





FULL ARTICLE

Machine learning of diffraction image patterns for accurate classification of cells modeled with different nuclear sizes

Jing Liu^{1,2}  | Yaohui Xu^{1,2}  | Wenjin Wang^{1,3} | Yuhua Wen^{1,3} |
Heng Hong⁴ | Jun Q. Lu^{1,5} | Peng Tian^{1,3*}  | Xin-Hua Hu^{1,5*} 

¹Institute for Advanced Optics, Hunan Institute of Science and Technology, Yueyang, Hunan, China

²School of Information Science and Engineering, Hunan Institute of Science and Technology, Yueyang, Hunan, China

³School of Physics & Electronic Science, Hunan Institute of Science and Technology, Yueyang, Hunan, China

⁴Department of Pathology and Comparative Medicine, Wake Forest School of Medicine, Wake Forest University, Winston-Salem, North Carolina

⁵Department of Physics, East Carolina University, Greenville, North Carolina

*Correspondence

Xin-Hua Hu, Department of Physics, East Carolina University, Greenville, NC 27858.
Email: hux@ecu.edu

Peng Tian, Institute for Advanced Optics, Hunan Institute of Science and Technology, Yueyang, Hunan 414006, China.
Email: tianpp815@163.com

Funding information

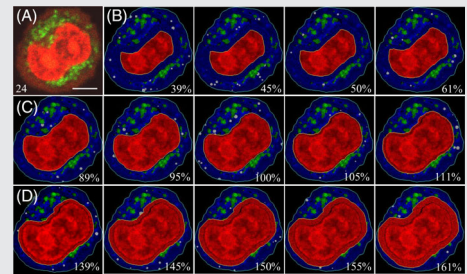
Education Department of Hunan Province, Grant/Award Number: 18B348; Hunan Provincial Science and Technology Department, Grant/Award Number: 19A198

Abstract

Measurement of nuclear-to-cytoplasm (N:C) ratios plays an important role in detection of atypical and tumor cells. Yet, current clinical methods rely heavily on immunofluorescent staining and manual reading. To achieve the goal of rapid and label-free cell classification, realistic optical cell models (OCMs) have been developed for simulation of diffraction imaging by single cells. A total of 1892 OCMs were obtained with varied nuclear volumes and orientations to calculate cross-polarized diffraction image (p-DI) pairs divided into three nuclear size groups of OCM_S, OCM_O and OCM_L based on three prostate cell structures. Binary classifications were conducted among the three groups with image parameters extracted by the algorithm of gray-level co-occurrence matrix. The averaged accuracy of support vector machine (SVM) classifier on test dataset of p-DI was found to be 98.8% and 97.5% respectively for binary classifications of OCM_S vs OCM_O and OCM_O vs OCM_L for the prostate cancer cell structure. The values remain about the same at 98.9% and 97.8% for the smaller prostate normal cell structures. The robust performance of SVM over clustering classifiers suggests that the high-order correlations of diffraction patterns are potentially useful for label-free detection of single cells with large N:C ratios.

KEYWORDS

cell modeling, cytology, diffraction imaging, light scattering



1 | INTRODUCTION

It is widely recognized that nuclear size change in terms of nuclear-to-cytoplasm (N:C) ratio offers a powerful cue for cancer diagnosis [1–5]. Histologic and immunofluorescent

microscopy remain the tools of choice to examine nuclear morphology in details. Despite recent advances in microscopy technology and machine learning, these imaging modalities still require staining, time-consuming acquisition and reading by trained specialists for decisive interpretation [5]. Various label-free methods of cellular imaging attract intensive research efforts for their practical

Jing Liu and Yaohui Xu contributed equally to this study.

benefits of little disturbance to imaged cells and much reduced preparation labors. Among these, imaging with coherent light scattered by cells stands out for strong signals and ability to profile internal structures by the 3D distribution of refractive index (RI) [6–12]. As we have discussed previously [13], 3D reconstruction of RI distribution consists of interferogram or diffraction image acquisition, error-prone phase unwrapping and computationally expensive tomographic reconstruction. Accomplishment of these steps is very challenging, which needs modeling nucleated cells with numerous and highly heterogeneous intracellular organelles of substantially irregular shapes. Discussion of contradictory results on the differences between RI values of cytoplasm and nucleus provided an illustrating example [9, 10, 14].

We have developed a single-shot method of polarization diffraction imaging flow cytometry (p-DIFC) for rapid assay of biological cells [15–19]. The p-DIFC method acquires one pair of cross-polarized diffraction images (p-DI) per cell by splitting the scattered light into s- and p-polarized components. Instead of RI reconstruction, the purpose of p-DI acquisition and analysis is for cell assay and classification by extraction of embedded diffraction pattern features. It has been shown experimentally that p-DI data contain rich information related to the cellular structure, and machine learning of embedded patterns allow accurate and rapid classification of cells in different prototypes [18, 20–25]. Extension of p-DIFC or other methods to the clinically important problems of nuclear size change detection, however, remains very difficult since comparisons have to be among cells of similar non-nuclear organelles. An effective and practical approach is to accurately simulate diffraction imaging of single cells with realistic optical cell models (OCMs) for calculations of p-DI pairs comparable to the measured ones. We have developed a set of tools to model single cells and simulate the diffraction imaging process [26, 27]. An OCM is first reconstructed from a stack of confocal images with large organelles of nucleus, mitochondria and cytoplasm membrane stained with fluorescent dyes [28–30]. Then the fluorescent intensities imaged from the stained organelles are used to derive RI values for subsequent simulations of light scattering and objective based imaging. With this method, we have investigated the dependence of diffraction patterns embedded in p-DI data on the 3D RI distributions of different prostate cells [13]. Compared to OCMs built by mixed spheres and ellipsoids [31–33], calculations of p-DI pairs with realistic OCMs provide an accurate means to quantitatively correlate the complex morphology of a cell and the characteristic diffraction patterns that can be quantified for classification.

In this report, we present a study on cell modeling and comparison of machine learning algorithms on classification of cells with different nuclear sizes using realistic OCMs built from confocal image stacks. A gray-level co-occurrence matrix (GLCM) algorithm has been selected to quantify textures of the calculated p-DIs [23, 24, 34]. We show that the supervised algorithm of support vector machine (SVM) performs well on binary cell classifications among three OCM groups of different nuclear sizes. In contrast, the unsupervised algorithms by clustering in the GLCM parameter space perform poorly that include hierarchical clustering, Gaussian mixture model (GMM) and k-means methods. Results of confusion matrix and scatter plot analysis are also presented to demonstrate the clear advantages of the SVM method for solving highly nonlinear and non-Gaussian type of classification problems in the GLCM parameter space.

2 | METHODS

2.1 | Confocal imaging and optical modeling of prostate cells

Confocal fluorescent image stacks were acquired from prostate cells of human prostate cancer cell line of PC3 (CRL-1435, ATCC) and human prostate normal epithelial cell type of PCS (PCS440010, ATCC). After the cells reached the logarithmic phase of growth in culture medium, they were detached from plate by trypsin and doubly stained by Syto-61 (S11343, ThermoFisher) for nucleus and MitoTracker Orange (M-7510, ThermoFisher) for mitochondria, which were chosen for their important roles on light scattering by cells [35]. Image stacks were acquired with a laser scanning confocal microscope (LSM 510, Zeiss) using a 63× water-immersion objective of 1.2 in NA and a 4× digital zoom. The pixel intensity of each fluorescence image slice was recorded in red and green channels, respectively, as F_r for Syto-61 and F_g for Mito-Tracker. Reconstruction was performed by an in-house developed code using the MATLAB platform (2019a, MathWorks) to output organelle type identifier and the fluorescence intensities of each voxel. The nuclear (or mitochondrial) voxels carry the F_r (or F_g) values only while the cytoplasm voxels hold both intensities due to small cytoplasmic concentrations of target molecules. We selected three cell structures for this study as PC3-a, PCS-a and PCS-b with the PC3 cell significantly larger than the PCS cells [22].

