PROFILING EFFUSION CELLS BY QUANTITATIVE ANALYSIS OF MORPHOLOGY AND DIFFRACTION IMAGING PATTERNS

by

Safaa Al-Qaysi

January, 2019

Director of Dissertation: Dr. Xin-Hua Hu

Major Department: Physics

Abstract

Profiling cells in human pleural and peritoneal effusion (PPE) samples is an essential task for cytological diagnosis of cancers and patient management. Conventional PPE cytology of a PPE sample is labor intensive and its efficacy depends heavily on the experience of trained specialists in addition to low sensitivity. In this dissertation research project, we focused our effort on application of a label-free method of polarization diffraction imaging flow cytometry (p-DIFC) for quantitative profiling of cell morphology in PPE samples and their correlations to the texture features of cross-polarized diffraction image (p-DI) pairs. To establish the morphology implications of the measured (p-DI) pairs, the 3D structures of PPE cells were measured by using confocal microscopy and quantified with 27 parameters for characterization and analysis of the cellular structures by the conventional fluorescent imaging method. Furthermore, realistic optical cell models (OCM) have been developed as virtual PPE cells and used for accurate simulation of diffraction imaging process to obtain calculated p-DI pairs. This approach allows us to correlate p-DI texture feature parameters quantified by the gray-level co-occurrence matrix (GLCM) algorithm and 3D morphology parameters and investigate various approaches of morphology based cell classification. Clustering algorithms of hierarchical clustering (HC) and Gaussian mixture model (GMM) have been investigated to develop a robust classification method for profiling of the PPE

cells' morphological features by the (GLCM) parameters of p-DI pairs. Correlations between the morphological feature parameters and p-DI feature parameters of the imaged PPE cells have been analyzed to gain insights on the morphology implications of image texture patterns and GLCM parameters of the measured p-DI pair data acquired from live and unstained PPE cells of patients of lung and ovarian cancers. Through this dissertation study we have utilized and developed a suite of image processing and analysis tools and obtained results that demonstrate the strong capability of the p-DIFC method to yield big data for profiling PPE cells acquired from cancer patients and the potential to detect malignant PPE cells in the future.

PROFILING EFFUSION CELLS BY QUANTITATIVE ANALYSIS OF MORPHOLOGY AND DIFFRACTION IMAGING PATTERNS

A Dissertation

Presented to the Faculty of the Department of Physics

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy in Biomedical Physics

by

Safaa Al-Qaysi

January, 2019

PROFILING EFFUSION CELLS BY QUANTITATIVE ANALYSIS OF MORPHOLOGY AND DIFFRACTION IMAGING PATTERNS

by

Safaa Al-Qaysi

APPROVED BY:

DIRECTOR OF
DISSERTATION:———————————————————————————————

Xin-Hua Hu, Ph.D.

CO-DIRECTOR
OF DISSERTATION:———————————————————————————————-

Yuanming Feng, Ph.D.

COMMITTEE MEMBER:———————————————————————————————-

Michael Dingfelder, Ph.D.

COMMITTEE MEMBER:———————————————————————————————-

Heng Hong, M.D. Ph.D.

COMMITTEE MEMBER:———————————————————————————————-

Jefferson Shinpaugh, Ph.D.

CHAIR OF THE DEPARTMENT
OF PHYSICS:———————————————————————————————

Jefferson Shinpaugh, Ph.D.

DEAN OF THE
GRADUATE SCHOOL:———————————————————————————————

Paul J. Gemperline, PhD

DEDICATION

I dedicate this dissertation:

To my father soul and to my mother

To my brothers

To my wife and to my son.

ACKNOWLEDGEMENTS

I want to express my sincere thanks and deep gratitude to my advisor Dr. Xin-Hua Hu for suggesting and advising the present work and for his continued support, guidance, and encouragement throughout the research. I want to thank my co-advisor Dr. Yuanming Feng and the advisory committee members Dr. Michael Dingfelder, Dr. Heng Hong, and Dr. Jefferson Shinpaugh for their valuable cooperation during this study. I am thankful for Dr. Heng Hong for his help on providing the effusion samples and the pathological information related to them. I thank Dr. Diana Dai for her helps with cytopathological slides images, and with identifying different types of cells within the samples. I have a deep thanks to Dr. Li Yang for opening his lab's door for me to do the cells preparation and confocal imaging. I am very thankful to Dr. Douglas Weidener for his help on the confocal microscope. I am grateful Dr. Kenneth Jacobs for his help with cross-polarized diffraction imaging flow cytometer. I am thankful to Mrs. Marion Greene for her support in my first steps in this research. Also, I would like to thank Dr. Wen for sharing his codes of 3D reconstruction and clustering. I thank Dr. Wang for sharing his codes on the optical cell model and diffraction image calculation. I thank Mr. John Jones for his assistance with ECU high-performance computer cluster.

