VĚDECKÉ SPISY VYSOKÉHO UČENÍ TECHNICKÉHO V BRNĚ

Edice PhD Thesis, sv. 856

ISSN 1213-4198

Ing. Lenka Štrbková

Biophysical Interpretation of Quantitative Phase Image

VYSOKÉ UČENÍ
TECHNICKÉ
V BRNĚ

STŘEDOEVROPSKÝ TECHNOLOGICKÝ INSTITUT

CENTRAL EUROPEAN INSTITUTE OF TECHNOLOGY

BIOPHYSICAL INTERPRETATION OF QUANTITATIVE PHASE IMAGE

BIOFYZIKÁLNÍ INTERPRETACE KVANTITATIVNÍHO FÁZOVÉHO ZOBRAZENÍ

ZKRÁCENÁ VERZE PH.D. THESIS

OBOR	Pokročilé nanotechnologie a mikrotechnologie
AUTOR PRÁCE	Ing. Lenka Štrbková
VEDOUCÍ PRÁCE	prof. RNDr. Radim CHmelík, Ph.D.
OPONENTI	prof. RNDr. Michal Kozubek, Ph.D. Dr. Andreas Hoppe
DATUM OBHAJOBY	13. března 2018

Brno 2018

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

Místo uložení práce:
Thesis is placed at the CEITEC Brno University of Technology, Purkyňova 123 612 00 Brno, Czech Republic.

ŠTRBKOVÁ, L. *Biophysical Interpretation of Quantitative Phase Image.* Brno: Brno University of Technology, Central European Institute of Technology, 2017. 31 p. Thesis supervisor prof. RNDr. Radim Chmelík, Ph.D.

ISBN 978-80-214-5633-4
ISSN 1213-4198

Table of Contents

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 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, and 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. The other approach to machine learning is unsupervised learning. Here, the computer system does not rely on the prior knowledge. Instead, the system finds new patterns and subdivides the data by using a set of pre-defined general rules. This work focuses solely on the supervised machine learning.

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]. 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.

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. Rajpoot wrote paper [10] about hyperspectral colon tissue cell classification based on supervised support vector machines (SVM). Machine learning approach for classification of erythrocytes was presented by Das et al. [11]. In those 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 physilogical changes while being fixed, which could possibly affect the measurement.

Another study [12] presents application of machine learning to analysis of cell morphology in phase-contrast microscopy. 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. Several publications have focused on classification of cells in the images gained by fluorescent microscopy. Automated scoring of diverse cell morphologies was described in [13]. Several automated image analysis methods for high-content screening were summarised in [14]. 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 [15–17], 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 [18–20]. However, 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 [21]. Automated detection and classification of living organisms in drinking water resources using DHM was performed in [22]. But to my present knowledge, none of the publications studied the potential of QPI for the classification of live adherent eukaryotic cells.

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 is a follow-up to existing results reached in the research area. 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. 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 [15,23,24], but also for technical specimens [25]. The main asset of this technique is the ability to provide quantitative phase image [4].

The optical set-up of the microscope is based on Mach-Zehnder-type interferometer modified for achromatic off-axis holographic microscopy. Microscope consists of two separated nearly identical optical arms – reference and object arm. 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 numerical reconstruction of the hologram is performed using the house-built software using the 2D Fourier transform [26] and phase unwrapping algorithm [27,28].

4.1 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 φ_o is phase in the object arm, φ_r is phase in the reference arm, λ 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 Δn is the difference between the refractive indices of the cellular material and the medium.

The phase can be also interpreted in terms of cell dry mass. The value of the cell dry mass is dependent mainly on the protein concentration within the cell [29]. It has been shown that the refractive index of the cellular material is directly proportional to the dry mass of the cell with the constant γ 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 $g.ml^{-1}$). It has been published years ago that the measured phase corresponds to the dry mass distribution within the cells [6,7,29,30]. The dry mass density of the cell (units of $pg.\mu m^{-2}$) can be obtained from the measured phase as follows:

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

5. Machine Learning in QPI

5.1 Introduction to Machine Learning

Machine learning explores the algorithms that have the ability to automatically learn and improve from the data and subsequently make predictions on the unknown data [1]. Machine learning algorithms are often categorized as supervised or unsupervised. In supervised machine learning [31], the algorithm relies on the prior knowledge and is trained from labelled training data. 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. Unsupervised learning [32] does not rely on any prior knowledge. Only supervised machine learning will be employed in this work. Supervised learning problems can be divided into regression and classification. This work focuses solely on the classification.

5.2 Classification in QPI

The approach for automated classification of cells in the quantitative phase images obtained by CCHM is proposed here. 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 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 and serve as an input for the classification. After the training of the classification algorithm, the testing unlabelled data can be fed into the classifier. The overview of the proposed classification process based on QPI is shown in Figure 1. In the following chapters, the steps of the classification process will be introduced in detail.

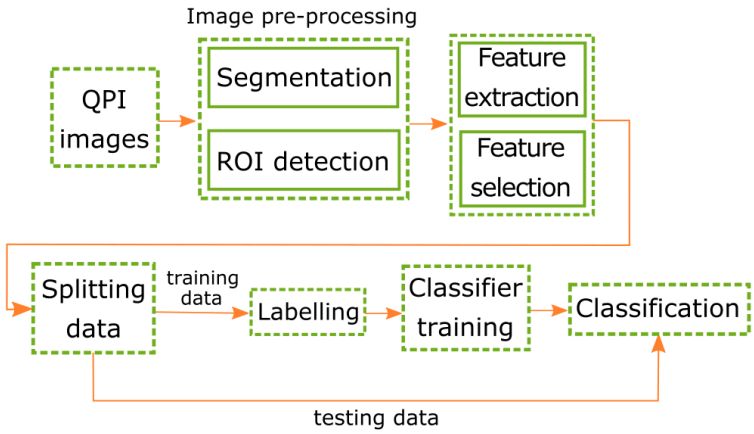

Figure 1: Overview of the proposed classification process based on QPI.