The OCMs for this study are based on fluorescence intensity stored in the 3D voxel array reconstructed from a confocal image stack to determine the intracellular RI distribution given by $n_\eta(\mathbf{r})$ for voxels of organelle type η at \mathbf{r} . We set the RI values by real numbers for all

intracellular organelles since the prostate cells absorb little light at the wavelength of $\lambda = 532$ nm for p-DIFC measurement. It was further assumed that the RI increments of an organelle from a baseline level of n_{c0} is dominated by water and linearly proportional to the dry mass or number of molecules targeted by fluorescent stains. The following equation has been employed because of the linear relation between fluorescent intensity F_r and/or F_g and dry mass of target molecules

$$n_\eta(\mathbf{r}) = n_{c0} + b_r F_r(\mathbf{r}) + b_g F_g(\mathbf{r}) \quad \forall \mathbf{r} \in \Omega_\eta \quad (1)$$

In Equation (1) we designate n_{c0} as the baseline RI of aqueous component in organelles such as cytosol in cytoplasm, b_r or b_g as the specific RI increment coefficients by the two fluorescent stain concentrations per voxel and Ω_η as the set of voxels with η as the organelle type index. One can easily show by Equation (1) that b_r is related to $n_{n,av}$ as the averaged RI value of nuclear voxels given by

$$b_r = \frac{n_{n,av} - n_{c0}}{F_{rn,av}}, \quad (2)$$

while b_g can be expressed in term of $n_{m,av}$ as the average RI value of mitochondrial voxels by

$$b_g = \frac{n_{m,av} - n_{c0}}{F_{gm,av}}, \quad (3)$$

where $F_{rn,av}$ is the averaged value of Syto-61 fluorescence intensity F_r saved in red channel for all nuclear voxels and $F_{gm,av}$ is that of MitoTracker Orange F_g saved in green channel for all mitochondrial voxels. With only three adjustable parameters of n_{c0} , $n_{n,av}$ and $n_{m,av}$, the above OCM equations provide a practical and objective approach to model the optical response of “molecular composition” using a realistic cell structure. To further improve the OCM defined above, small spheres with Gaussian distributions of RI and radius were inserted in cytoplasm to simulate the effect of lysosomes in light scattering by prostate cells [7]. The volume ratio of lysosomes to cell in the OCMs developed here was set to 1% with mean and SD values of radius given by 0.2 and 0.08 μm , and those of RI are denoted as $n_{l,av}$ and $\Delta n_{l,av}$.

We have recently applied the GLCM algorithm to quantify the diffraction patterns embedded in the p-DI pairs calculated by OCMs of different values for n_{c0} , $n_{n,av}$ and $n_{m,av}$. It has been found that the shape irregularity and RI heterogeneity in the organelles play very significant roles in the diffraction patterns in comparison to the average RI values of the organelles [13]. Consequently, we present results here with only one set of n_{c0} , $n_{n,av}$, $n_{m,av}$ and

$n_{l,av}$ for all OCMs selected from simulation data calculated with different set of RI values.

The cell structures of PC3-a, PCS-a and PCS-b were selected from the reconstruction 3D voxel arrays to obtain OCMs for this study. The nucleus of each cell structure was varied by dilation or erosion of integer number of nuclear membrane voxel layers to modify OCMs. We define a nuclear volume variation ratio r_{nv} to characterize modified OCMs against the one built with the original cell structure. All OCMs of either PC3-a or PCS-a combined with PCS-b structures were divided into three nuclear size groups. The OCM_S group consists of those with r_{nv} ranging from 39% to 61%, the OCM_O group from 89% to 111% and the OCM_L group from 139% to 161%. For the PCS-b cell structure, the upper bounds of OCM_O and OCM_L groups were reduced to 109% and 159% to shorten simulation time. Each group has 5 to 11 OCMs of different r_{nv} values with variable stepsizes of 2%, 3% or 4% so that 50%, 100% and 150% is at the middle of range. Furthermore, the orientation of each OCM is characterized by (α, β, γ) as the Euler angles in the coordinate system defined by the incident light and imaging directions as shown in Figure 1A. A total of 22 orientations were applied for calculation of p-DI pairs to assess the effect of orientation on diffraction patterns of p-DI data and classification. Based on our previous p-DIFC studies, orientations of flowing cells in are not randomized since different cell types of high morphological similarity can be accurately classified by the measured p-DI data [18, 20, 22, 25]. One can therefore conclude that the cells are rather aligned around certain preferred direction. For this study we limited the orientation change by variation of α only from 25.0° to 35.5° with a stepsize of 0.5°. Table 1 lists the morphology parameters of the two cell structures used in this study and the total number of OCMs for each of the three size groups.

2.2 | Configuration of diffraction imaging and simulation for calculated p-DI pairs

For p-DI measurement, suspension samples of detached prostate cells were prepared in a concentration of about 2×10^6 cells/mL and injected into the core fluid stream of a p-DIFC system [15–18, 22]. A continuous-wave and linear polarized laser beam of $\lambda = 532$ nm in wavelength is propagated along the z-axis and focused on the core fluid that carries the cell suspension through the focal spot around 30 μm . A flowing cell scatters the incident coherent light as it passes through the focus due to RI mismatch between cell and core fluid as host medium.

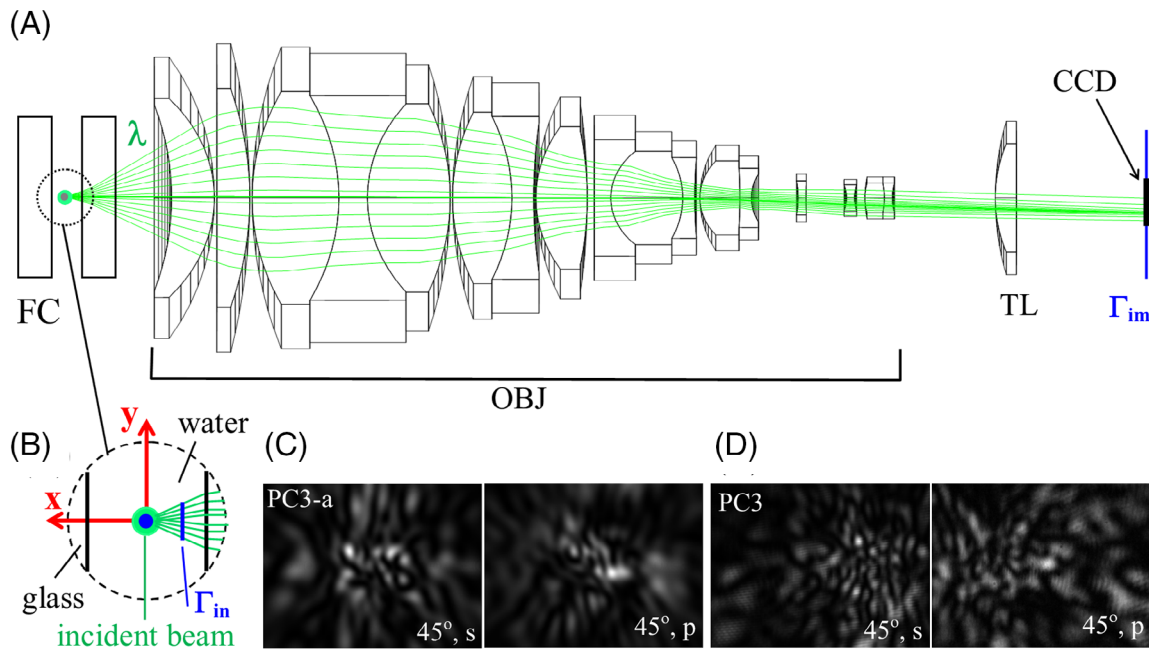


FIGURE 1 A, The imaging configuration of one polarized branch for p-DI acquisition and calculation with the green lines representing the incident and scattered light of $\lambda = 532$ nm; FC, flow chamber; TL, tube lens; CCD, imaging sensor; Γ_{in} , imaging plane (blue line); B, magnified view of the scattering configuration and coordinate axes (red lines) with incident beam along the z-axis; Γ_{in} : input plane defined inside the flow chamber (blue line); C, one calculated p-DI pair by an OCM of PC3 cell with the polarizations of incident beam and scattered light marked; D, one measured p-DI pair of a PC3 cell. OCM, optical cell model

