training data are labelled by expert biologist. Since the final time-lapse feature vectors are of substantial size, the next step is the selection of features with the highest potential to distinguish between the given classes, which would then form an input for the machine learning classification algorithms.

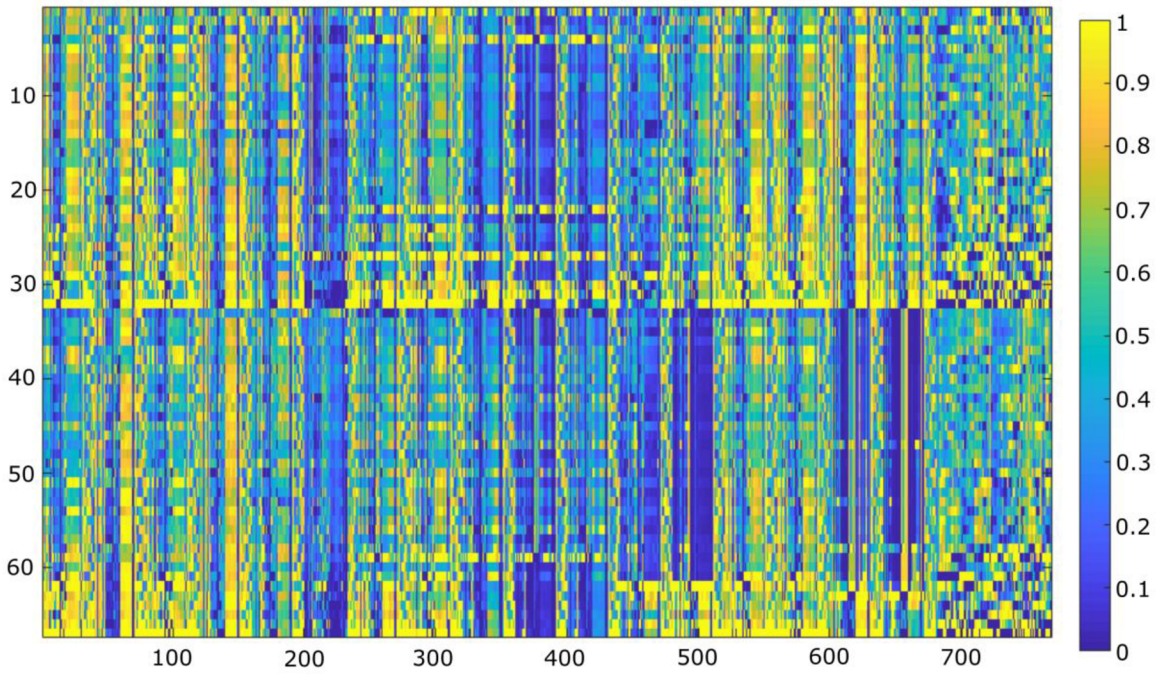


**Figure 27:** Example of the final time-lapse feature vectors concatenated into matrix. Elements of the matrix contain the (normalized) time-lapse feature values, and are visualized using colour: from blue (low values) to yellow (high values). First 32 rows represent time-lapse feature vectors extracted from epithelial cells and the other 35 rows from mesenchymal cells.

## 8.6 Feature Selection

In case of classification of cells based on static QPI, the feature selection was done by statistical analysis of each feature, which would estimate the potential of the particular feature to distinguish between given classes. Since here the time-lapse feature vectors are composed of considerably higher number of features (more than 660 and 500 features in feature-based and value-based approach, respectively), the feature selection is performed in an automated and more effective manner.

Moreover, in this case, when the number of observations is limited in comparison to large number of features, the large amount of features is not desirable for producing a desired learning result and the limited observations may lead the learning algorithm to overfit to the noise. Reducing the number of features is therefore in this case an essential step before the classification. Moreover, the reduction of features leads to lower computation complexity, which makes the whole process less time-consuming.

Several methods exist for the feature selection [67] and can be clustered into two groups: filter methods and wrapper methods. Filter methods depend on general characteristics of the data in the evaluation and selection of features, while not involving the learning algorithm. On the other hand, the wrapper methods use the performance of the chosen learning algorithm to evaluate the potential of individual features. Wrapper methods search for features better fit for the chosen learning algorithm, which leads to the final reduced set of feature optimized for the specific classification algorithm. Since we use several algorithms for the classification in this work, this is not desirable. Moreover, the

wrapper methods can be significantly slower than filter methods. For that reason, we apply the filter approach for the feature selection.

Firstly, the *t*-test was applied on each feature and the *p*-value for each feature was compared as a measure of the feature's ability to discriminate between the two classes. To estimate the order of class separation by the features, the empirical cumulative distribution function (CDF) of the *p*-values was plotted. CDF of the *p*-values for the feature-based approach is shown in Figure 28. There are approximately 15% of features, which have the *p*-values close to zero and 30% of features having the *p*-values smaller than 0.05. It can be concluded that there are roughly 200 features in the original time-lapse feature set, which have a potential to separate the two cell classes.

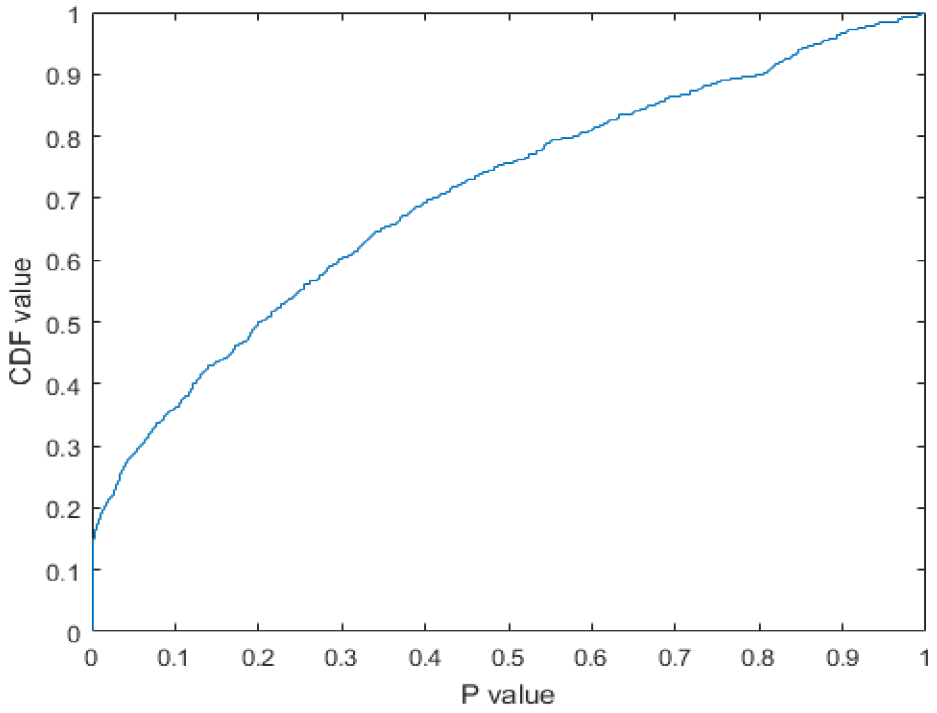


**Figure 28:** Cumulative distribution function of the *p*-values for all time-lapse features gained by feature-based approach.