5.2.1 Image Pre-processing

The cells in the quantitative phase image are firstly segmented from the background. Several methods for the segmentation exist [33], in this work the marker-controlled watershed segmentation approach [34], implemented in Q-Phase software (TESCAN ORSAY HOLDING a.s.), is applied. The segmented cells are then identified as separate ROIs, while each of them is labelled by a unique integer number.

5.2.2 Feature Extraction

After the image pre-processing, each ROI is represented by a set of cell features. The process, in which the input data are transformed into a reduced representation by feature vector, is often termed feature extraction [35]. 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.

(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.

(ix) Extent (EX) is given by the ratio of pixels in the cell region to pixels in the total bounding box. 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 contain quantitative information about dry mass density distribution within the cell.

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

$$\varphi_{total} = \sum_{i=1}^{k} \varphi_i, \tag{6}$$

where *k* is the number of pixels of the cell region and φ_i is the phase value in the *i*th pixel belonging to the region of the cell.

(ii) Average phase (μ_φ) 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_φ) and standard deviation of the phase (σ_φ) 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:

$$Var_\varphi = \frac{1}{k-1}\sum_{i=1}^{k}(\varphi_i - \mu_\varphi)^2, \tag{7}$$

and

$$\sigma_\varphi = \sqrt{Var_\varphi}. \tag{8}$$

(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:

$$Skew_\varphi = \frac{\sum_{i=1}^{k}(\varphi_i - \varphi_{avg})^3}{(k-1)\sigma_\varphi^3}. \quad (9)$$

(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

$$Kurt_\varphi = \frac{\sum_{i=1}^{k}(\varphi_i - \varphi_{avg})^4}{(k-1)\sigma_\varphi^4}. \quad (10)$$

All extracted features are summarized into feature vectors, each representing one cell. Each cell feature vector is then assigned one of the class labels. Additional step before the classification is the feature scaling. The main advantage of scaling is to avoid features in greater numeric ranges dominating those in smaller numeric ranges [35].

5.2.3 Feature Selection

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 [36] was chosen for that purpose. 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. Since the parametric *t*-test is based on the assumption that the data follows normal distribution [37], the normality was verified by box-whisker plots and Shapiro-Wilk test [38].

5.2.4 Supervised Classification Algorithms

It is well known that the performance of the classification is highly dependent on the selection of the classification algorithm [39] and thus we employ several supervised machine learning algorithms in this work to correctly assess the performance of the classification. 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 [40], 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 [41] 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 [42] classifies data by finding the best discriminating hyperplane that separates objects with different class membership. 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 closest data points to the margin of separation are called support vectors. The support vectors thus specify the discrimination function. SVM can handle both linearly

separable data and non-linearly separable data using kernel functions. Here we used linear, quadratic, cubic and Gaussian kernel.

(d) K-nearest neighbour (KNN) classifier. The principle behind *k* nearest neighbour method [43] 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).

(e) Ensemble classifiers. Ensemble methods [44] combine multiple learning algorithms in order to improve generalization and robustness over a single learning algorithm. 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 [45] was inspired by the human learning process and is based on combinations of neurons, each of which takes a number of inputs and generates an output. Each neuron is associated with adaptive weight that is tuned by a learning algorithm. The output of the neuron is a function of the weighted sum of inputs. Neuron is associated with activation function, which defines the output of that neuron given an input or set of inputs. A 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. 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.

5.2.5 Classifier Performance Evaluation

For indication of the performance of a classification algorithm, confusion matrices are a widely used tool [46]. The confusion matrix compares training class labels with output class labels determined by the classification algorithm. Several performance parameters can be calculated from the confusion matrix for a classification algorithm: accuracy, precision, recall and F-score. *K*-fold cross-validation was used to evaluate the performance of the classification. The data were partitioned into *k* randomly chosen subsets of roughly equal size. One subset was used to validate the classifier, which had been trained on the remaining subsets. This process was repeated *k* times, such that each subset was used for the validation (we used $k=5$).

6. 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. Here, the proposed methodology exploiting the quantitative information about the dry mass density distribution within the cell is applied in the experiment.

6.1 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 approach is compared with a commonly used methods based on morphometric features.

6.2 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. 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.

6.3 Image Acquisition

The 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. At least 100 images were acquired from each sample in pursuit of collecting enough data for the classification.

Morphological changes of cells appeared in the order of minutes after the application of PBS. 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 (Figure 2). 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. 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).