TABLE 1 Morphology parameters of three prostate cells and number of OCMs^a

Cell ID	V_c (μm^3)	SVr_c (μm^{-1})	ER_c (μm)	$\langle R_c \rangle$ (μm)	Vr_{nc} (%)	SVr_n (μm^{-1})	Vr_{mc} (%)	SVr_m (μm^{-1})	# of OCMs ^a		
									OCM _S	OCM _O	OCM _L
PC3-a	2250	0.500	8.13	8.35	37.0	0.716	7.66	5.86	11 × 22	11 × 22	11 × 22
PCS-a	1118	0.614	6.44	6.56	30.1	0.964	4.71	6.42	11 × 22	11 × 22	11 × 22
PCS-b	1622	0.607	7.29	7.62	24.6	0.957	28.1	3.93	10 × 22	5 × 22	5 × 22

Abbreviations: OCMs, optical cell models; $\langle R_c \rangle$, average distance of cell membrane voxels to centroid; ER_c , equivalent radius of cell; SVr_c (SVr_n or SVr_m), surface-to-volume ratio of cell (nucleus or mitochondrion); V_c , cell volume; Vr_{nc} (Vr_m), volume ratio of nucleus/mitochondrion-to-cell.

^aTotal number of OCMs in each group are given by the number of OCMs of different nuclear sizes multiplied by 22 as the number of orientations.

Collection of the scattered light from the glass flow chamber is achieved with an imaging unit along the side directions within a cone angle as shown in Figure 1B. The imaging unit consists of an infinity-corrected $\times 50$ objective of 0.55 in NA (378-805-3, Mitutoyo), a polarizing beam splitter for separating scattering light into s- and p-polarized branches for p-DI acquisition with one CCD camera (LM075, Lumenera) per branch. The unit, with camera sensor fixed to the focal plane of tube lens, can be translated away from the focused position by Δx toward the flow chamber ($\Delta x > 0$) to increase the

contrast and vary collection angular cone [16, 26]. Acquisition and simulation of p-DI data were performed with Δx set to 150 μm for PC3 and PCS cells.

To simulate diffraction imaging, we first employed an open-source code of ADDA based on the discrete-dipole approximation to determine the angular distribution of scattered light [36]. Each OCM listed in Table 1 was used as the input data for the ADDA code with incident light wavelength set to $\lambda = 532$ nm. The simulations were performed with ratios of wavelength to voxel size or dipole-per-wavelength values ranging from 5 to 10 and output

data in the forms of angle-resolved 4×4 Mueller matrices. The imaging configuration shown in Figure 1 is implemented in a ray-tracing optical design software (Zemax, 2009) to trace scattered light rays. The ray tracing starts from the virtual input plane Γ_{in} defined in the flow chamber with a distance of $150 \mu\text{m}$ from the scatterer located at the origin. It proceeds through the host medium of water, chamber glass, air and imaging unit to the imaging plane Γ_{im} as depicted in Figure 1A,B. The calculated p-DI pairs were obtained after completion of ray tracing for all scattering angles within the collection cone of the objective. Different combinations of incident beam and scattered light polarizations for each p-DI pair were determined with linear combinations of the Mueller matrix elements produced by ADDA. More details of diffraction imaging simulation and validation against measured data can be found elsewhere [13, 24, 27, 37].

2.3 | Image characterization by GLCM parameters and classification

The second-order statistical algorithm of GLCM has been applied to characterize each input image of 8-bit pixel intensity [34]. The algorithm outputs a matrix of rank $256 (=2^8)$ has its elements defined as the relative frequency of paired pixels having their intensities given by the row and column numbers of the elements. For this study, the pixel distance in each pair was set to 1 and four matrices were obtained for the directions of pixel pairs given by 0° , 45° , 90° and 135° . The input image textures are quantified by 15 GLCM parameters calculated from each matrix and the averaged GLCM parameters over the four directional matrices were used for subsequent classification. Each calculated p-DI pair of 16-bit pixels was normalized into images of 8-bit pixels to determine a total of 32 GLCM parameters with definitions and symbols of GLCM parameters given in the Support Information.

We have investigated unsupervised and supervised classifiers for binary classification of p-DI pairs calculated with the OCMs of two nuclear size groups for each of the PC3 and PCS cell structures. Different clustering algorithms of k-means, hierarchical and GMM have been applied as the unsupervised classifiers [38, 39]. It has been found that the best performing clustering classifier is given by combining the hierarchical with GMM clustering algorithms for high stability [37]. The combined clustering classifier is denoted in this report as hGMM, which starts hierarchical clustering by assuming each p-DI pair being a cluster of its own in the GLCM parameter space. It then iterates by linking two clusters of the shortest distance at a time and iteration ends when the

total number of clusters is reduced to $k = 2$. The output is imported into the GMM algorithm to obtain a Gaussian probability density function (pdf) for each of the two clusters in the parameter space. Classification is optimized by maximizing iteratively a likelihood function L defined as the logarithmic sum of pdf's over the p-DI data represented by the GLCM parameters assigned to different clusters. Each iteration varies the cluster assignment and pdf parameters to increase L until its value stabilizes [40]. Afterwards, each cluster, C_1 or C_2 , was assigned to one OCM size group as defined in Table 1 that yields highest classification accuracy A_{hGMM} . The standard definition of classification accuracy, A_{hGMM} or A_{SVM} , is adopted to measure the performance of either hGMM or SVM classifiers that is given by the number ratio of correctly identified p-DI pairs to the total number of p-DI pairs.

Different from unsupervised methods of clustering, an SVM classifier requires training to learn image patterns characterized by the GLCM parameters in our case. The algorithm maps the GLCM parameters into a feature space defined by a kernel function and solve a quadratic optimization problem using the training data. The solution yields an optimized kernel function and a decision function which can be applied to the test data to determine $A_{SVM,tes}$. We employed the SVM tool provided by Matlab for this study and both the hGMM code and interface code to the SVM tool were developed in the same platform.

3 | RESULTS

3.1 | Confocal imaging of cell morphology and OCMs of varied nuclear sizes

Confocal image stacks were acquired from PC3 and PCS cells with a stepsize of $0.5 \mu\text{m}$ along the direction perpendicular to the object plane of the microscope objective. Each image slice has 512×512 pixels which was segmented into three sets of nucleus, mitochondria and cytoplasm in the red and green color channels recording the two fluorescent intensities. Interpolated slices were added during reconstruction to obtain a 3D array of nearly cubic voxels for each imaged cell [29, 30]. With the confocal image stacks of PC3-a, PCS-a and PCS-b selected for reconstruction, simulations of diffraction imaging process were performed with a total of 1892 OCMs as listed in Table 1 with 726 for the PC3 (PC3-a) and 1166 for the PCS (PCS-a and PCS-b) structures. Each OCM is specified by its cell ID, nuclear volume variation ratio r_{nv} and orientation angles (α, β, γ) . Light scattering by a cell was

modeled by assuming an OCM placed in the host medium of water with RI given by $n_h = 1.336$ and a linearly polarized incident beam of three polarization directions and $\lambda = 532$ nm. We have investigated different values of averaged RI for nuclear and mitochondrial voxels of OCMs as defined in Equations (2) and (3) from 1.39 to 1.55 and found GLCM parameters of the calculated p-DI data exhibiting very weak dependence on averaged RI values [13]. Based on these results, the input parameters for all OCMs in this study were set to $n_{c0} = 1.345$, $n_{n,av} = 1.442$, $n_{m,av} = 1.450$, $n_{l,av} = 1.440$ and $\Delta n_{l,av} = 0.02$ [7–11]. Simulations with another set of higher RI values of $n_{c0} = 1.361$, $n_{n,av} = 1.486$, $n_{m,av} = 1.500$, $n_{l,av} = 1.480$ and $\Delta n_{l,av} = 0.02$ have been carried out and the above finding was confirmed with similar classification accuracies.