I like to express my profound gratitude to Dr. Michael Dingfelder, the assistant chair of graduate studies, for his valuable help and support from my first day as a Ph.D. student till this moment. I would like to appreciate the chair and the family of the Physics department, faculty, staff, and graduate students, for their valuable collaboration.

In addition, I want to record my affection and thanks to my family and friends for their moral support and patients throughout my study. Last but not least, I would like to express my heartfelt thank and appreciation to my lovely wife for her unlimited help and support to me to achieve the success.

Contents

List of Tables

List of Figures

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
ADDA	Amsterdam discrete dipole approximation
AIC	Akaike information criterion
AJCC	American joint commission on cancer
BIC	Bayes information criterion
BLL	Biomedical laser laboratory
BSA	Bovine serum albumin
CC	Cancer cell
CCD	Charge coupled device
CIMA	Cell image and morphology analysis
CT	Computed tomography
DDA	Discrete dipole approximation
DI	Diffraction image
EM	Electromagnetic
EMax	Expectation maximization
FCM	Flow cytometry
FDTD	Finite-difference time-domain

FOV	Field of view
FS	Forward scatter
GLCM	Gray-level co-occurrence matrix
GMM	Gaussian mixture model
HC	Hierarchical clustering
IFC	Imaging flow cytometer
LSM	Laser scan microscope
ME	Malignant effusion
MLPE	Maximum-likelihood parameter estimation
MRI	Magnetic resonance imaging
NC	Normal cell
NSCLC	Non-small cell lung cancer
OCM	Optical cell model
p-DI	Cross-polarized diffraction image
p-DIFC	Polarization diffraction imaging flow cytometer
PBS	Phosphate buffer saline
PDF	Probability density function
PET	Positron emission tomography
PMT	Photo-multiplier tube
PPE	Pleural and peritoneal effusion
RBC	Red blood cell
RI	Refractive index
ROI	Region of interest
SCLC	Small cell lung cancer

SS Side scatter

STD Standard deviation

TDI Time delay and integration

Chapter 1 Introduction

Manual evaluation of pleural and peritoneal effusions (PPEs) is an essential part of cytological diagnosis of cancers. Traditionally, PPE samples are prepared in the form of smear, cytospin, and liquid-based-cytology slides for evaluation by cytopathologists. Even with immunofluorescence staining, PPE cytology examination of PPE samples has its efficacy depending heavily on the experience of specialists, is labor intensive and often of low sensitivity. Malignant PPEs are defined as the presence of neoplastic cells [1, 2] and their detection presents a challenging problem in patients with lung, breast and ovarian cancers due to the diversity in cell morphology and the complexity of the staining process. For malignant pleural effusions alone, the annual incidence is around 175,000 in the United States, and they are present in 30% of lung cancer patients [3] and 15% of patients who die of cancers [4]. Existing approaches for diagnosis of malignancies in PPEs suffer from low sensitivity averaging at only about 58% despite its high specificity that often reaches above 95% [5, 6].

The long-term goal of the dissertation research described here is to develop a label-free method of diffraction imaging for quantitative profiling of PPE cells and detection of malignant ones. For this dissertation, our research is focused on the quantitative profiling of cell morphology in PPEs and their correlations to the texture features of diffraction images acquired by a method of polarization diffraction imaging flow cytometry (p-DIFC). In addition to acquiring the diffraction images, the 3D morphology of PPE cells was quantitatively measured by using confocal microscopy for characterization and comparison of the cellular structures in terms of the cytoplasm, nucleus, and mitochondria. A method of realistic optical cell model (OCM) for developing virtual PPE cells was used for accurate simulation of diffraction imaging process to obtain calculated cross-polarized

diffraction image (p-DI) pairs. This allows us to correlate p-DI feature parameters and 3D morphology feature parameters and investigate morphology based cell classification. The hierarchical clustering (HC) and Gaussian mixture model (GMM) algorithms have been investigated to develop a clustering method for study of cell classification by the morphological feature parameters and p-DI feature parameters. Correlations between the morphological feature parameters and p-DI feature parameters of the imaged PPE cells have been analyzed to gain insights on the image parameters of p-DI pair data. In the last part of this dissertation research, we have acquired p-DI data from live and unstained PPE cells extracted from patients of lung and ovarian cancers and applied the clustering method developed for analyzing the calculated p-DI data to profile and analyze the PPE cells by their texture patterns.