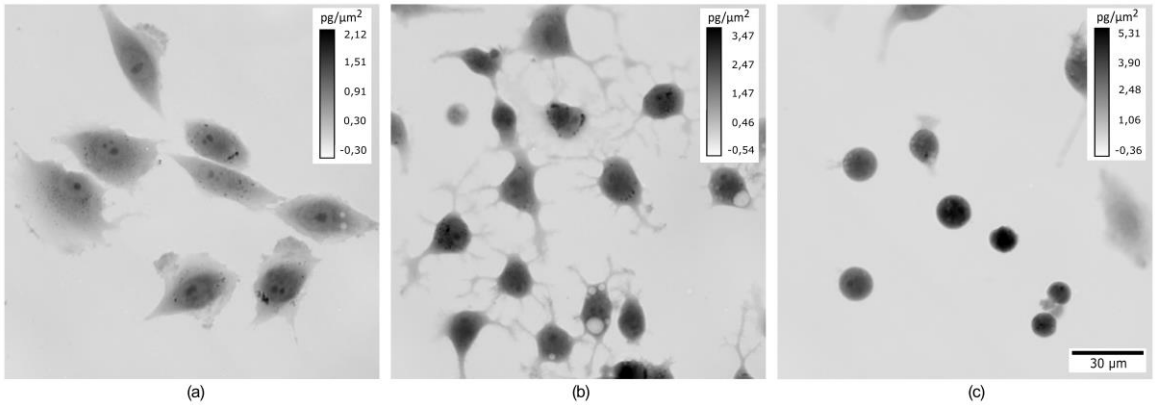

Figure 2: Morphological changes of LW13K2 cells induced by PBS. (a) Viable cells, (b) semi-deprived cells and (c) deprived cells. Quantitative phase images are shown in grayscale in units of pg/µm^2 recalculated from phase (in radians) according to Davies [7].

6.4 Image Pre-processing and Feature Extraction

The cells in the quantitative phase images were segmented from the background by marker-controlled watershed segmentation approach. 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. The image processing was performed in Matlab.

6.5 Feature Selection

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). Shapiro-Wilk test was performed to confirm that the values of cell features are normally distributed. For visual confirmation, box-whisker plots of the feature value distributions are shown in Figure 3.

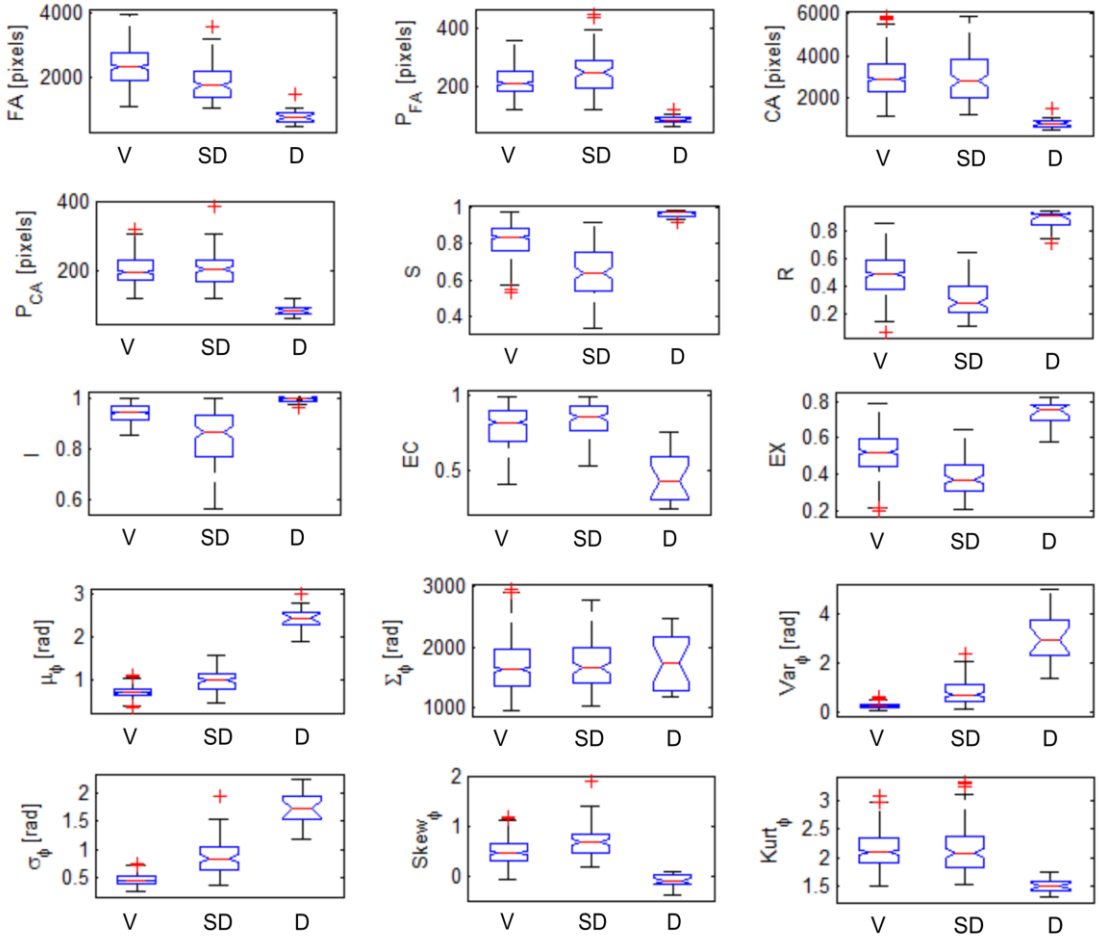

Figure 3: Box-whisker plots of the values of features (from the top to the right): footprint area (*FA*), perimeter of the footprint area (P_{FA}), convex area (*CA*), perimeter of the convex area (P_{CA}), solidity (*S*), roundness (*R*), indentation (*I*), eccentricity (*EC*), extent (*EX*), average phase value (μ_φ), total phase value (φ_{total}), variance of the phase (Var_φ), standard deviation of the phase (σ_φ), skewness ($Skew_\varphi$) and kurtosis ($Kurt_\varphi$).

Afterwards, the *t*-test was performed for each feature and between all possible pairs in order to investigate whether the defined features are reliable for distinguishing between the cell classes. The results show that nearly all features except φ_{total} have potential for the

discrimination between the 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 3). It should be noted that QPI cell features derived from the phase (except for φ_{total}) showed much lower *p*-value (mostly one order of magnitude lower) in the *t*-test than morphometric cell features. This leads to assumption, that QPI features could enhance classification performance.