Figure 2 compares a center confocal image slice acquired from the PC3-a structure and corresponding false-color center image slices of OCMs of different r_{nv} values. Similar comparisons are shown in Figure 3 for the PCS-a and PCS-b structures. For OCMs of $r_{nv} < 100\%$, the RI values of selected nuclear voxels near the membrane were replaced by those of randomly chosen cytoplasmic voxels calculated by Equation (1) for the nuclear

voxels. For OCMs of $r_{nv} > 100\%$, the RI values of selected cytoplasmic or mitochondrial voxels were replaced by those of randomly chosen nuclear voxels calculated by Equation (1). These images show clearly that the use of OCM to investigate the effect of different nucleus volumes on p-DI patterns has the unique advantage of keeping other organelles unchanged or minimally changed which is difficult, if not impossible, to achieve experimentally. Since PC3 cells on average are much larger in both cellular and nuclear volumes than the PCS cells [22], we performed binary classification of p-DI data among the three OCM groups of different nuclear sizes built with PC3 and PCS cell structures separately. The OCM groups of the PCS structures consist of those based on the PCS-a and PCS-b cells shown in Figure 3.

3.2 | Calculations of p-DI data with OCMs of different nuclear sizes and orientations

Calculations of p-DI pairs were performed with all OCMs listed in Table 1. Following our previous experimental and numerical investigations, three cases of linear

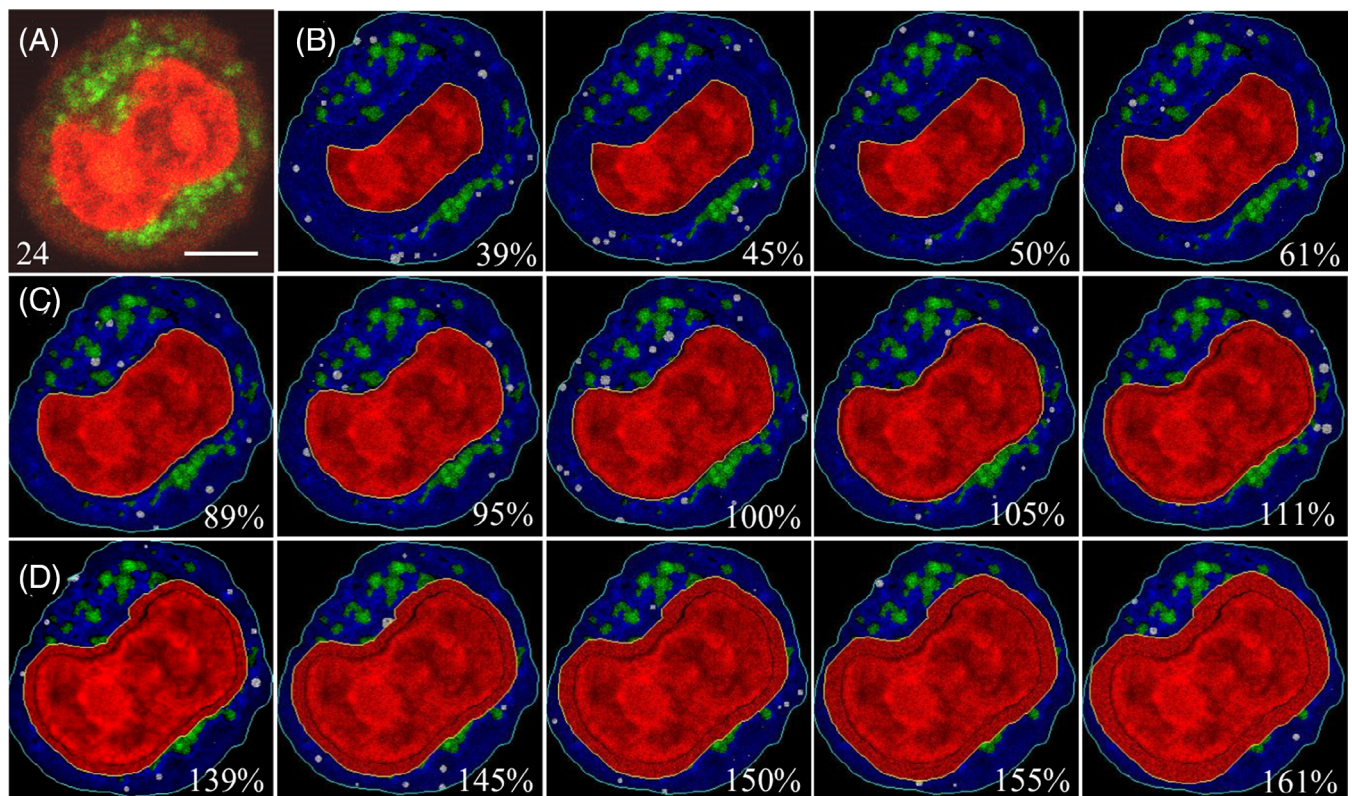


FIGURE 2 One center slice of: A, the confocal image stack acquired from the PC3-a cell with slice number marked; corresponding center slices of different OCMs in the group of B, OCM_S; C, OCM_O; D, OCM_L. Different voxel colors are used in OCM slices from B to D to mark organelles with brightness indicating RI values; nucleus: red, cytoplasm: blue, mitochondria: green, lysosome: white. Each OCM center slice is marked with r_{nv} values for nuclear volume variation ratio and white lines indicate cytoplasmic and nuclear membranes. Bar = 5 μ m. OCM, optical cell model