Through this project, we have acquired more than 449 confocal image stacks of PPE cells extracted from 12 patients suspected of cancers. The 3D morphological features of these cells have been quantified by 27 parameters obtained through reconstruction to investigate and compare their morphological differences in cytoplasm, nucleus, and mitochondria that are important to light scattering. We employed an image analysis software to extract 2D morphology of 560 PPE cells from cytopathological image slides of 5 patients that include immunochemically labeled normal and cancer cells for comparison to the confocal imaging results. Realistic optical cell models OCMs have been constructed based on 449 PPE cell structures to obtain calculated p-DI pairs and examined the effect of intracellular morphology and refractive index heterogeneity on the diffraction patterns of calculated p-DI against the measured data. An experimental p-DIFC system was used to acquire p-DI pairs of PPE cells from 12 patients with three incident beam polarizations. With the insight obtained from the analysis of 3D morphology and calculated p-DI data, we have applied the clustering methods based on the HC and GMM algorithms to classify the measured p-DI pairs and corresponding PPE cells into three groups. The results of this study demonstrate that the p-DIFC method has the capacity to yield big data for profiling PPE cells acquired from cancer patients and the potential to detect malignant PPE cells in the future.

This dissertation is organized as follows: Chapter 2 provides a background overview of the

theory of light scattering, cell morphology and its connection to light scattering theory through optical cell modeling, light scattering simulation and diffraction image flow cytometry, and effusion cell evaluation. Chapter 3 describes the experimental methods for cell extraction, preparation, confocal imaging acquisition, 3D reconstruction, 2D feature extraction, and the machine learning tools for cell clustering. Chapter 4 presents the methods of optical cell model construction, light scattering simulation and simulation of diffraction imaging process based on a ray-tracing approach to obtain calculated p-DI data. It also includes the analysis of texture patterns in the p-DI data for clustering and correlation study with the morphology. Chapter 5 reports the results of p-DIFC measurement and analysis of acquired p-DI data from PPE cells based on the results of Chapters 3 and 4. Finally, Chapter 6 summarizes the results of the dissertation study and suggests future research directions.

Chapter 2 Background

This chapter provides an overview of light scattering by biological cells and cytology diagnosis of effusion cell which are to be profiled in this dissertation research by confocal and diffraction imaging. We will discuss the fundamental theory for modeling of light scattering and simulation of diffraction images that can be measured in our study by a flow cytometry method. The conventional approach of imaging flow cytometry will also be reviewed for comparison with our diffraction imaging approach.

2.1 Theory of light scattering

Light scattering is defined as the deviation of propagation for light wavefields from its incident direction due to the interaction with matter of heterogeneous optical property in terms of refractive index (RI). The incident light wavefields excite the molecules of a sample such as a particle and induce electric dipoles in the form of polarization vector, among other effects, resulting in a scattered light. The induced electric dipoles emit or re-radiate light wavefields as scattered light dominated by the elastic component that is of the same frequency as the incident light. Furthermore, the scattered light propagates in different directions from the illuminated particle or scatterer and thus form angular distributions, which for coherent excitation exhibit characteristic patterns and correlate highly with the morphology of the particle in terms of RI [7].

For particles of sizes much smaller than the wavelength of incident light, the intensity of the scattered light is inversely proportional to the fourth power of the wavelength as accounted for by the Rayleigh model [8]. Thus, incident light wavefields of shorter wavelengths have a much higher

probability of being scattered than those of long wavelengths. On the other hand, if the scatterers have sizes on scales same or larger than the wavelength, the scattered light tends to become stronger in the forward scattering directions as can be shown by the Mie theory [7]. Furthermore, as the degree of symmetry for the internal structure of the scatterer decreases, the angular distribution of scattered light exhibits increasingly complex patterns that are difficult to quantify but can be learned with powerful image processing tools. The angle-resolved or spatial distribution of the scattered light can, therefore, provide morphological information associated with the RI distribution inside a scatterer and also on the molecular composition. When a particle is illuminated by highly coherent light from a laser, the scattered light wavefields from different molecules inside the particle are coherent and present characteristic diffraction patterns [9].