6.6 Classification Results

Feature vectors representing the cells in the quantitative phase images form an input for the classification algorithms. Prior to classification, the feature vectors are filtered based on the feature selection results. Therefore, the feature φ_{total} is eliminated from the feature vectors.

Several supervised machine learning algorithms were employed to compare the performance of the classification with two different sets of features. In the first case, the feature vector consisted of morphometric features only. In the latter case, also QPI features were added. Performance measures of each classification algorithm were determined as a mean of the values obtained by 5-fold cross-validation. 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 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 [11,12,47]. 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. For comparison of the two classification approaches, the performance results were evaluated by Wilcoxon signed rank test [48]. 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 4.

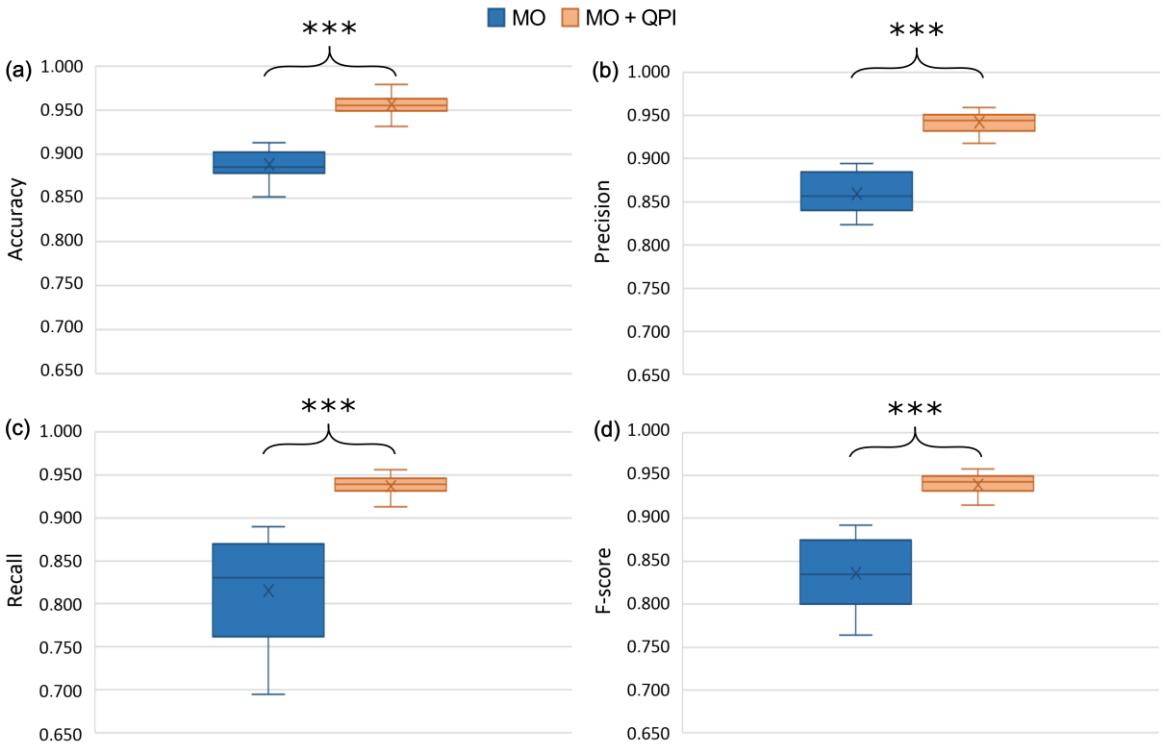

Figure 4: Box-whisker plots of overall classification performance for two types of feature vectors: (a) accuracy, (b) precision, (c) recall and (d) F-score. Wilcoxon signed rank test was used for the statistical analysis. Symbols indicating significance are placed above (***: $p < 0.001$).

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.

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 1 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 [48], which revealed no significant differences between the classification performance in the two experiments.

Table 1: 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. 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. The other option is to tune the parameters of the classification algorithms. Another options are the extraction of additional features or obtaining extra information from time-lapse QPI. Implementation of these two options will be objective of the following chapter.

7. 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. However, some complex dynamic processes demand time-resolved live-cell imaging in order to correctly interpret the cell states. 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. The proposed methodology will be demonstrated in the experiment with time-lapse images of live cells.

7.1 Experiment Design

The proposed approach was tested in the experiment with live adherent eukaryotic cells undergoing epithelial-mesenchymal transition (EMT) [49]. 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 5. 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. 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. The features with the highest potential to distinguish between the given classes are selected and form an input for the classifiers.

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, 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.

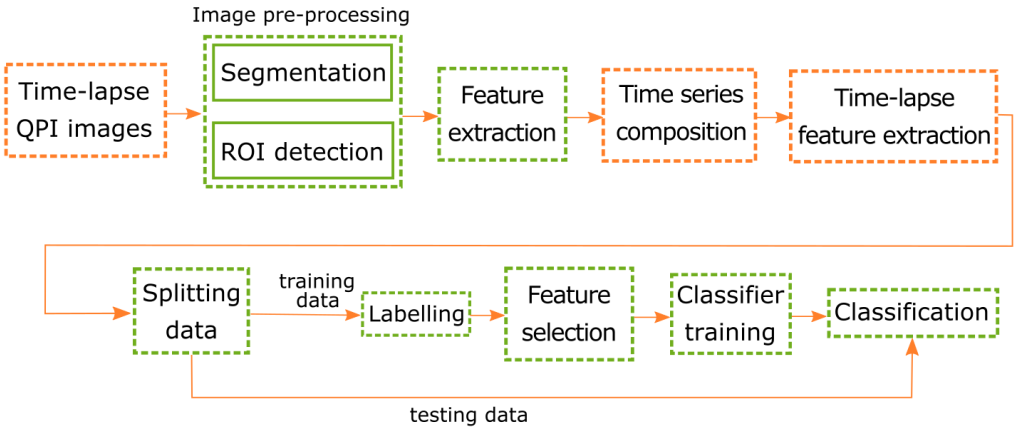

Figure 5: Overview of the proposed feature-based classification process based on time-lapse QPI.