In the value-based approach (Figure 29), there are approximately 18% of features, which have the *p*-values close to zero and 30% of features having the *p*-values smaller than 0.05. CDF of the *p*-values showed that there are roughly 150 features from the original time-lapse feature set having rather high discriminative power.

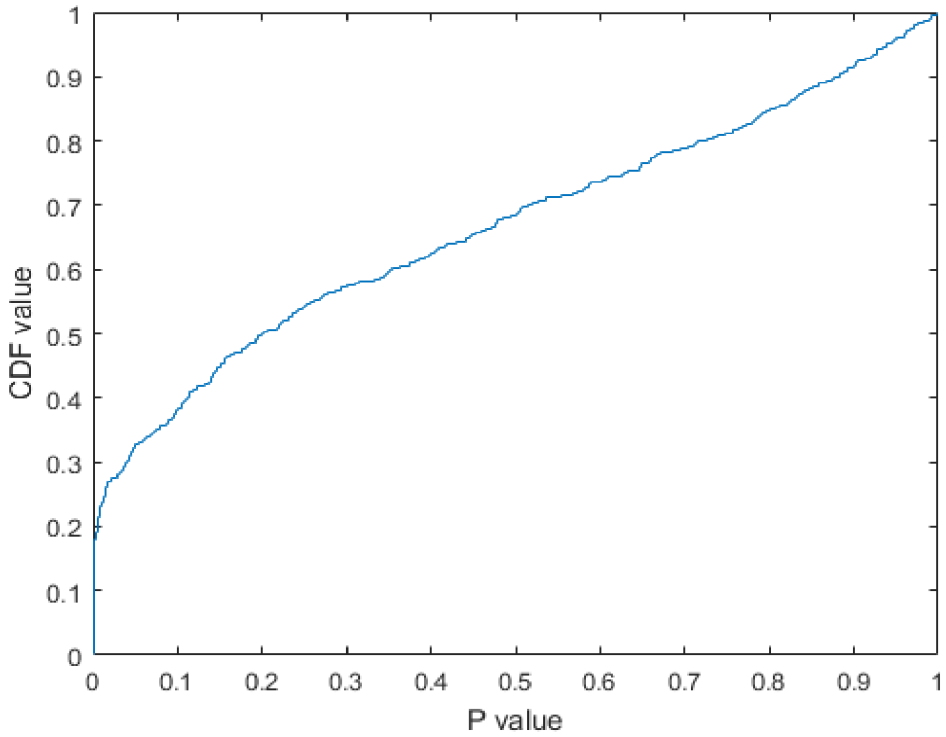


**Figure 29:** Cumulative distribution function of the *p*-values for all time-lapse features gained by value-based approach.

The features were subsequently ordered by their $p$-values. In order to define the appropriate number of features to be selected, the classification error (the number of misclassified observations divided by the number of observations) as a function of the number of features was plotted. To obtain the classification error, several classification algorithms were employed. The results of the classification error in feature-based and value-based approach when using SVM are shown in Figure 30 and Figure 31, respectively. The classification error was computed for different numbers of features between 2 and 30. The final number of selected features was determined as the mean value of the results produced by employing different classification algorithms.

In feature-based approach, the filter feature selection method obtains the smallest classification error when 10 features are engaged. Only these 10 features with the highest discriminative power are kept in the reduced time-lapse feature vectors used for the classification. In value-based approach, 12 features were determined as optimal.

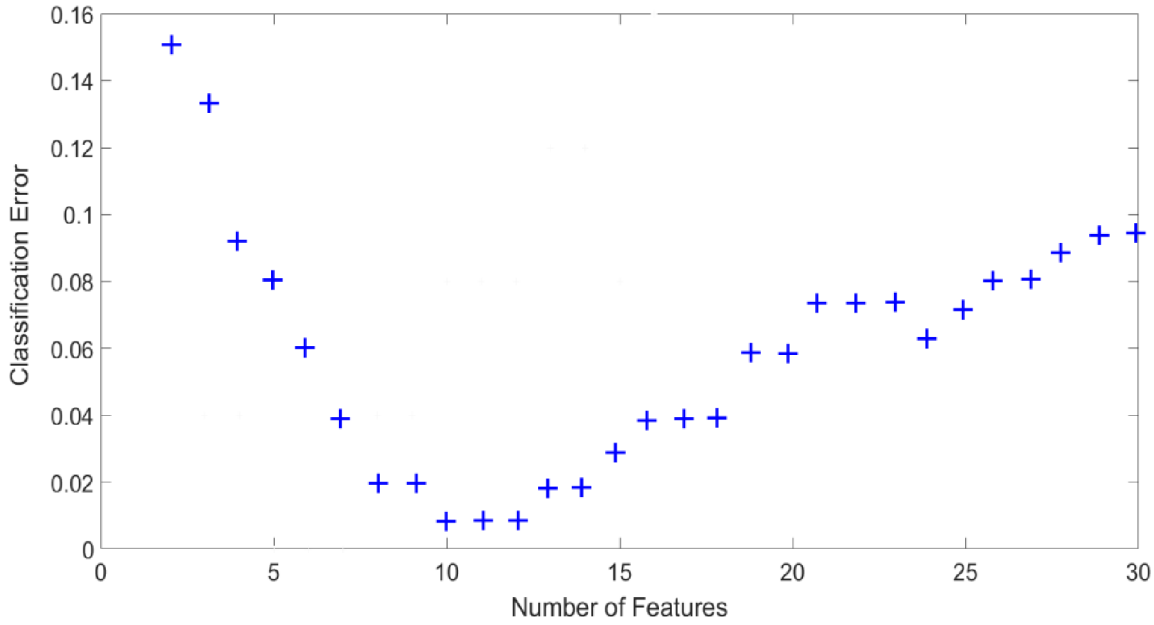


**Figure 30:** The classification error as a function of the number of features (using SVM classifier) in feature-based approach.

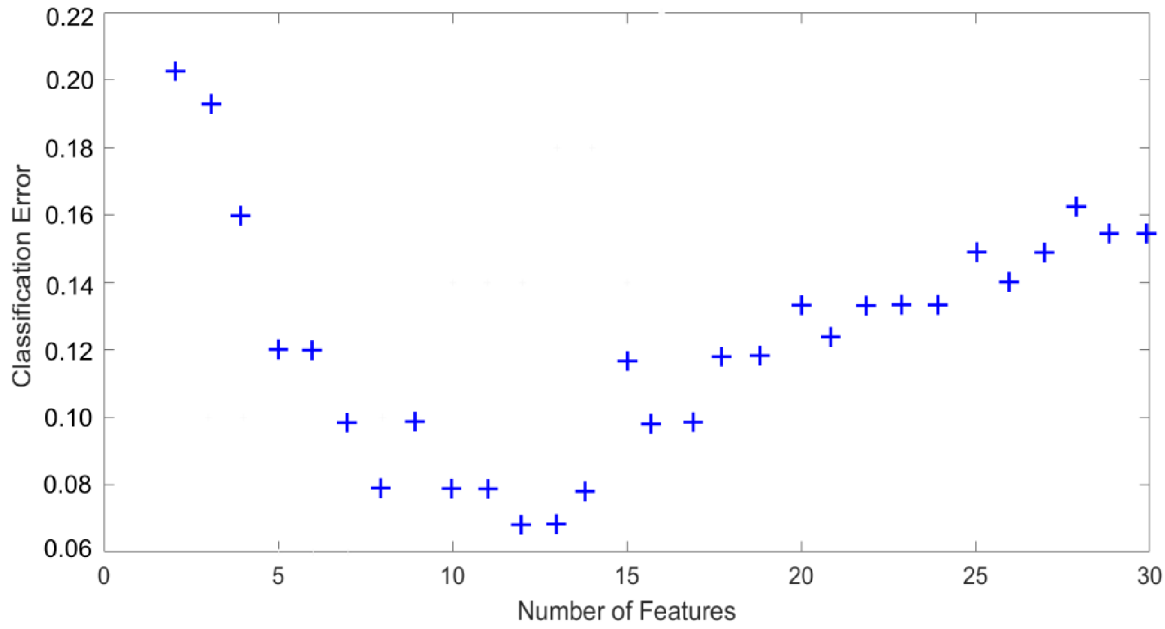


**Figure 31:** The classification error as a function of the number of features (using SVM classifier) in value-based approach.