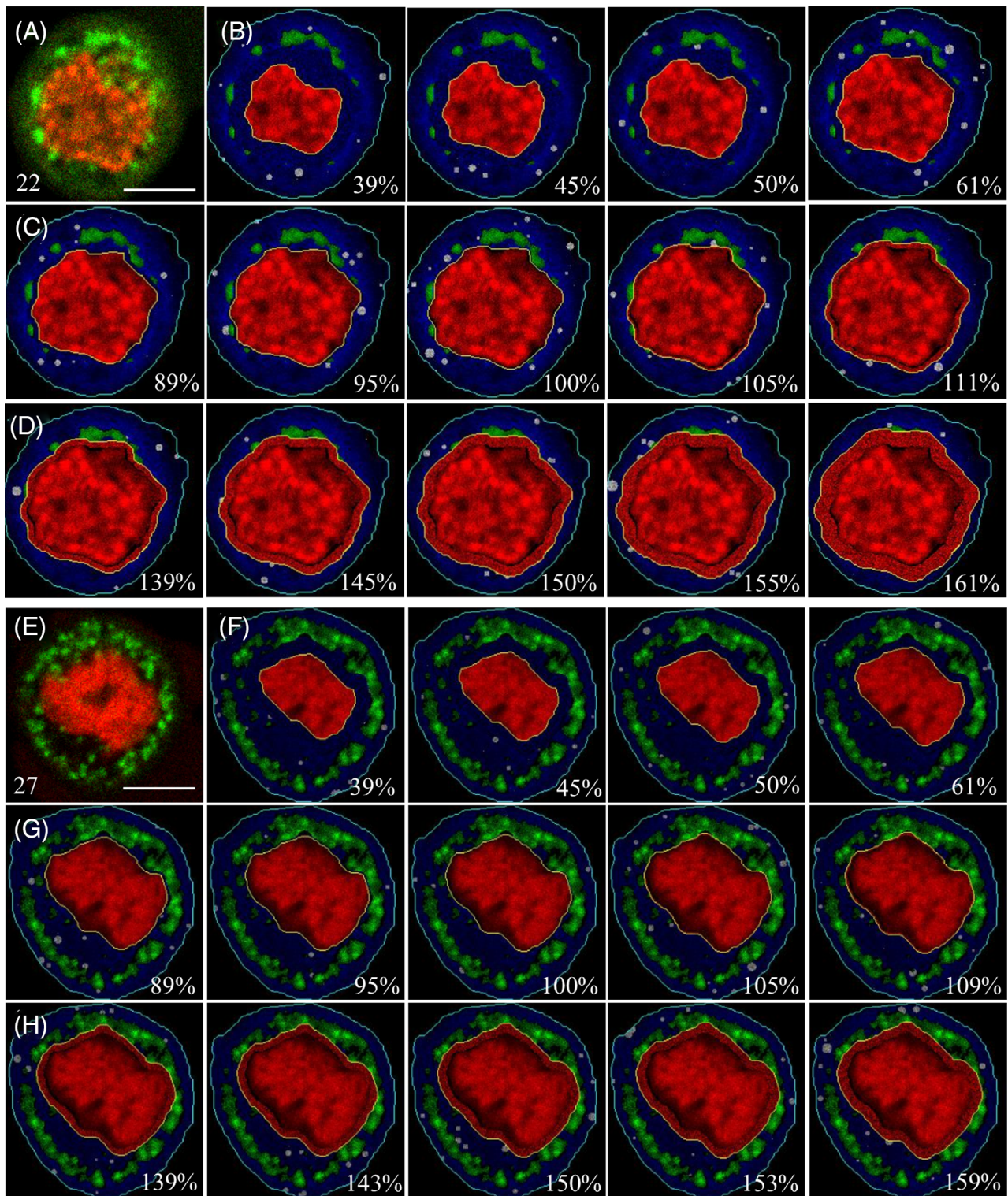


FIGURE 3 Similar to Figure 2 for the two PCS cells and corresponding OCMs in three groups with A to D for the PCS-a cell structure and E to H for the PCS-b cell structure. Bar = 5 μm . OCM, optical cell model

polarization direction were considered for the incident laser beam and denoted as ver for along the y-axis, hor for along the x-axis and 45° for between ver and hor in the x-y plane as depicted in Figure 1A. A total of 5676 p-DI pairs were calculated though light scattering

modeling with the ADDA code followed by image simulation with the Matlab and Zemax based ray-tracing code with the off-focus distance $\Delta x = 150 \mu\text{m}$. The ADDA simulations were performed on the computing cluster of the Institute for Advanced Optics, which took

about 1 to 4 hours to complete calculation of the angle-resolved Mueller matrices for one OCM of the PC3-a structure on one CPU (Xeon E5-2650V4, Intel). For the smaller PCS-a and PCS-b structures, p-DI calculations took less time with 10 to 40 minutes for one OCM. In comparison, projection and ray-tracing of the Mueller matrix elements can complete in about 10 minutes for three pairs of p-DI of different incident beam polarizations on one OCM with one CPU (i5-9400, Intel). Figures 4 and 5 show the calculated pairs of p-DI for the PC3-a and PCS-a plus PCS-b structures.

For each nuclear size group shown in the two figures, two p-DI pairs are included for visual comparison of pattern differences by two OCMs of same r_{nv} but different α values by 10.5° . It can be clearly seen that the orientational variations of OCM lead to pattern changes to a certain extent as a result of the nonspherical and highly heterogeneous intracellular organelles. We observe further a trend of speckle size decrease in p-DIs as the r_{nv} value increases that reflects the fundamental relation between scatterer size and angular size of diffraction speckles based on, for example, the Mie theory for spheres [24, 37]. The trend is more visible in Figure 5 for the cases of PCS-a and PCS-b cell structures of small volumes.

3.3 | Classification among different groups of nuclear size

The calculated p-DI data were divided into three groups of OCM_S , OCM_O and OCM_L for binary classification among the three groups of the same incident beam polarization and cell structure of PC3 or PCS. Each p-DI pair in a group was processed by the GLCM algorithm to convert each image in the pair into 16 parameters as defined

in the SI file with 15 for characterizing image texture and 1 as the mean pixel intensity. Thus, each p-DI pair was characterized by 32 GLCM parameters for subsequent classification. The GLCM parameters of p-DI data were pooled according to the paired OCM groups for binary classification based on the same cell structure and incident beam polarization. For example, the GLCM parameters of all p-DI data in the OCM_S group were combined with those in OCM_O group for one pool and with those in OCM_L for another pool. Then the values of each GLCM parameter for all p-DI pairs in a pool were normalized by the same maximum and minimum values of that parameter in the pool for binary classification by either classifier of hGMM or SVM. The unsupervised and supervised classification were carried out on a computer with one CPU (i5-9400, Intel) and the time to complete an hGMM classification was about 3 seconds while SVM training and test with a selected kernel function took about 20 and 2 seconds respectively.

Table 2 lists the values of N_{tot} as the total number of p-DI pairs in each pool for binary classifications by the hGMM method among the three OCM groups of different r_{nv} for the same cell structure and incident beam polarization. The numbers of training and test datasets for SVM classification are recorded together with N_{tot} . It also includes the values of classification accuracy A_{hGMM} which show poor performance of the clustering method with $A_{hGMM} < 80\%$. We have also investigated hGMM classification with different subsets of the 32 GLCM and intensity parameters which yielded similarly low values of A_{hGMM} . Additional tests of the k -means clustering algorithms have also been carried out which yielded poor performance as well. To understand these results by the clustering methods, we examined the distributions of p-DI pairs in different pools by scatter plots in various subspaces of GLCM parameters. Two examples are presented

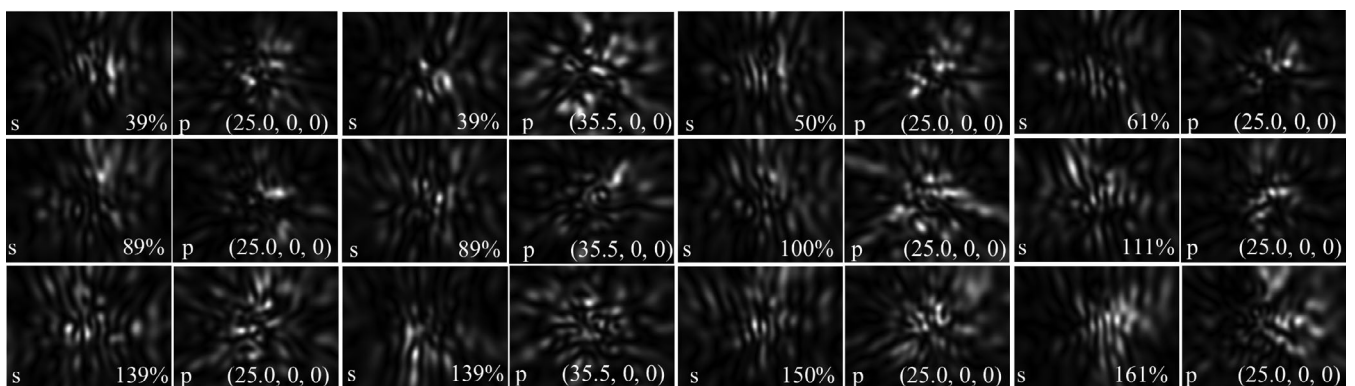


FIGURE 4 Normalized p-DI pairs calculated with PC3-a structure and different values of r_{nv} and (α, β, γ) as orientation angles: OCM_S (top row); OCM_O (middle row); OCM_L (bottom row). Each pair is marked with polarization of the scattered light and values of r_{nv} and (α, β, γ) . The incident beam polarization was set to ver. OCM, optical cell model