7.2 Epithelial–Mesenchymal Transition

The epithelium is one or more layers of cells with different functions. Epithelial cells are closely adjacent and their arrangement ensures integrity, which does not allow cells to migrate. The mesenchymal cells are characterized by increased migration capacity, invasiveness and increased production of extracellular matrix components.

While being an essential process during development, EMT is also occurring under pathological conditions 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. Additional transformation by epigenetic changes and genetic modifications leads to a carcinoma in situ. 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 allowing them the transport to distant organs. 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) [50].

The transformation of epithelial cells into mesenchymal (Figure 6) is regulated by a sequence of strictly controlled molecular processes. At first, the cytoskeleton is reorganized and the cell shape changes. 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. 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 [51]. 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 its study.

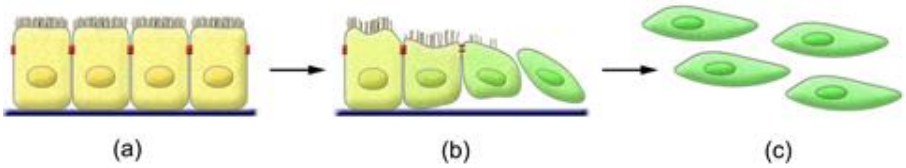

Figure 6: The steps of epithelial-mesenchymal transition (EMT). Polarized epithelial cells (a) lose their epithelial characteristics and reduce intercellular junctions and polarity (b). The cells acquire mesenchymal phenotype (c). The change is accompanied by degradation of the basal membrane [49].

7.3 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) 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 0,1mg/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. 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 5 ng/ml was added prior to the imaging.

7.4 Image Acquisition

The 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 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 represented the categories for 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 (Figure 7), the cells were labelled as either epithelial (48 cells) or mesenchymal (52 cells).

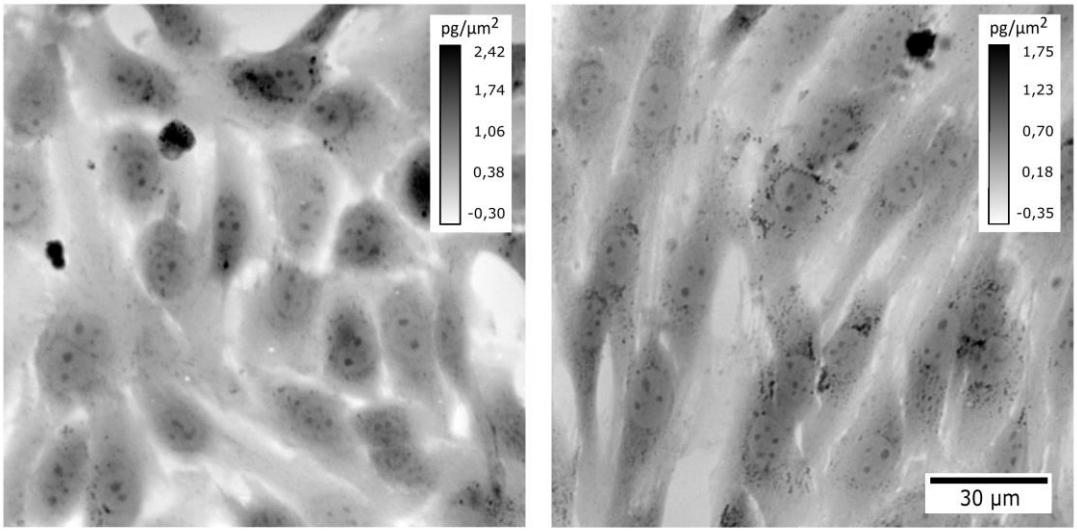

Figure 7: Examples of quantitative phase images of epithelial (a) and mesenchymal (b) phenotype.

7.5 Image Pre-processing and Feature Extraction

The cells in the time-lapse quantitative phase images were segmented from the background 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. We considered only cells staying in the field of view for the whole time. Subsequently, the cells were identified as separate ROIs. 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. All cell features therefore give rise to a multivariate time series. The example of the multivariate time series composed of 11 time-lapse images of one cell is shown in Figure 8.