## 8.7 Classification Results

After the features with the highest potential to distinguish between the epithelial and mesenchymal cell classes were selected, they create the input for the classification algorithms. The same set of supervised machine learning algorithms was used for the classification as in the previous case with static QPI.

The classification was firstly performed on the reduced time-lapse feature vectors gained by value-based approach. The same procedure was repeated for the reduced time-

lapse feature vectors gained by feature-based approach. Moreover, the classification was performed also on the features extracted from the static images in order to evaluate the potential of the methodology based on time-lapse QPI.

Performance measures (accuracy, precision, recall and *F*-score) of each classification algorithm were determined as a mean of the values obtained by 5-fold cross-validation. The overall performance of the classification was determined as the mean of performance measure values reached by all classification algorithms. The whole classification procedure was performed in Matlab.

The performance of the classification implementing the value-based approach is summarized in Table 4. The overall accuracy of the classification was 0.924 ± 0.054. The overall precision, recall and *F*-score were 0.908 ± 0.052, 0.883 ± 0.089 and 0.894 ± 0.071, respectively.

**Table 4:** Performance of the classification by different supervised machine learning algorithms using the value-based approach.

| | Accuracy | Precision | Recall | F-score |
|---|---|---|---|---|
| Decision trees (complex) | 0.782 | 0.832 | 0.628 | 0.716 |
| Decision trees (medium) | 0.883 | 0.842 | 0.828 | 0.835 |
| Decision trees (simple) | 0.893 | 0.862 | 0.831 | 0.846 |
| Linear discriminant analysis | 0.972 | 0.958 | 0.954 | 0.956 |
| Quadratic discriminant analysis | 0.915 | 0.896 | 0.886 | 0.891 |
| SVM (linear) | 0.950 | 0.941 | 0.933 | 0.937 |
| SVM (quadratic) | 0.971 | 0.961 | 0.952 | 0.956 |
| SVM (cubic) | 0.982 | 0.978 | 0.972 | 0.975 |
| SVM (Gaussian medium) | 0.986 | 0.982 | 0.979 | 0.980 |
| KNN (fine) | 0.948 | 0.939 | 0.924 | 0.931 |
| KNN (medium) | 0.918 | 0.896 | 0.889 | 0.892 |
| KNN (cosine) | 0.936 | 0.892 | 0.876 | 0.884 |
| KNN (cubic) | 0.891 | 0.854 | 0.843 | 0.848 |
| KNN (weighted) | 0.882 | 0.832 | 0.828 | 0.830 |
| Bagged trees | 0.822 | 0.826 | 0.703 | 0.760 |
| Subspace discriminant | 0.954 | 0.938 | 0.943 | 0.940 |
| Subspace KNN | 0.978 | 0.961 | 0.959 | 0.960 |
| Boosted trees | 0.936 | 0.914 | 0.908 | 0.911 |
| Neural networks | 0.958 | 0.941 | 0.939 | 0.940 |
| **MEAN ± SD** | 0.924 ± 0.054 | 0.908 ± 0.052 | 0.883 ± 0.089 | 0.894 ± 0.071 |

The performance of the classification using the feature-based approach is summarized in Table 5. Representing the cell behaviour by the time-lapse features led to higher performance of the classifier as in the case of value-based approach, with the overall accuracy of the classification reaching 0.976 ± 0.011. In this case, the overall precision, recall and *F*-score were 0.966 ± 0.014, 0.960 ± 0.013 and 0.963 ± 0.014, respectively.

**Table 5:** Performance of the classification by different supervised machine learning algorithms using the feature-based approach.

| | Accuracy | Precision | Recall | F-score |
|---|---|---|---|---|
| Decision trees (complex) | 0.979 | 0.968 | 0.961 | 0.964 |
| Decision trees (medium) | 0.985 | 0.978 | 0.971 | 0.974 |
| Decision trees (simple) | 0.988 | 0.982 | 0.979 | 0.980 |
| Linear discriminant analysis | 0.966 | 0.958 | 0.952 | 0.955 |
| Quadratic discriminant analysis | 0.979 | 0.963 | 0.957 | 0.960 |
| SVM (linear) | 0.948 | 0.948 | 0.942 | 0.945 |
| SVM (quadratic) | 0.988 | 0.984 | 0.979 | 0.981 |
| SVM (cubic) | 0.987 | 0.986 | 0.972 | 0.979 |
| SVM (Gaussian medium) | 0.989 | 0.986 | 0.981 | 0.983 |
| KNN (fine) | 0.968 | 0.958 | 0.942 | 0.950 |
| KNN (medium) | 0.986 | 0.982 | 0.965 | 0.973 |
| KNN (cosine) | 0.984 | 0.979 | 0.969 | 0.974 |
| KNN (cubic) | 0.966 | 0.954 | 0.948 | 0.951 |
| KNN (weighted) | 0.968 | 0.949 | 0.946 | 0.947 |
| Bagged trees | 0.985 | 0.98 | 0.969 | 0.974 |
| Subspace discriminant | 0.978 | 0.961 | 0.952 | 0.956 |
| Subspace KNN | 0.959 | 0.94 | 0.938 | 0.939 |
| Boosted trees | 0.968 | 0.951 | 0.948 | 0.949 |
| Neural networks | 0.976 | 0.965 | 0.958 | 0.961 |
| **MEAN ± SD** | 0.976 ± 0.011 | 0.966 ± 0.014 | 0.960 ± 0.013 | 0.963 ± 0.014 |

In order to correctly evaluate the benefit of incorporating the temporal information over the classification based solely on the static QPI, the classification was performed also on the static quantitative phase images of cell undergoing EMT. The static QPI images were obtained from the time-lapse data by selecting one image from each time-lapse sequence. The classification of epithelial and mesenchymal cells based on the static QPI was performed according to the methodology described in Section 6.2. The performance of the classification based on single-time-point QPI is summarized in Table 6. The overall accuracy of the classification was 0.890 ± 0.052. The overall precision, recall and *F*-score were 0.874 ± 0.054, 0.839 ± 0.100 and 0.855 ± 0.078, respectively.