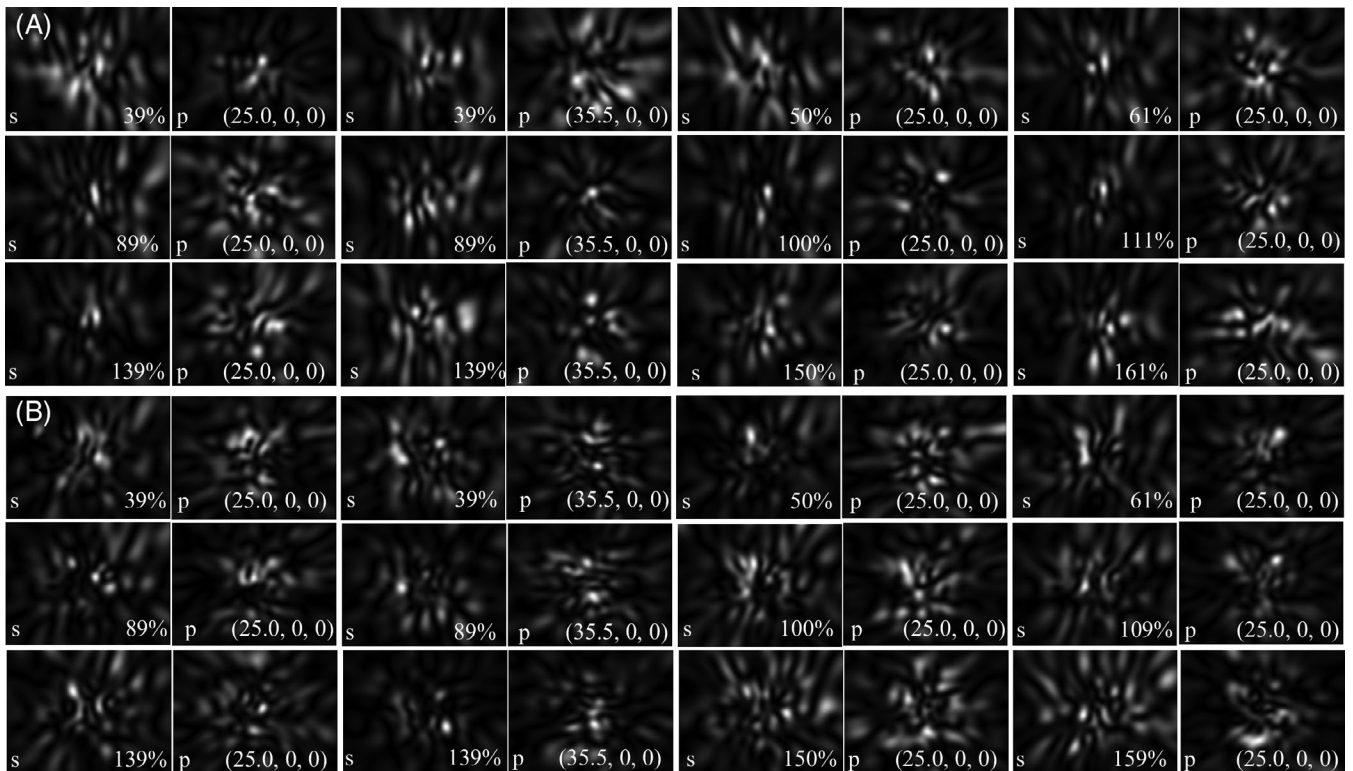


FIGURE 5 Similar to Figure 4 with: A, PCS-a structure; B, PCS-b structure

in Figure 6 to show unambiguously that p-DI pairs in the same pool are not linearly separable in the GLCM parameter space. It becomes obvious from these plots that the classification problems encountered here are highly nonlinear and the distributions of p-DI data are non-Gaussian. These characteristics lead to the poor performance of hGMM and other clustering methods.

To improve performance of binary classification, we conducted SVM based supervised machine learning of diffraction patterns by dividing the p-DI data in each pool randomly into training and test datasets for each combination of cell structures and incident beam polarization as listed in Table 1. A scheme of 5-fold cross-validation was employed to optimize SVM classifiers by the training dataset with a selected kernel function [22]. The averaged accuracy of binary classifications performed on the training dataset by the 5-fold cross-validation and that of the test dataset are presented respectively as $A_{SVM,tra}$ and $A_{SVM,tes}$ in Table 3. Comparison of the accuracy values in Tables 2 and 3 demonstrate clearly that SVM outperforms clustering algorithms significantly in classifying cells among the three nuclear size groups. The confusion matrices in Figure 7 presents additional details on performance of the two classifiers. It is interesting to note that the values of classification accuracy stay about the same for both cell structures or classifiers despite the much

larger differences in r_{nv} values between OCM_S and OCM_L than those of the other two pools.

4 | DISCUSSION

Clinical research signifies the utility of nuclear size estimation or measurement in terms of N:C ratio [2]. Figure 8 presents a cytology smear image of fine-needle aspirate from a patient with lung adenocarcinoma, which shows clearly tumor cells of large N:C ratios. Consequently, development of accurate and automated approaches for future technical development should have wide implications in diagnosis of atypia and malignancy. Compared to existing methods by conventional microscopy, single-shot diffraction imaging by coherent light may offer competitive advantages to probe without labeling cells for classification. A significant challenge is to develop tools for survey and evaluation of correlations between the nuclear size change and the diffraction image features that can be measured and extracted rapidly. By classifying p-DI pairs divided into three nuclear size groups using OCM tools, we have shown that the SVM classifiers perform well over clustering methods on distinguishing cells of different nuclear sizes and the potential of p-DIFC method for such applications. For the larger PC3-a

TABLE 2 Datasets and classification accuracy by hGMM method

Cell structure	pol	Binary groups ^a	N_{tot}	N_{tra}	N_{tes}	$A_{\text{hGMM}} (\%)$		
						S vs O	O vs L	S vs L
PC3-a	ver	All	2×242	2×170	2×72	66.1	58.1	55.0
	hor	All	2×242	2×170	2×72	51.9	70.3	76.2
	45°	All	2×242	2×170	2×72	63.6	71.9	77.5
PCS-a & PCS-b	ver	S vs O or L	$462 + 352$	2×285	$177 + 67$	46.3		57.9
		O vs L	2×352	2×285	2×67		53.6	
	hor	S vs O or L	$462 + 352$	2×285	$177 + 67$	63.3		58.6
		O vs L	2×352	2×285	2×67		50.9	
	45°	S vs O or L	$462 + 352$	2×285	$177 + 67$	56.5		52.1
		O vs L	2×352	2×285	2×67		65.1	

Abbreviations: N_{tes} , number of p-DI pairs in test dataset; N_{tot} , total number of p-DI pairs of a pool for binary classification; N_{tra} , number of p-DI pairs in training dataset; pol, incident beam polarization.

^aFor binary classification of OCM groups: S = OCM_S , O = OCM_O , L = OCM_L .