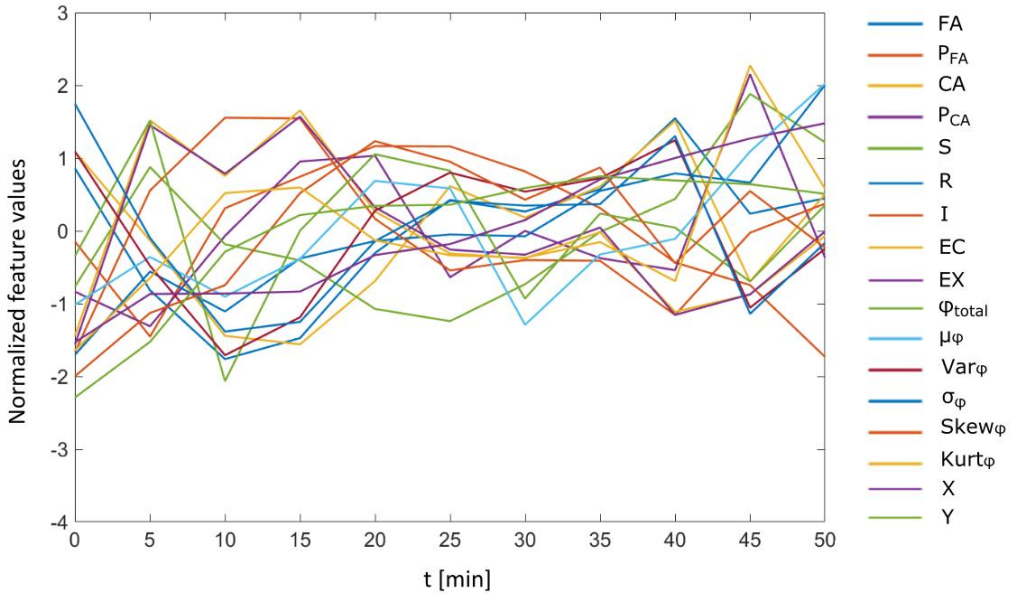

Figure 8: Example of the multivariate time series segment. Each univariate time series is composed of the feature values obtained within 50 min with 5 min interval. Footprint area (*FA*), perimeter of the footprint area (P_{FA}), convex area (*CA*), perimeter of the convex area (P_{CA}), solidity (*S*), roundness (*R*), indentation (*I*), eccentricity (*EC*), extent (*EX*), total phase of the cell (φ_{total}), average phase (μ_φ), variance (Var_φ) and standard deviation of the phase (σ_φ), skewness ($Skew_\varphi$), kurtosis ($Kurt_\varphi$), centroid X (*X*), centroid Y (*Y*).

7.5.1 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 $\mathbf{X} = \{X_1, X_2, ..., X_Q\}$ 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 9. The multivariate time series X_i can be written as

$$X_i = \{X_{ijt}\}, \qquad \text{for } j = 1,..., d;\ t = 1,..., n, \tag{11}$$

with the total number of observations dnQ.

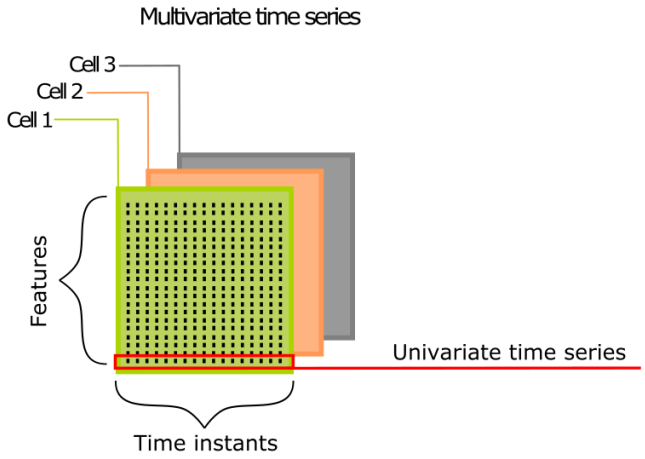

Figure 9: Illustrative demonstration of the multivariate time series representing cell behaviour. Each cell is represented by the multivariate time series composed of univariate time series (formed by cell feature values obtained within a defined time period).

We will consider the j-th component of the i-th time series $X_{ij} = \{X_{ij1}, ..., X_{ijn}\}$ 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 $M = (m_1, m_2, ..., m_L)$ 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.

There are several feature extraction techniques used for dealing with feature-based representation of the time series. The employed 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. The Fourier transform [52] therefore generates an approximation to a time series using as a basis cosine and sine functions. 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 for each frequency. However, many of the Fourier coefficients have 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.

(c) Wavelet transformation features

The wavelet transform [53] uses a basis containing waveforms that are localized in space and, therefore, is suitable for approximating time series including local structures. 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. 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. Algorithm computing discrete wavelet transform (DWT) was employed for the representation. The features extracted by the wavelet transform contain the approximation coefficients for each transform. 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.

(f) Symbolic aggregate approximation features

The symbolic aggregate approximation (SAX) method [54] 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) [55] and the conversion of a PAA sequence into a string composed of letters.

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 10.

Figure 10: 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 were added on the tail as shown in Figure 11. The principal components analysis and motion features were extracted in a different way, which will be described in the next sections.

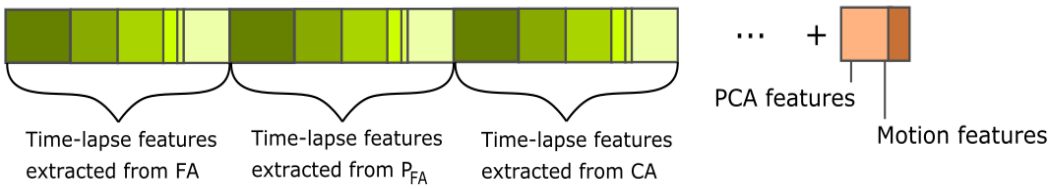

Figure 11: Final time-lapse feature vector construction. The final feature vector representing a single cell is formed by concatenation of partial time-lapse feature vectors obtained from univariate time series of QPI and morphological cell features. In addition, the motion and PCA features are added.