**Table 6:** Performance of the classification by different supervised machine learning algorithms using the static QPI.

| | Accuracy | Precision | Recall | F-score |
|---|---|---|---|---|
| Decision trees (complex) | 0.886 | 0.837 | 0.832 | 0.834 |
| Decision trees (medium) | 0.892 | 0.863 | 0.885 | 0.874 |
| Decision trees (simple) | 0.926 | 0.931 | 0.921 | 0.926 |
| Linear discriminant analysis | 0.958 | 0.944 | 0.941 | 0.942 |
| Quadratic discriminant analysis | 0.895 | 0.845 | 0.823 | 0.834 |
| SVM (linear) | 0.916 | 0.899 | 0.882 | 0.890 |
| SVM (quadratic) | 0.885 | 0.963 | 0.951 | 0.957 |
| SVM (cubic) | 0.945 | 0.935 | 0.928 | 0.931 |
| SVM (Gaussian medium) | 0.938 | 0.915 | 0.908 | 0.911 |
| KNN (fine) | 0.897 | 0.872 | 0.846 | 0.859 |
| KNN (medium) | 0.763 | 0.811 | 0.620 | 0.703 |
| KNN (cosine) | 0.895 | 0.866 | 0.852 | 0.859 |
| KNN (cubic) | 0.794 | 0.795 | 0.622 | 0.698 |
| KNN (weighted) | 0.793 | 0.760 | 0.658 | 0.705 |
| Bagged trees | 0.869 | 0.838 | 0.772 | 0.804 |
| Subspace discriminant | 0.893 | 0.852 | 0.841 | 0.846 |
| Subspace KNN | 0.942 | 0.922 | 0.902 | 0.912 |
| Boosted trees | 0.938 | 0.921 | 0.912 | 0.916 |
| Neural networks | 0.889 | 0.842 | 0.838 | 0.840 |
| **MEAN ± SD** | 0.890 ± 0.052 | 0.874 ± 0.054 | 0.839 ± 0.100 | 0.855 ± 0.078 |

The performance of the classification obtained by the mentioned classification approaches were compared by statistical hypothesis testing. The Wilcoxon signed rank test was used in order to reveal the significant differences between the three distributions. The null hypothesis is that the median difference between pairs of observations is zero. *P*-value 0.05 was considered to be statistically significant. The test revealed very significant differences between the feature-based and value-based time-lapse classification approaches ($p < 0.001$) in terms of all performance parameters (accuracy, precision, recall and *F*-score). Significantly different results ($p < 0.001$) were obtained also from the classification based on static QPI and the classification based on time-lapse QPI employing the feature-based approach. According to the test, the classification based on static QPI and the classification based on time-lapse QPI using the value-based approach provided different performance of the classification with a lower significance ($p < 0.01$ for precision and $p < 0.05$ for other performance parameters). The methodology based on time-lapse QPI employing the feature-based approach appears superior in terms of the classification performance in comparison to other two approaches. The classification based on time-lapse QPI using the value-based approach reached slightly lower performance, however it outperforms the classification based on static QPI, which does not consider the temporal information. The performance results of all approaches are shown in the form of box-whisker plots in Figure

32. It should be noted that the methodology based on time-lapse QPI employing the feature-based approach shows much lower variance of the classification performance achieved by different algorithms than other two approaches.

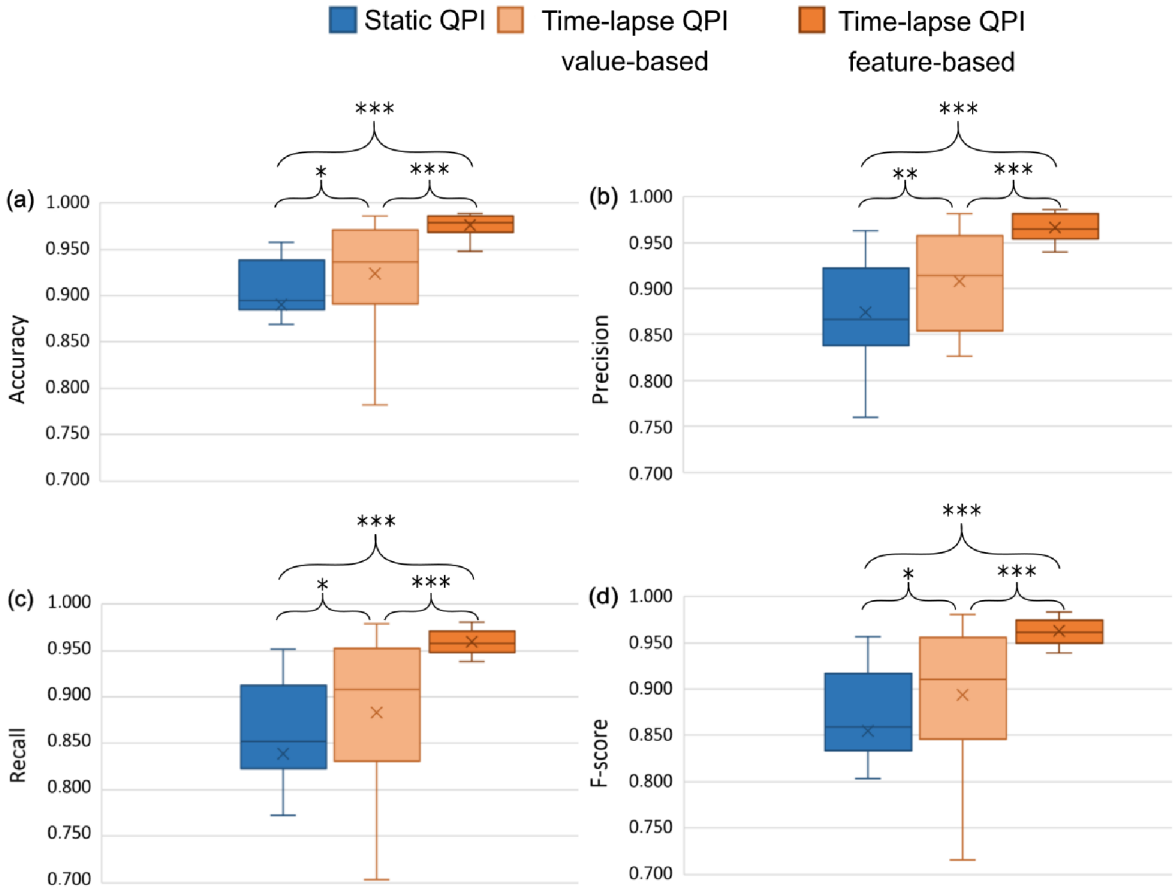


**Figure 32:** Box-whisker plots of overall classification performance of classification based on static QPI, time-lapse QPI (value-based and feature-based approach): (a) accuracy, (b) precision, (c) recall and (d) *F*-score. Wilcoxon signed rank test was used for the statistical analysis. Symbols indicating significance are placed above (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).

Several conclusions can be drawn from the results of the classification. The classification based on time-lapse QPI using either value-based or feature-based approach outperforms the classification based on static QPI, which does not consider the temporal information. Hence, taking into account the time information appears to improve the classification of the two cell phenotypes.

However, when it comes to the classification based on time-lapse QPI, the feature-based approach outperforms the value-based approach. The low performance values in the value-based approach can be a consequence of many factors. The main reason might be that the features, which are in this case the raw time series data, do not fully represent the cell behaviour. The other possibility is the increased sensitivity of this approach to the amount of noise in the time series.

Although the performance of the classification based on time-lapse QPI using the feature-based approach was rather high, further improvement could be achieved by

enlargement of the time-lapse QPI dataset, which would allow the classification algorithms to improve the training based on more extensive data.

Even though the methodology was demonstrated only in the experiment with cells undergoing EMT, the approach might also contribute to higher performance when it comes to different classification tasks.