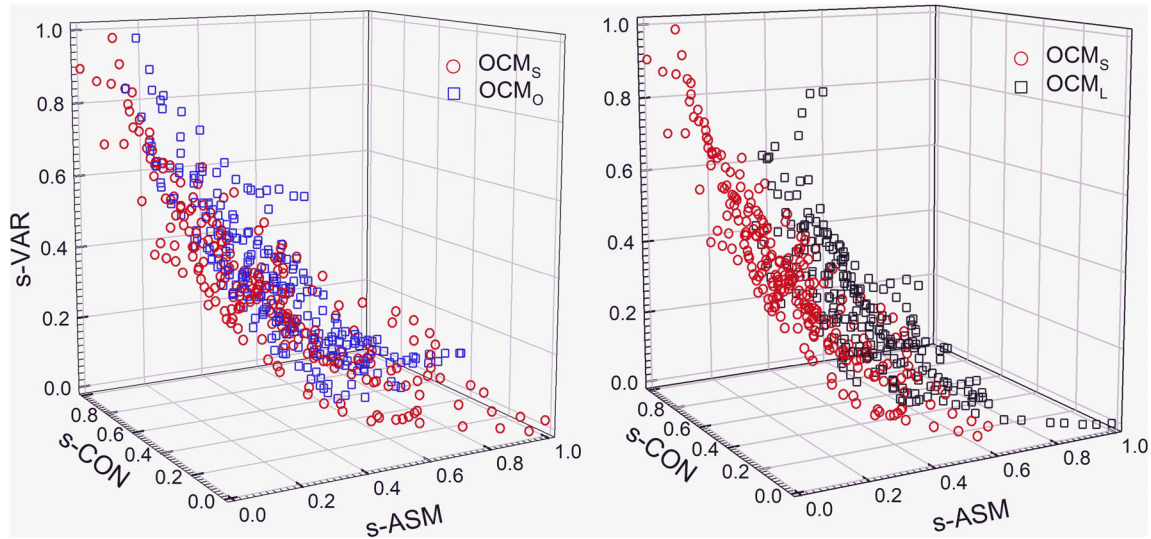


FIGURE 6 Scattering plots of s-polarized images calculated with PC3-a structure and incident beam polarization set to ver for the pools of OCM_S vs OCM_O (left) and of OCM_S vs OCM_L (right) in the subspace of GLCM parameters by ASM (angular second moment), CON (contrast) and VAR (variance). Note that the same data set of OCM_S appears differently in the two plots due to the pool dependence of GLCM parameter normalization. OCM, optical cell model. GLCM, gray-level co-occurrence matrix

structure, the averaged value and standard deviation of A_{SVM} for the test dataset over the 3 incident beam polarizations is $98.8\% \pm 0.7\%$ and $97.5\% \pm 2.3\%$ respectively for the binary classifications of OCM_S vs OCM_O groups and OCM_O vs OCM_L groups. For the smaller PCS structure with the PCS-a and PCS-b cells, the corresponding A_{SVM} values remain about the same at $98.9\% \pm 0.5\%$ and $97.8\% \pm 0.6\%$ respectively.

The robust performance of the SVM method can be attributed to its ability for effectively solving nonlinear classification problems in the feature space instead of the GLCM parameter space. By training a kernel function to

optimize mapping, the final feature space enables linear separability of the input GLCM data and the classifier is evaluated further with the held-out test data. The significant differences between the two sets of confusion matrices for hGMM and SVM in Figure 7 provide strong evidences for using feature space to classify cells of different nuclear sizes with GLCM parameters of non-Gaussian distribution. Since SVM can be represented as a shallow neural network of single hidden layer [41], it becomes apparent that the features extracted by SVM form a parameter space defined by the correlations among the GLCM parameters. Since GLCM characterizes

TABLE 3 Classification parameters and accuracy by SVM method^a

Cell structure	pol	$A_{SVM,tra}$; $A_{SVM,tes}$ (%) and KF ^b					
		S vs O	KFs	O vs L	KFs	S vs L	KFs
PC3 (PC3-a)	ver	100; 99.3	C, Q	99.7; 94.4	C	100; 100	C
	hor	98.5; 97.9	C	98.2; 97.9	C	99.4; 98.6	C
	45°	100; 99.3	C, Q, L, G	99.7; 100	Q	100; 99.3	C, Q, G
PCS (PCS-a & PCS-b)	ver	98.4; 99.6	C	98.2; 97.0	C	98.9; 98.4	C
	hor	98.9; 98.8	C	98.8; 98.5	C	98.2; 98.0	C
	45°	99.5; 98.4	C	99.3; 97.8	C	97.7; 97.5	C, G

^aThe number of p-DI pairs in the training and test datasets are in Table 2, pol = incident beam polarization, S = OCM_S, O = OCM_O, L = OCM_L.

^bThe kernel functions (KFs) used for the best classification accuracy of training ($A_{SVM,tra}$) and test data ($A_{SVM,tes}$): C, cubic; Q, quadratic; L, linear; G, medium Gaussian; definitions of KFs are provided on the Mathworks website.

FIGURE 7 Confusion matrices of binary classification of nuclear size groups with incident beam polarization set to ver and group notation of S = OCM_S, O = OCM_O, L = OCM_L; A and B, hGMM method on all data; C and D, SVM method on test datasets. Rows represent ground truth and the blue squares indicate zero elements. OCM, optical cell model

(A) PC3 structure; hGMM method	Group	C ₁	C ₂	Group	C ₁	C ₂	Group	C ₁	C ₂
	S	179	63	O	110	132	S	127	115
(B) PCS structure; hGMM method	Group	C ₁	C ₂	Group	C ₁	C ₂	Group	C ₁	C ₂
	S	240	222	O	270	82	S	334	128
(C) PC3 structure; SVM method	Group	S	O	Group	O	L	Group	S	L
	S	72		O	67	5	S	72	
(D) PCS structure; SVM method	Group	S	O	Group	O	L	Group	S	L
	S	176	1	O	64	3	S	173	4

the second-order correlations of the input images, our results thus demonstrate the strong ability of GLCM combined with SVM to solve the classification problem here. These data further reveal the capacity of the p-DIFC method to extract high-order correlations of the diffraction patterns embedded in p-DI pairs. These conclusions are consistent with those of our recent study on classification of five cell types by a convolutional neural network on measured p-DI data with a minimum of three convolutional layers [25]. Taken together, the presented results suggest that the p-DIFC method may provide in the future an automated approach to classify atypical cells of large nuclear size against the normal ones by extraction of high-order correlations from p-DI data. Research is

underway to apply the tools developed here for detection of cells with large N:C ratios in human pleural effusion samples by the measured and calculated p-DI data.

5 | CONCLUSION

We have developed useful tools through this study to calculate p-DI pairs using realistic OCMs of human prostate cells and compare different algorithms for binary classification. The OCMs have been developed to represent nucleus, mitochondria, lysosomes and cytoplasm as intracellular RI distributions for simulation of diffraction imaging and calculation of p-DI pairs. With the GLCM

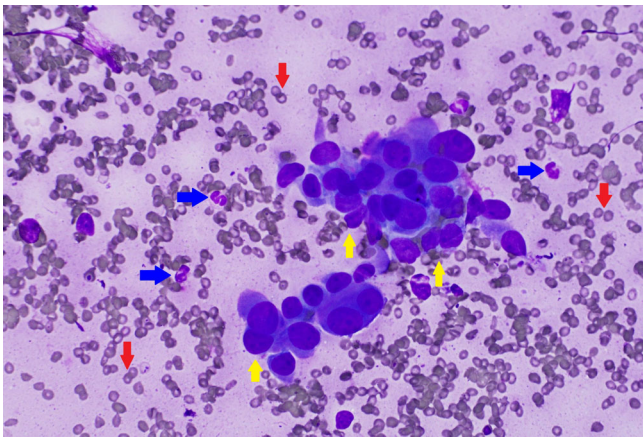


FIGURE 8 A Diff-Quik stained cytology smear of fine-needle aspirate from a lung cancer patient. The stained malignant cells exhibit large nuclear sizes in purple color and examples are marked by yellow-up arrows. Examples of many red cells of small sizes in background are marked by red-down arrows and several neutrophils are marked by blue-right arrows

algorithm to characterize the second order of pixel intensity in the p-DI data, we have shown that the SVM algorithm outperforms various clustering methods by solving the nonlinear classification problems in a feature space. It was further demonstrated that the ability to extract high-order correlation features from diffraction image patterns is critical for SVM to achieve high classification accuracies for the prostate cell structures and all three polarization directions of the incident laser beam. These results present a proof-of-concept for future development of label-free methods to automate accurate assay of nuclear size and conditions in biological cells.

ACKNOWLEDGMENTS

J. L. acknowledges grant supports of #19A198 and to Hunan Key Laboratory of Photonics and Optical Communications by the Science and Technology Department of Hunan Province and W. W. acknowledges grant supports of #18B348 by the Education Department of Hunan Province.