(g) Principal components analysis features

Principal components analysis (PCA) [56] is a statistical technique used to eliminate the less significant components 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. 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 and is calculated as

$$d_a = \sum_{i=1}^{n} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}\,, \tag{12}$$

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. *Velocity* of the cell motion is calculated as the overall distance travelled by the cell over the elapsed time. *Directionality* of the cell motion is calculated by comparing the *Euclidian distance* to the *accumulated distance.*

The motion features 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 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. The example of a set of final time-lapse feature vectors gained by feature-based approach can be seen in Figure 12.

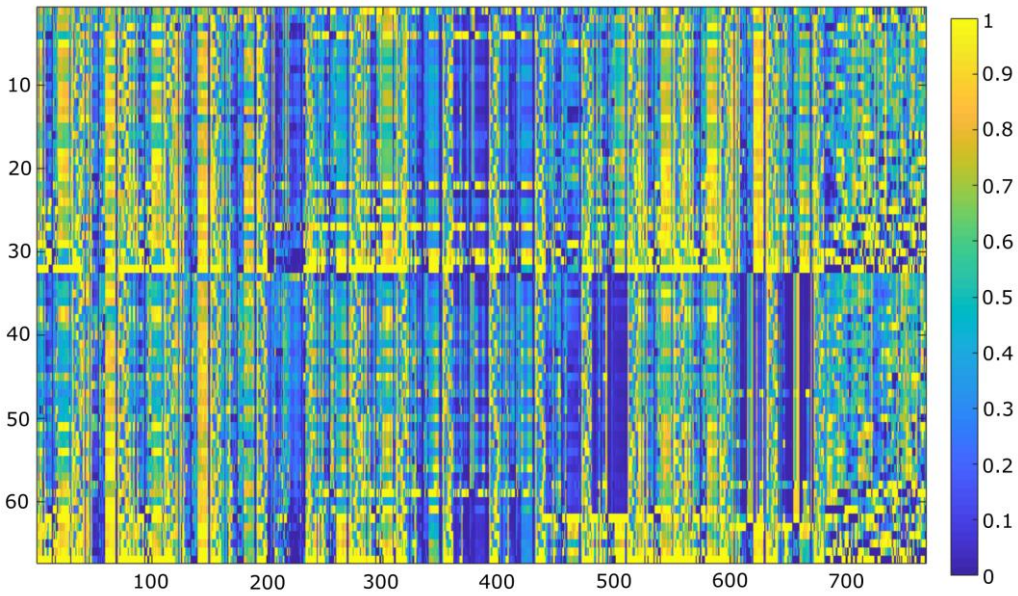

Figure 12: Example of the final time-lapse feature vectors concatenated into matrix. Elements of the matrix contain the (normalized) time-lapse feature values, and are visualized using colour: from blue (low values) to yellow (high values). First 32 rows represent time-lapse feature vectors extracted from epithelial cells and the other 35 rows from mesenchymal cells.

7.6 Feature Selection

Since the time-lapse feature vectors are composed of considerably higher number of features, the feature selection is performed in automated and more effective manner as in classification based on static QPI. Moreover, since the number of observations is small in comparison to large number of features, this may lead to overfitting of the learning algorithm to the noise. Reducing the number of features is therefore essential. Several methods exist for the feature selection, the filter approach was applied here. 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. There are approximately 15% of features in the feature-based approach, which have the *p*-values close to zero and 30% of features having the *p*-values smaller than 0.05. In the value-based approach, CDF of the *p*-values showed similar results. The features were subsequently ordered by their *p*-values. In order to define the appropriate number of features to be selected, the classification error as a function of the number of features was plotted. To obtain the classification error, several classification algorithms were employed. The result of the classification error in feature-based approach when using SVM is shown in Figure 13.

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 feature selection 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.

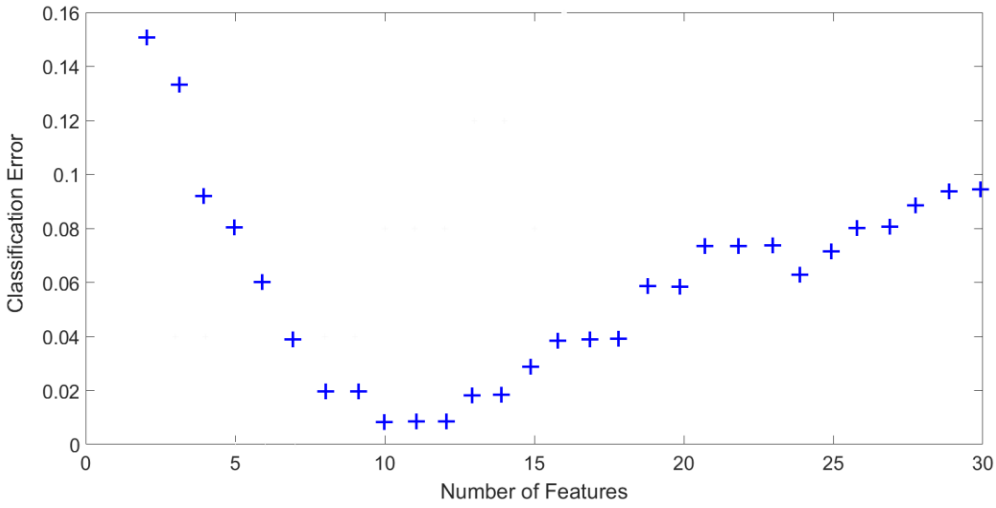

Figure 13: Classification error as a function of feature count (using SVM classifier) in feature-based approach.

7.7 Classification Results

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 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 performance of the classification implementing the value-based approach is summarized in Table 2. The overall accuracy of the classification was 0.924 ± 0.054.

Table 2: Performance of the classification by different algorithms using the value-based approach.