Furthermore, it should be noted that the results of the experiment also have the significant meaning for the study of EMT. Even these days, the process of EMT is still not well understood and therefore it is a subject for many currently performed studies. To my present knowledge, this has been the first time the cells undergoing EMT were monitored by digital holographic microscopy. The classification of cell phenotypes and therefore determining the EMT stages based on QPI may contribute to the study of EMT mechanisms and help understand the whole process, which would unquestionably play important role for the cancer research.

# 9. Conclusions

The thesis focuses on the application of supervised machine learning for the interpretation of the quantitative phase images obtained by CCHM. The objective was to define a methodology, which would assist during the analysis of live cell behaviour by means of CCHM, exploiting the quantitative nature of the images it provides. Several partial steps were achieved towards this objective.

Firstly, the methodology for the classification of cells in the single-time-point quantitative phase images was proposed. Two types of cell features characterising the cell behaviour were defined and extracted from the quantitative phase images. Commonly used morphometric features, describing mostly the shape and morphology of the cell, represent the first type. The ability of CCHM to provide quantitative phase contrast enables to extract the other type of features, which are based on the phase distribution in the cell region and provide information about the dry mass density distribution within the cells. These features are referred to as QPI features. The performance of the proposed methodology was demonstrated in the experiment with deprived cells, while three types of cell morphologies were being distinguished. After the pre-processing steps, both mentioned types of cell features were extracted from the quantitative phase images and their potential to discriminate between defined classes was assessed both visually using box-whisker plots and statistically by Welch's *t*-test. Several supervised machine learning algorithms were used for the classification. The results from the classification based on QPI features were compared with the classification based on commonly used morphometric features. Based on the results it was assumed that the classification employing also quantitative phase information outperforms the commonly used method based solely on the morphometric features.

However, in order to take into account also the dynamics of monitored cells, the methodology based on time-lapse quantitative phase images was proposed. In this method, the time-lapse features representing the cell dynamic behaviour were proposed and extracted, with the best features selected for the classification. Two approaches for the time-lapse feature extraction were used: value-based and feature-based approach. Both approaches were tested and compared in the experiment with live cells undergoing epithelial-mesenchymal transition. Moreover, both approaches were compared with the methodology based on the single-time-point quantitative phase images. The results showed, that the methodology based on time-lapse images exploiting the feature-based approach outperforms the other methods. Despite of the challenging time-lapse feature extraction, the proposed approach based on incorporating the temporal information into the classification process provides a novel, yet efficient way to classify the cells in quantitative phase images with promising performance results.

However, it is worth to note that even though considering of the temporal context resulted in better performance in this particular experiment, it is highly possible, that the methodology based on single-time-point images may be sufficient or even become superior in case of different classification tasks. Such outcome may occur in the classification tasks, where the temporal information does not play important role and the morphology of cells belonging to different classes is clearly distinguishable from one image. In such case, where

the single-time-point image is a sufficient source of information, one should prefer this approach in order to avoid the extra complexity of the time-lapse approach. The performance of both methodologies might be, therefore, dependent on the particular classification task.

Even though the interpretation of cell behaviour in quantitative phase images by means of supervised machine learning was presented only for the two specific classification tasks, both proposed methodologies might also contribute to higher performance when it comes to different tasks. There are several applications for which the methodology could mean a valuable help, e.g. the monitoring of cell live cycle, cell death, reaction of cells to treatment, interaction of cells with material (biocompatibility testing), detection of different experimental conditions or distinguishing different cell lines. The detailed description of possible applications are provided in Section 10.

Although the performance of the classification in the experiments was rather high, there are several options for the further improvement, e.g. enlargement of the training set, tuning of the classification algorithms or extraction of additional features. The foremost goal of the future work will be the implementation of these proposals, which will be further discussed in Section 10.

The overall outcomes suggest that CCHM offers strong preconditions for an accurate automated analysis of live cell behaviour, while the main asset of the technique lies in the quantitative nature of the images it provides. I believe that this work might provide a stepping stone for the high-throughput automated analysis of specific cell behaviour by means of CCHM. The future aim is to define a complex tool, which would provide assistance during the analysis of live cell behaviour in the laboratory of Experimental Biophotonics. I believe that such tool could strengthen the role of CCHM as a valuable microscopy technique for automated analysis of live cell behaviour, and contribute to the promoting this microscopy technique as a standard diagnostic method in biology and medicine. The next steps, which are necessary for further progress towards this direction are summarized in the future outlook.

It should be noted, however, that this work does not represent a complete summary of my work during the PhD study, but rather a major part of it. During that period of time, I focused on several other projects in the Experimental Biophotonics research group, e.g. the study of adhesion of normal human dermal fibroblasts to the cyclopropylamine plasma polymers by CCHM [28], quantitative phase imaging of plasmonic metasurfaces [30] or vortex topographic microscopy for full-field reference-free imaging and testing [68]. Part of the results gained during my PhD study were published in 4 peer-reviewed scientific journals with impact factor and presented at 8 conferences (6 foreign and 2 domestic). The complete list of publications can be found in Section 12.

# 10. Future Outlook

The proposition and implementation of the methodology for classification of cells by means of machine learning in both time-lapse and single-time-point quantitative phase images serves as a solid foundation for continuing research in the field of automated interpretation of quantitative phase images. As was already mentioned in the previous section, there are several options for the further improvement. This section contains a discussion of some of those options along with the proposed solutions.

First of all, future work could focus on adding new types of features found to be useful for time series classification, since there are several other features discussed in the literature, which could lead to higher performance of the classification.

It is possible that different classification algorithms will achieve different results depending on the classification task. It would be beneficial to know which algorithm performs the best in a specific task. To address this issue, experiments can be performed to map the most appropriate classification algorithms to the specific applications, so that each application would have a preferred algorithm that is likely to be most effective.

The other option is the refinement of the classification algorithms by tuning their parameters. However, the parameters of an algorithm might be also dependent on the classification task. Further experiments could clarify the best possible tuning for specific applications.

So far, the classification of time series data was performed on rather small number of examples. The enlargement of the database of quantitative phase images and, therefore of the training set, might enable the classifier to improve the performance by training based on more extensive data. This could however lead to overtraining and worse generalisation for the new examples. The additional experiments are needed to evaluate the potential contribution of additional training data.

Another important issue regarding the classification based on time-lapse data is the determination of appropriate intervals for image acquisition. The suitable intervals might be different when it comes to diverse dynamic cell processes and each classification task should have the determined interval to be used. In future, intervals with variable length will be tested and the appropriate intervals will be estimated for the specific classification tasks.

Furthermore, the feature selection process could be enhanced. So far, only filter approach was implemented in order to select the features with the highest discriminative power. Future work might therefore involve the application of wrapper methods, which would consider the performance of a specific classification algorithm for a particular classification task.

So far, only the supervised machine learning algorithms were implemented for the classification. In the future, the potential of unsupervised machine learning and deep learning (by convolutional neural networks) will be studied.

Another possible improvement can be seen in optimization of the Matlab code. Maximizing code performance might speed up the whole process.

As it has been already mentioned previously, the presented methodology can also be used for investigating other cell states or events. Example of further applications may include the following:

- *Detection of apoptosis*. Apoptotic cells exhibit characteristic changes related to cell shrinkage and nuclear fragmentation. Detection of these changes could contribute to automated analysis of cell viability during various experiments.