ORCID

Jing Liu  <https://orcid.org/0000-0001-6742-069X>

Yaohui Xu  <https://orcid.org/0000-0003-1976-8151>

Peng Tian  <https://orcid.org/0000-0003-3757-8298>

Xin-Hua Hu  <https://orcid.org/0000-0002-4353-9028>

REFERENCES

- [1] D. Zink, A. H. Fischer, J. A. Nickerson, *Nat. Rev. Cancer* **2004**, *4*, 677.
- [2] L. J. Vaickus, R. H. Tambouret, *Cancer Cytopathol.* **2015**, *123*, 524.
- [3] T. Takaki, M. Montagner, M. P. Serres, M. Le Berre, M. Russell, L. Collinson, K. Szuhai, M. Howell, S. J. Boulton, E. Sahai, M. Petronczki, *Nat. Commun.* **2017**, *8*, 16013.
- [4] N. Agarwal, A. M. Biancardi, F. W. Patten, A. P. Reeves, E. J. Seibel, *J. Med. Imaging (Bellingham)* **2014**, *1*, 017501.
- [5] P. J. McIntire, J. T. Snow, S. S. Elsoukkary, L. Soong, J. Sweeney, B. D. Robinson, M. T. Siddiqui, *Cancer Cytopathol.* **2019**, *127*, 120.
- [6] N. Lue, W. Choi, G. Popescu, K. Badizadegan, R. R. Dasari, M. S. Feld, *Opt. Express* **2008**, *16*, 16240.
- [7] O. C. Marina, C. K. Sanders, J. R. Mourant, *Biomed. Opt. Express* **2012**, *3*, 296.
- [8] T. Kim, R. Zhou, M. Mir, S. D. Babacan, P. S. Carney, L. L. Goddard, G. Popescu, *Nat. Photon.* **2014**, *8*, 256.
- [9] M. Schürmann, J. Scholze, P. Müller, J. Guck, C. J. Chan, *J. Biophotonics* **2016**, *9*, 1068.
- [10] Z. A. Steelman, W. J. Eldridge, J. B. Weintraub, A. Wax, *J. Biophotonics* **2017**, *10*, 1714.
- [11] M. Habaza, M. Kirschbaum, C. Guernth-Marschner, G. Dardikman, I. Barnea, R. Korenstein, C. Duschl, N. T. Shaked, *Adv. Sci.* **2017**, *4*, 1600205.
- [12] M. M. Villone, P. Memmolo, F. Merola, M. Mugnano, L. Miccio, P. L. Maffettone, P. Ferraro, *Lab Chip* **2017**, *18*, 126.
- [13] S. Wang, J. Liu, J. Q. Lu, W. Wang, S. A. Al-Qaysi, Y. Xu, W. Jiang, X. H. Hu, *J. Biophotonics* **2019**, *12*, e201800287.
- [14] M. A. Yurkin, *J. Biophotonics* **2018**, *11*, e201800033.
- [15] K. M. Jacobs, J. Q. Lu, X. H. Hu, *Opt. Lett.* **2009**, *34*, 2985.
- [16] K. M. Jacobs, L. V. Yang, J. Ding, A. E. Ekpenyong, R. Castellone, J. Q. Lu, X. H. Hu, *J. Biophotonics* **2009**, *2*, 521.
- [17] Y. Sa, Y. Feng, K. M. Jacobs, J. Yang, R. Pan, I. Gkigkitzis, J. Q. Lu, X. H. Hu, *Cytometry A* **2013**, *83*, 1027.
- [18] Y. Feng, N. Zhang, K. M. Jacobs, W. Jiang, L. V. Yang, Z. Li, J. Zhang, J. Q. Lu, X. H. Hu, *Cytometry A* **2014**, *85*, 817.
- [19] H. Wang, Y. Feng, Y. Sa, Y. Ma, R. Pan, J. Q. Lu, X. H. Hu, *Appl. Optics* **2015**, *54*, 5223.
- [20] K. Dong, Y. Feng, K. M. Jacobs, J. Q. Lu, R. S. Brock, L. V. Yang, F. E. Bertrand, M. A. Farwell, X. H. Hu, *Biomed. Opt. Express* **2011**, *2*, 1717.
- [21] X. Yang, Y. Feng, Y. Liu, N. Zhang, W. Lin, Y. Sa, X. H. Hu, *Biomed. Opt. Express* **2014**, *5*, 2172.
- [22] W. Jiang, J. Q. Lu, L. V. Yang, Y. Sa, Y. Feng, J. Ding, X. H. Hu, *J. Biomed. Opt.* **2016**, *21*, 071102.
- [23] H. Wang, Y. Feng, Y. Sa, J. Q. Lu, J. Ding, J. Zhang, X.-H. Hu, *Pattern Recognit.* **2017**, *61*, 234.
- [24] W. Wang, J. Liu, J. Q. Lu, J. Ding, X. H. Hu, *Opt. Express* **2017**, *25*, 9628.
- [25] J. Jin, J. Q. Lu, Y. Wen, P. Tian, X. H. Hu, *J. Biophotonics* **2020**, *13*, e201900242.
- [26] R. Pan, Y. Feng, Y. Sa, J. Q. Lu, K. M. Jacobs, X. H. Hu, *Opt. Express* **2014**, *22*, 31568.
- [27] J. Zhang, Y. Feng, W. Jiang, J. Q. Lu, Y. Sa, J. Ding, X. H. Hu, *Opt. Express* **2016**, *24*, 366.
- [28] R. S. Brock, X. H. Hu, D. A. Weidner, J. R. Mourant, J. Q. Lu, *J. Quant. Spectrosc. Radiat. Transfer* **2006**, *102*, 25.
- [29] Y. Zhang, Y. Feng, C. R. Justus, W. Jiang, Z. Li, J. Q. Lu, R. S. Brock, M. K. McPeck, D. A. Weidner, L. V. Yang, X. H. Hu, *Integr. Biol.* **2012**, *4*, 1428.
- [30] Y. Wen, Z. Chen, J. Lu, E. Ables, J. L. Scemama, L. V. Yang, J. Q. Lu, X. H. Hu, *PLoS One* **2017**, *12*, e0184726.

- [31] H. Shahin, M. Gupta, A. Janowska-Wieczorek, W. Rozmus, Y. Y. Tsui, *Opt. Express* **2016**, *24*, 28877.
- [32] J. Stark, T. Rothe, S. Kiess, S. Simon, A. Kienle, *Phys. Med. Biol.* **2016**, *61*, 2749.
- [33] X. Lin, N. Wan, L. Weng, Y. Zhou, *Appl. Optics* **2017**, *56*, 3608.
- [34] R. M. Haralick, *Proc. IEEE* **1979**, *67*, 786.
- [35] M. Kalashnikov, W. Choi, C.-C. Yu, Y. Sung, R. R. Dasari, K. Badizadegan, M. S. Feld, *Opt. Express* **2009**, *17*, 19674.
- [36] M. A. Yurkin, A. G. Hoekstra, *J. Quant. Spectrosc. Radiat. Transfer* **2011**, *112*, 2234.
- [37] W. Wang, Y. Wen, J. Q. Lu, L. Zhao, S. A. Al-Qaysi, X.-H. Hu, *J. Quant. Spectrosc. Radiat. Transf.* **2019**, *224*, 453.
- [38] P. Willett, *Inf. Process. Manag.* **1988**, *24*, 577.
- [39] H. Permuter, J. Francos, I. Jermyn, *Pattern Recognit.* **2006**, *39*, 695.
- [40] N. Shental, A. Bar-Hillel, T. Hertz, D. Weinshall, *Proc. Neural Inf. Process. Sys.* **2003**, *16*, 465.
- [41] Y. Tan, Y. Xia, J. Wang, in *Proc. IEEE-INNS-ENNS Int. Joint Conf. Neural Netw. (IJCNN)*, IEEE, Como, Italy **2000**, pp. 411–416.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Liu J, Xu Y, Wang W, et al. Machine learning of diffraction image patterns for accurate classification of cells modeled with different nuclear sizes. *J. Biophotonics*. 2020; e202000036. <https://doi.org/10.1002/jbio.202000036>