	Accuracy	Precision	Recall	F-score
MEAN ± SD	0.924 ± 0.054	0.908 ± 0.052	0.883 ± 0.089	0.894 ± 0.071

Performance of the classification using the feature-based approach is summarized in Table 3. Representing the cell behaviour by time-lapse features led to higher performance of the classifier as in the value-based approach, with the overall accuracy of the classification reaching 0.976 ± 0.011.

Table 3: Performance of the classification by different algorithms using the feature-based approach.

	Accuracy	Precision	Recall	F-score
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 5.2. The performance of the classification based on single-time-point QPI is summarized in Table 5. The overall accuracy of the classification was 0.890 ± 0.052.

Table 4: Performance of the classification by different algorithms using the static QPI.

	Accuracy	Precision	Recall	F-score
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 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 14.

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. 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.

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.

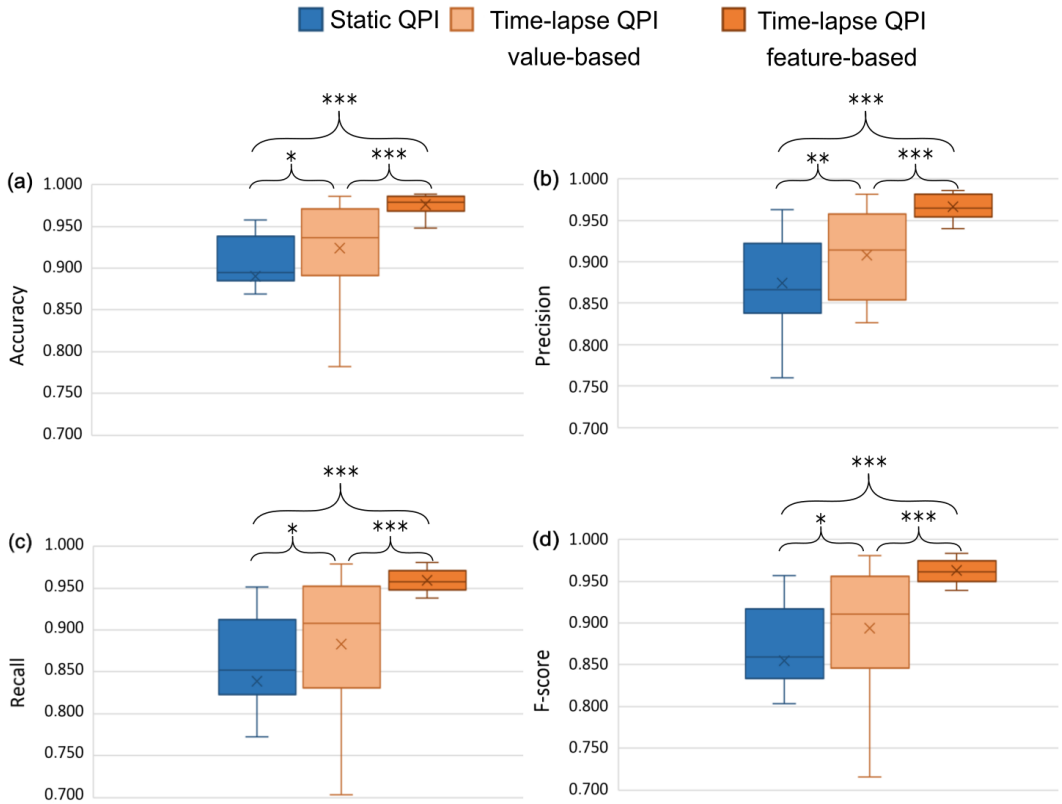

Figure 14: Box-whisker plots of overall classification performance based on static QPI, time-lapse QPI (value-based and feature-based approach): (a) accuracy, (b) precision, (c) recall and (d) *F*-score. Wilcoxon signed rank test was used for the statistical analysis (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).

8. 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 represent the first type. Quantitative phase images enable to extract features, which are based on phase distribution in the cell region and provide information about the dry mass density within the cell. 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 image pre-processing, both mentioned types of cell features were extracted and their potential to discriminate between the classes was assessed

both visually and statistically. 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.

In order to take into account also the dynamics of monitored cells, the methodology based on time-lapse quantitative phase images was proposed. 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.

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.

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 goal of the future work will be the implementation of these proposals.

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 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 [23], quantitative phase imaging of plasmonic metasurfaces [25] or vortex topographic microscopy for full-field reference-free imaging and testing. 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 10.

9. References

[1] C.M. Bishop, Pattern recognition and machine learning (Information Science and Statistics), 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) 14747–14762.

[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] K. Rajpoot, N. Rajpoot, SVM optimization for hyperspectral colon tissue cell classification, in: Barillot C., Haynor D.R., Hellier P. (Eds.), Medical Image Computing and Computer-Assisted Intervention, Berlin, Heidelberg, 2004: pp. 829–837.

[11] 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.

[12] 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.

[13] 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) 1826–1831.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

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

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

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

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

[31] S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised machine learning: A review of classification techniques, in: Emerging Artificial Intelligence Applications in Computer Engineering, IOS Press, Washington, DC, 2007: pp. 407.

[32] 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).

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

[34] 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).

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

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

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

[38] 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.

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

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

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

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

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

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

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

[46] 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.

[47] 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.

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

[49] R. Kalluri, R.A. Weinberg, The basics of epithelial-mesenchymal transition, Journal of Clinical Investigation. 119 (2009) 1420–1428.

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

[51] J. Xu, S. Lamouille, R. Derynck, TGF-beta-induced epithelial to mesenchymal transition, Cell Research. 19 (2009) 156–172.

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

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

[54] 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.

[55] 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.

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

10. 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, 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)

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.