- *Classification of apoptotic versus oncotic cells.* The classification of these two types of cell death could be important for diagnostics, dose-response, and toxicological studies. Some experiments have already been performed regarding the distinction between apoptosis and oncosis by means of QPI [69], however machine learning has not been employed so far.

- *Drug testing*. Testing the drugs by monitoring of cell reaction to the applied treatment is a commonly used method in cancer therapy. The classification of treated versus resistant cells would help to automate drug testing, which would make the personalised cancer therapy more feasible in the future.

- *Biocompatibility assessment.* The interaction of cells with engineered biomaterials plays an important role for the biomaterials development and bioengineering. Here the automated evaluation of material biocompatibility could be applied by classifying viable versus less viable cells interacting with the material.

- *Mitosis detection*. Detection of mitosis, or its stages could contribute to automated monitoring of cell live cycle.

- *Cancer diagnostics*. Several classification tasks that could contribute to diagnosis of cancer are proposed here:
    - classification of healthy versus cancer cells,
    - classification of primary cancer cells versus metastatic cells (the capability of QPI to differentiate between the classes has already been studied in [70]),
    - distinguishing cells with different metastatic potential (the pilot study has been performed in [71]).

- *Detection of different experimental conditions* based on altered cell behaviour.

- *Indication of diseased cells.* It has been shown that information measured by QPI can be used as an effective indicator to quantitatively analyse the

physical and chemical alterations in diseased red blood cells [72]. The process has not been yet automatized by machine learning.

The future work will focus on employing the proposed methods in the mentioned experiments, while the ultimate goal will be development of the complex system for the assistance during the analysis and interpretation of live cell behaviour by CCHM.

# 11. References


[1] C.M. Bishop, Pattern recognition and machine learning (Information Science and Statistics), 1st edn. 2, Springer, New York, 2006.

[2] T. Slabý, P. Kolman, Z. Dostál, M. Antoš, M. Lošťák, R. Chmelík, Off-axis setup taking full advantage of incoherent illumination in coherence-controlled holographic microscope, Optics Express. 21 (2013) 14747–62.

[3] P. Kolman, R. Chmelík, Coherence-controlled holographic microscope, Optics Express. 18 (2010) 21990–22003.

[4] G. Popescu, Quantitative phase imaging of cells and tissues, McGraw Hill, New York, 2011.

[5] M. Mir, B. Bhaduri, R. Wang, R. Zhu, G. Popescu, Quantitative phase imaging, in: E. Wolf (Ed.), Progress in Optics, Elsevier, 2012: pp. 133–217.

[6] R. Wayne, Light and video microscopy, Elsevier/Academic Press, 2013.

[7] H. Davies, M. Wilkins, Interference microscopy and mass determination, Nature. 169 (1952) 541.

[8] R.C. Mellors, R. Silver, A microfluorometric scanner for the differential detection of cells: application to exfoliative cytology, Science. 114 (1951).

[9] D. Comaniciu, P. Meer, D.J. Foran, Image-guided decision support system for pathology, Machine Vision and Applications. 11 (1999) 213–224.

[10] B. Swolin, P. Simonsson, S. Backman, I. Lofqvist, I. Bredin, M. Johnsson, Differential counting of blood leukocytes using automated microscopy and a decision support system based on artificial neural networks - evaluation of DiffMasterTM Octavia, Clinical and Laboratory Haematology. 25 (2003) 139–147.

[11] K. Rajpoot, N. Rajpoot, SVM optimization for hyperspectral colon tissue cell classification, in: Springer, Berlin, Heidelberg, 2004: pp. 829–837.

[12] D.K. Das, C. Chakraborty, B. Mitra, A.K. Maiti, A.K. Ray, Quantitative microscopy approach for shape-based erythrocytes characterization in anaemia, Journal of Microscopy. 249 (2013) 136–149.

[13] Z. Saeedizadeh, A. Mehri Dehnavi, A. Talebi, H. Rabbani, O. Sarrafzadeh, A. Vard, Automatic recognition of myeloma cells in microscopic images using bottleneck algorithm, modified watershed and SVM classifier, Journal of Microscopy. 261 (2016) 46–56.

[14] D.H. Theriault, M.L. Walker, J.Y. Wong, M. Betke, Cell morphology classification and clutter mitigation in phase-contrast microscopy images using machine learning, Machine Vision and Applications. 23 (2012) 659–673.

[15] T.R. Jones, A.E. Carpenter, M.R. Lamprecht, J. Moffat, S.J. Silver, J.K. Grenier,

A.B. Castoreno, U.S. Eggert, D.E. Root, P. Golland, et al., Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning, Proceedings of the National Academy of Sciences of the United States of America. 106 (2009) 1826–31.

[16] A. Shariff, J. Kangas, L.P. Coelho, S. Quinn, R.F. Murphy, Automated image analysis for high-content screening and analysis, Journal of Biomolecular Screening. 15 (2010) 726–734.

[17] A. Křížova, J. Čolláková, Z. Dostál, L. Kvasnica, H. Uhlířová, T. Zikmund, P. Veselý, R. Chmelík, Dynamic phase differences based on quantitative phase imaging for the objective evaluation of cell behavior, Journal of Biomedical Optics. 20 (2015) 111214.

[18] J. Čolláková, A. Křížová, V. Kollárová, Z. Dostál, M. Slabá, P. Veselý, R. Chmelík, Coherence-controlled holographic microscopy enabled recognition of necrosis as the mechanism of cancer cells death after exposure to cytopathic turbid emulsion, Journal of Biomedical Optics. 20 (2015) 111213.

[19] P. Marquet, B. Rappaz, P.J. Magistretti, E. Cuche, Y. Emery, T. Colomb, C. Depeursinge, Digital holographic microscopy: a non-invasive contrast imaging technique allowing quantitative visualization of living cells with subwavelength axial accuracy, Optics Letters. 30 (2005) 468.

[20] B. Kemper, A. Bauwens, A. Vollmer, S. Ketelhut, P. Langehanenberg, J. Müthing, H. Karch, G. von Bally, Label-free quantitative cell division monitoring of endothelial cells by digital holographic microscopy, Journal of Biomedical Optics. 15 (2010) 36009.

[21] P. Girshovitz, N.T. Shaked, Generalized cell morphological parameters based on interferometric phase microscopy and their application to cell life cycle characterization, Biomedical Optics Express. 3 (2012) 1757.

[22] D. Bettenworth, P. Lenz, P. Krausewitz, M. Brückner, S. Ketelhut, D. Domagk, B. Kemper, Quantitative stain-free and continuous multimodal monitoring of wound healing in vitro with digital holographic microscopy, PLoS One. 9 (2014) e107317.

[23] M. Mir, A. Bergamaschi, B.S. Katzenellenbogen, G. Popescu, Highly sensitive quantitative imaging for monitoring single cancer cell growth kinetics and drug response., PLoS One. 9 (2014) e89000.

[24] F. Yi, I. Moon, B. Javidi, Cell morphology-based classification of red blood cells using holographic imaging informatics, Biomedical Optics Express. 7 (2016) 2385.

[25] A. El Mallahi, C. Minetti, F. Dubois, Automated three-dimensional detection and classification of living organisms using digital holographic microscopy with partial spatial coherent source: application to the monitoring of drinking water resources, Applied Optics. 52 (2013) A68.

[26] H. Majeed, M.E. Kandel, K. Han, Z. Luo, V. Macias, K. Tangella, A. Balla, G.