The strong correlation between the wavelength-scaled scatterer's 3D morphology and patterns of scattered light wavefields allows the determination of morphological features through diffraction imaging. To quantify the angular distribution of scattered light, one can define the polar angle θ_{sca} and azimuthal scattering angle ϕ_{sca} relative to the incident direction which is usually set as the z-axis. For scatterers of complex morphology, measurement and analysis of the spatial distributions of its coherent scattered light yield powerful tools due to the highly correlated relations between the RI distribution of the scatter and diffraction patterns. This correlation could thus provide fundamental morphology information of the scatterer [10].

The intensity and angular distribution of light scattered from a single particle depends on the characteristics of incident light wavefields and the scattering particle properties (shape, size, and the material) [11]. Figure 2.1 shows light scattering by a particle of arbitrary shape. The direction of scattering ($\hat{e}_{sca}$) is measured from the direction of propagation of the incident light ($\hat{e}_{inc}$) by the polar angle θ_{sca} and azimuth angle ϕ_{sca} with the unit vectors of $\hat{e}_{sca}$ and $\hat{e}_z$ forming the scattering plane. In addition, the polarization of scattered light propagating in a direction can be quantified by its s-polarized and p-polarized components defined as perpendicular and parallel polarized light against the scattering plane [7].

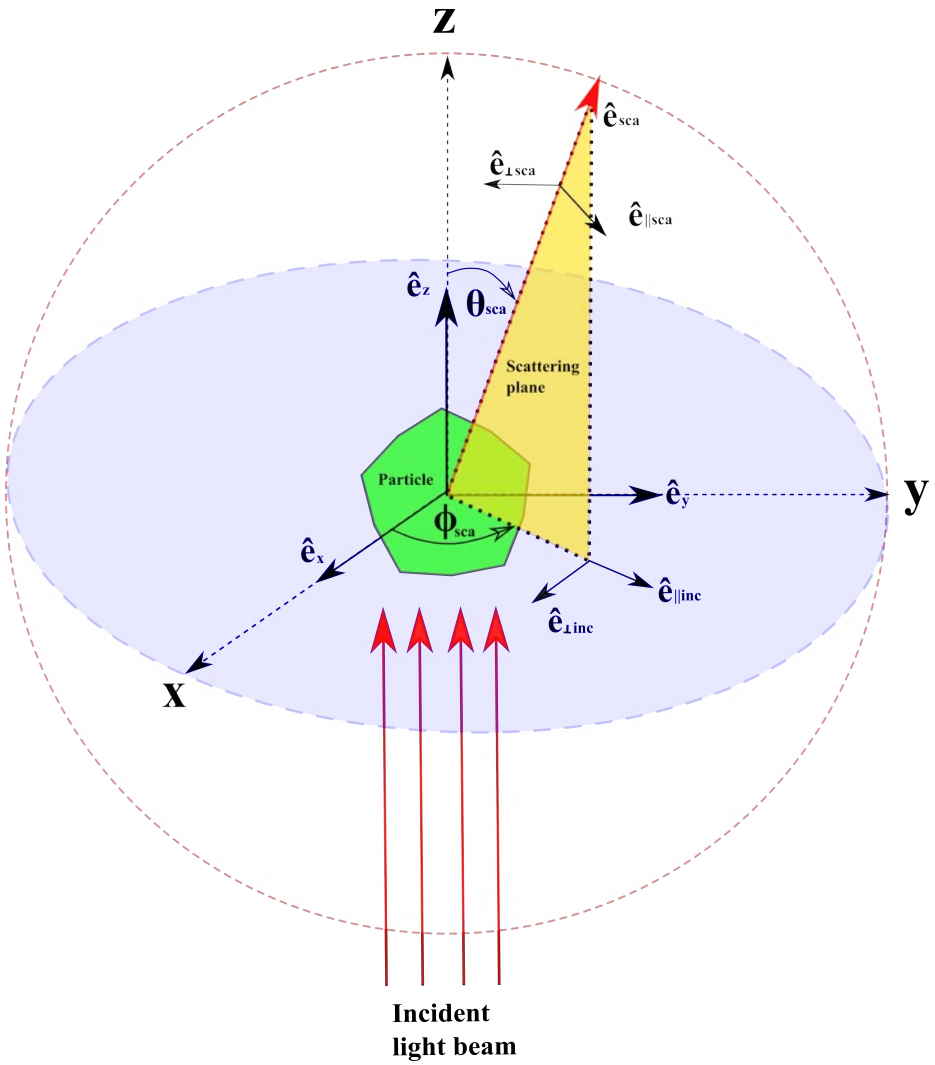

Figure 2.1 Scattering problem

Any incident light wavefields can be represented by sums of plane-wave electromagnetic (EM) components of different propagation directions and frequencies using the Fourier transform technique. So we can consider a simple case of incident light as a harmonically oscillating plane-wave that propagates in vacuum, which is described by its electric field and magnetic field given by [7]:

$$\begin{aligned} \vec{E}_{inc}(\vec{r},t) &= \vec{E}_0 e^{(i\vec{k}\cdot\vec{r}-i\omega t)} \\ \vec{H}_{inc}(\vec{r},t) &= \vec{H}_0 e^{(i\vec{k}\cdot\vec{r}-i\omega t)} \end{aligned} \tag{2.1}$$