Popescu, Breast cancer diagnosis using spatial light interference microscopy., Journal of Biomedical Optics. 20 (2015) 111210.

[27] T.H. Nguyen, S. Sridharan, V. Macias, A.K. Balla, M.N. Do, G. Popescu, Prostate cancer diagnosis using quantitative phase imaging and machine learning algorithms, in: G. Popescu, Y. Park (Eds.), SPIE BiOS, International Society for Optics and Photonics, 2015: p. 933619.

[28] L. Štrbková, A. Manakhov, L. Zajíčková, A. Stoica, P. Veselý, R. Chmelík, The adhesion of normal human dermal fibroblasts to the cyclopropylamine plasma polymers studied by holographic microscopy, Surface and Coatings Technology. 295 (2015) 70–77.

[29] V. Kollárová, J. Čolláková, Z. Dostál, P. Veselý, R. Chmelík, Quantitative phase imaging through scattering media by means of coherence-controlled holographic microscope, Journal of Biomedical Optics. 20 (2015) 111206.

[30] J. Babocký, A. Křížová, L. Štrbková, L. Kejík, F. Ligmajer, M. Hrtoň, P. Dvořák, M. Týč, J. Čolláková, V. Křápek, et al., Quantitative 3D phase imaging of plasmonic metasurfaces, ACS Photonics. 4 (2017) 1389–1397.

[31] T. Kreis, Digital holographic interference-phase measurement using the Fourier-transform method, Journal of the Optical Society of America A. 3 (1986) 847–855.

[32] D. Ghiglia, M. Pritt, Two-dimensional phase unwrapping: theory, algorithms, and software, John Wiley & Sons, New York, 1998.

[33] R. Goldstein, H. Zebker, C. Werner, Satellite radar interferometry: Two-dimensional phase unwrapping, Radio Science. 23 (1988) 713–720.

[34] T. Zikmund, L. Kvasnica, M. Týč, A. Křížová, J. Čolláková, R. Chmelík, Sequential processing of quantitative phase images for the study of cell behaviour in real-time digital holographic microscopy, Journal of Microscopy. 256 (2014) 117–25.

[35] Z. El-Schich, A. Mölder, H. Tassidis, P. Härkönen, M. Falck Miniotis, A. Gjörloff Wingren, Induction of morphological changes in death-induced cancer cells monitored by holographic microscopy, Journal of Structural Biology. 189 (2015) 207–212.

[36] J. Kühn, F. Charrière, T. Colomb, E. Cuche, F. Montfort, Y. Emery, P. Marquet, C. Depeursinge, Axial sub-nanometer accuracy in digital holographic microscopy, Measurement Science and Technology. 19 (2008) 74007.

[37] R. Barer, Interference microscopy and mass determination, Nature. 169 (1952) 366–367.

[38] T.A. Zangle, M.A. Teitell, Live-cell mass profiling: an emerging approach in quantitative biophysics, Nature Methods. 11 (2014) 1221–1228.

[39] S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised machine learning: A review of classification techniques, in: Emerg. Artif. Intell. Appl. Comput. Eng., IOS Press, Washington, DC, 2007: p. 407.

[40] A.K. Jain, M.N. Murty, P.J. Flynn, A. Rosenfeld, K. Bowyer, N. Ahuja, A. Jain, Data clustering: a review, ACM Computing Surveys. 31 (1999).

[41] M. Sonka, V. Hlavac, R. Boyle, Image processing, analysis, and machine vision, Cengage Learning, 2014.

[42] K. Parvati, P. Rao, M.M. Das, Image segmentation using gray-scale morphology and marker-controlled watershed transformation, Discrete Dynamics in Nature and Society. (2009).

[43] R. Duda, P. Hart, D. Stork, Pattern classification, John Wiley & Sons, 2012.

[44] R. Johnson, D. Wichern, Applied multivariate statistical analysis, Prentice-Hall, New Jersey, 2014.

[45] A. Ghasemi, S. Zahediasl, Normality tests for statistical analysis: a guide for non-statisticians, International Journal of Endocrinology and Metabolism. 10 (2012) 486–489.

[46] A. Elliott, W. Woodward, Statistical analysis quick reference guidebook: With SPSS examples, SAGE Publications, 2007.

[47] N. Razali, Y. Wah, Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests, Journal of statistical modelling and analytics. 2 (2011) 21–33.

[48] F.E. Satterthwaite, An approximate distribution of estimates of variance components, Biometrics Bulletin. 2 (1946) 110.

[49] D.H. Wolpert, The lack of a priori distinctions between learning algorithms, Neural Computation. 8 (1996) 1341–1390.

[50] J.R. Quinlan, Induction of decision trees, Machine Learning. 1 (1986) 81–106.

[51] G.J. McLachlan, Discriminant analysis and statistical pattern recognition, Wiley-Interscience, 2004.

[52] C. Cortes, V. Vapnik, Support vector machine, Machine Learning. 20 (1995) 273–297.

[53] P. Cunningham, S.J. Delany, K-nearest neighbour classifiers, Technical Report UCD-CSI-2007-4, 2007: 1–17.

[54] T.G. Dietterich, Ensemble methods in machine learning, in: J. Kittler, F. Roli (Eds.), Multiple classifier systems, Springer, 2001: pp. 1–15.

[55] G.P. Zhang, Neural networks for classification: a survey, IEEE Transactions on Systems, Man, and Cybernetics, Part C Applications Rev. 30 (2000) 451–462.

[56] M. Makhtar, D. Neagu, M. Ridley, Comparing multi-class classifiers: on the similarity of confusion matrices for predictive toxicology applications, in: Proceedings of Intelligent Data Engineering and Automated Learning-IDEAL, 2011: pp. 252–261.

[57] V. Piuri, F. Scotti, Morphological classification of blood leucocytes by microscope images, in: Proceedings of Computational Intelligence for Measurement Systems and Applications, 2004: pp. 103–108.

[58] M. Hollander, D.A. Wolfe, E. Chicken, Nonparametric statistical methods., John Wiley & Sons, New Jersey, 2013.

[59] R. Kalluri, R.A. Weinberg, The basics of epithelial-mesenchymal transition, Journal of Clinical Investigation. 119 (2009) 1420–8.

[60] J.P. Thiery, Epithelial–mesenchymal transitions in tumour progression, Nature Reviews Cancer. 2 (2002) 442–454.

[61] J. Xu, S. Lamouille, R. Derynck, TGF-beta-induced epithelial to mesenchymal transition., Cell Research. 19 (2009) 156–72.

[62] A. Oppenheim, Discrete-time signal processing, Pearson Education India, 1999.

[63] R. Young, Wavelet theory and its applications, Springer Science & Business Media, New York, 2012.

[64] J. Lin, E. Keogh, S. Lonardi, B. Chiu, A symbolic representation of time series, with implications for streaming algorithms, in: Proceedings of 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery, ACM Press, 2003: pp. 2-11.

[65] E.J. Keogh, M.J. Pazzani, Scaling up dynamic time warping for datamining applications, in: Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining, ACM Press, 2000: pp. 285–289.

[66] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdisciplinary Reviews: Computational Statistics. 2 (2010) 433–459.

[67] Y. Saeys, I. Inza, P. Larranaga, A review of feature selection techniques in bioinformatics, Bioinformatics. 23 (2007) 2507–2517.

[68] P. Bouchal, L. Štrbková, Z. Dostál, Z. Bouchal, Vortex topographic microscopy for full-field reference-free imaging and testing, Optics Express. 25 (2017) 21428.

[69] J. Balvan, A. Křížová, J. Gumulec, M. Raudenská, Z. Sládek, M. Sedláčková, P. Babula, M. Sztalmachová, R. Kizek, R. Chmelík, et al., Multimodal holographic microscopy: distinction between apoptosis and oncosis, PLoS One. 10 (2015) e0121674.

[70] D. Roitshtain, L. Wolbromsky, E. Bal, H. Greenspan, L.L. Satterwhite, N.T. Shaked, Quantitative phase microscopy spatial signatures of cancer cells, Cytometry Part A. 91 (2017) 482–493.

[71] V.L. Calin, M. Mihailescu, E.I. Scarlat, A. V. Baluta, D. Calin, E. Kovacs, T. Savopol, M.G. Moisescu, Evaluation of the metastatic potential of malignant cells by image processing of digital holographic microscopy data, FEBS Open Bio. (2017) 1-12.

[72] K. Lee, K. Kim, J. Jung, J. Heo, S. Cho, S. Lee, G. Chang, Y. Jo, H. Park, Y. Park, Quantitative phase imaging techniques for the study of cell pathophysiology: from principles to applications, Sensors. 13 (2013) 4170–91.

# 12. Author Publications and Other Outputs

**PUBLICATIONS:**

L. Štrbková, A. Manakhov, L. Zajíčková, A. Stoica, P. Veselý, R. Chmelík. The adhesion of normal human dermal fibroblasts to the cyclopropylamine plasma polymers studied by holographic microscopy. Surface and Coatings Technology, Vol. 295 (2016), pp. 70-77 (**Q1, IF = 2.589**)

M. Antoš, P. Bouchal, Z. Dostál, L. Štrbková, L. Kvasnica, P. Kolman, R. Chmelík. Mikroskopie v Laboratoři experimentální biofotoniky. Jemná mechanika a optika, 2016, Vol. 61, No. 6, pp. 135-139. ISSN: 0447-6441.

J. Babocký, A. Krížová, L. Štrbková, L. Kejík, F. Ligmajer, M. Hrtoň, P. Dvořák, M. Týč, J. Čolláková, V. Křápek, R. Kalousek, R. Chmelík, T. Šikola. Quantitative 3D phase imaging of plasmonic metasurfaces. ACS Photonics, 2017, Vol. 4, No. 6, p. 1389-1397. ISSN: 2330-4022 (**Q1, IF = 6.756**)

L. Štrbková, D. Zicha, P. Veselý, R. Chmelík. Automated classification of cell morphology by coherence-controlled holographic microscopy. Journal of Biomedical Optics, 2017, Vol. 22, No. 8. ISSN: 1083-3668 **(Q2, IF = 2.53)**

P. Bouchal, L. Štrbková, Z. Dostál, Z. Bouchal. Vortex topographic microscopy for full-field reference-free imaging and testing. Optics Express, 2017, Vol. 25, No. 18. ISSN: 1094-4087 **(Q1, IF = 3.307)**

**OTHER OUTPUTS:**

L. Štrbková, L. Methodics of Characterisation for the Cold-Field Emission Sources Intended for Electron Microscopy. In *Proceedings of the 19th conference Student EEICT*, Vol. 2, 2013, Brno, Czech Republic. (awarded by 2nd place)

L. Štrbková, A. Krížová, J. Čolláková, P. Veselý, R. Chmelík. Dynamic phase differences method for the assessment of cellular dynamic processes. In *Proceedings of the Microscience Microscopy Congress,* 2014, Manchester, UK. ISBN 978-80-210-7159-9

J. Čolláková, A. Krížová, Z. Dostál, L. Štrbková, M. Lošťák, L. Kvasnica, T. Slabý, P. Kolman, M. Antoš, P. Veselý, R. Chmelík. Cell biology by Coherence Controlled Holographic Microscope (CCHM). In *Proceedings of the 18th International Microscopy Congress,* 2014, Prague, Czech Republic. ISBN 978-80-260-6720-7

V. Kollárová, M. Lošťák, M. Slabá, T. Slabý, J. Čolláková, Z. Dostál, A. Krížová, L. Štrbková, R. Chmelík. Imaging of 2D objects in diffuse media by coherence-controlled holographic microscope. In *Digital Holography and Three-Dimensional Imaging*, 2014 Seattle, Washington, United States.

L. Štrbková, A. Manakhov, A. Stoica, L. Zajíčková, P. Veselý, R. Chmelík. Biocompatibility Assessment of Cyclopropylamine Plasma Polymers Studied by

Coherence-Controlled Holographic Microscopy. In *Proceedings of Focus on Microscopy,* 2015, Göttingen, Germany. ISBN 978-3-95404-942-4

L. Štrbková, A. Manakhov, A. Stoica, L. Zajíčková, P. Veselý, R. Chmelík. Biocompatibility Assessment of Cyclopropylamine Plasma Polymers Studied by Q-Phase. In *Proceedings of Frontiers in Material and Life Science, Creating Life in 3D,* 2014, Brno, Czech Republic.

L. Štrbková, A. Manakhov, Veselý, R. Chmelík. Biocompatibility of Thin Films Studied by Q-Phase. In *Proceedings of 3D Image Acquisition and Display: Technology, Perception and Applications,* 2016, Heidelberg, Germany. ISBN: 978-1-943580-15-6.

L. Štrbková, L.; P. Veselý; R. Chmelik. The role of digital holographic microscopy in the classification of cellular morphologies. In *Proceedings of Focus on Microscopy,* 2017, Bordeaux, France.

J. Babocký, A. Krížová, L. Štrbková, L. Kejík, F. Ligmajer, M. Hrtoň, P. Dvořák, M. Týč, J. Čolláková, V. Křápek, R. Kalousek, R. Chmelík, T. Šikola. Quantitative 3D phase imaging of plasmonic metasurfaces. In *Proceedings of the 8th International Conference on Surface Plasmon Photonics*, 2017, Taipei, Taiwan.

B. Diederich, L. Štrbková, F. Mucha, B. Cao, J. Peychl, R. Heintzmann. Machine Learning to Reconstruct 3D Scattering Data from Partially Coherent Imaging Data. In *Proceedings of the Quantitative BioImaging Conference*, 2018, Göttingen, Germany. **submitted conference abstract**

**PROJECTS**:

- Junior research grant 2015 "Automatic detection of cell apoptosis", FSI/STI-J-15-2752 - **principal investigator**
- BUT Molecular biotechnology (2015-2016), CZ.1.05/3.1.00/14.0311 - research team member
- Junior research grant 2016 "Recognition of dynamic cellular processes in quantitative phase images", STI-J-16-3796 - **principal investigator**

**FOREIGN INTERNSHIP:**

- Doctoral internship (1.10. – 1.11.2014) at **University of Illinois at Urbana-Champaign,** Beckman Institute for Advanced Science and Technology, 405 N Mathews Ave, Urbana, IL 6180, USA (Placement: Quantitative Light Imaging Laboratory)
- Doctoral internship (5.8. – 5.10.2016) **Max Planck Institute of Molecular Cell Biology and Genetics**, Pfotenhauerstraße 108, 01307 Dresden, Germany (Placement: Light Microscopy Facility)