where $\vec{E}_{inc}$ and $\vec{H}_{inc}$ are the incident electric field and magnetic field, respectively, $\vec{E}_0$ and $\vec{H}_0$ are amplitudes, $k = \frac{2\pi}{\lambda}$ is the wavenumber in vacuum, λ is the wavelength of the incident light in

vacuum, ω is the angular frequency, and t is the time. The incident light propagates undisturbed until it enters into a region containing an object with a RI different from that of the host medium leading to scattering of the incident wavefields. The EM fields scattered by the object, ($\vec{E}_{sca}$ and $\vec{H}_{sca}$), is the difference between the total field in the presence of the object, $\vec{E}$ and $\vec{H}$ and the incident field, $\vec{E}_{inc}$ and $\vec{H}_{inc}$ that would exist in the absence of the object [7]:

$$\begin{aligned} \vec{E}_{sca} &= \vec{E} - \vec{E}_{inc} \\ \vec{H}_{sca} &= \vec{H} - \vec{H}_{inc} \end{aligned} \tag{2.2}$$

The total fields must satisfy Maxwell's equations [12]:

$$\begin{aligned} \nabla \cdot \vec{D} &= 0 \\ \nabla \cdot \vec{B} &= 0 \\ \nabla \times \vec{E} &= -\frac{\partial \vec{B}}{\partial t} \\ \nabla \times \vec{H} &= -\frac{\partial \vec{D}}{\partial t} \end{aligned} \tag{2.3}$$

where $\vec{D} = \epsilon(\vec{r})\vec{E}$ is the electric displacement, $\epsilon(\vec{r})$ is the electric permittivity with spatial variation, $\vec{B} = \mu_\circ \vec{E}$ is the magnetic induction, and $\mu_\circ$ is the magnetic permeability of free space. From Maxwell's equations, it is straightforward to obtain the Helmholtz wave equation for $\vec{E}$ after removing the harmonic time dependence:

$$\begin{aligned} -4\pi \vec{f}(\vec{r},\omega) &= \nabla^2 \vec{E}(\vec{r},\omega) - k^2 \vec{E}(\vec{r},\omega) \\ \vec{f}(\vec{r},\omega) &= \frac{1}{4\pi}\left[k^2\left(\frac{\epsilon(\vec{r})}{\epsilon_\circ} - 1\right)\vec{E}(\vec{r},\omega) + \nabla\left(\frac{\vec{E}(\vec{r},\omega)}{\epsilon(\vec{r})} \cdot \nabla\epsilon(\vec{r})\right)\right] \end{aligned} \tag{2.4}$$

where $k^2 = \omega^2/c^2 = \mu_\circ \omega^2 \epsilon_\circ$, c is the speed of light and $\epsilon_\circ$ is the permittivity of free space. The solution for $\vec{E}$ has the form [13]:

$$\begin{aligned}\vec{E}(\vec{r},\omega) &= \vec{E}_{inc}(\vec{r},\omega) + \int_V \vec{f}(\vec{r}',\omega)G(\vec{r},\vec{r}',\omega)d^3\vec{r}' \\ G(\vec{r},\vec{r}',\omega) &= \frac{\exp(ik|\vec{r}-\vec{r}'|)}{|\vec{r}-\vec{r}'|}\end{aligned} \tag{2.5}$$

where V is the volume of the region containing by the scatterer and G is the spherical Green's function. The second term is the scattered field, and in the far-field approximation it becomes [13]:

$$\vec{E}_{sca}(\vec{r},\omega) = \frac{\exp(ikr)}{\vec{r}} \int_V \vec{f}(\vec{r}',\omega)\exp(-ikr\vec{r}')d^3\vec{r}' \tag{2.6}$$

In the equation above, the function $\vec{f}(\vec{r}',\omega)$ is integrated over the volume of the scatterer. Since $\vec{f}(\vec{r}',\omega)$ is a function of electric field according to equation 2.4, it is apparent that the internal field must be known in order to solve for the scattered field. The scattered field depends not only on the intensity but also on the polarization of the incident field. The incident and scattered electric fields can be expressed as linearly polarized waves with components parallel ($E_\parallel$) and perpendicular ($E_\perp$) to the scattering plane [12]:

$$\begin{aligned}\vec{E}_{inc} &= E_{\parallel inc}\hat{e}_{\parallel inc} + E_{\perp inc}\hat{e}_{\perp inc} \\ \vec{E}_{sca} &= E_{\parallel sca}\hat{e}_{\parallel sca} + E_{\perp sca}\hat{e}_{\perp sca}\end{aligned} \tag{2.7}$$

where, $\hat{e}_\parallel$ and $\hat{e}_\perp$ are parallel and perpendicular basis vectors for the scattering and incident planes. As shown in equation 2.7, the amplitude of the scattered field and the amplitude of the incident field are linearly related. This relation can be expressed in matrix form [7]:

$$\begin{pmatrix} E_{\parallel sca} \\ E_{\perp sca} \end{pmatrix} = \frac{\exp ik(r-z)}{-ikr} \begin{pmatrix} S_2 & S_3 \\ S_4 & S_1 \end{pmatrix} \begin{pmatrix} E_{\parallel inc} \\ E_{\perp inc} \end{pmatrix} \tag{2.8}$$

where S_η ($\eta = 1,2,3,4$) is the amplitude scattering matrix. the elements of this matrix depend on the scattering angle θ, and the azimuthal angle ϕ.

Furthermore, the incident and scattered waves can also be described with the real-valued Stokes parameters, which can be compared directly to the measured light intensities. These parameters

are I, Q, U and V, where I is the total intensity (flow of energy per unit area) of radiation, Q is polarization at 0° or 90° to the scattering plane, U is polarization at ±45° to the scattering plane, and V is left or right circular polarization [14]. The Stokes parameters of light scattered a particle are related to the electric field vectors in terms of time averages [12]:

$$\begin{aligned} I &= \langle E_{\parallel sca} E^*_{\parallel sca} + E_{\perp sca} E^*_{\perp sca} \rangle \\ Q &= \langle E_{\parallel sca} E^*_{\parallel sca} - E_{\perp sca} E^*_{\perp sca} \rangle \\ U &= \langle E_{\parallel sca} E^*_{\perp sca} + E_{\perp sca} E^*_{\parallel sca} \rangle \\ V &= i\langle E_{\parallel sca} E^*_{\perp sca} - E_{\perp sca} E^*_{\parallel sca} \rangle \end{aligned} \tag{2.9}$$

The incident Stokes vector is transformed by the scattering matrix to the scattered vector as shown in the following relation [7]:

$$\begin{pmatrix} I_{sca} \\ Q_{sca} \\ U_{sca} \\ V_{sca} \end{pmatrix} = \frac{1}{k^2 r^2} \begin{pmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{13} & S_{23} & S_{33} & S_{34} \\ S_{14} & S_{24} & S_{34} & S_{44} \end{pmatrix} \begin{pmatrix} I_{inc} \\ Q_{inc} \\ U_{inc} \\ V_{inc} \end{pmatrix} \tag{2.10}$$

where *inc* and *sca* stand for the incident and the scattered light, respectively, k is the wavenumber, r is the distance from the scatterer, S_{ij} , the elements of the Mueller matrix, are a function of the scattering angle θ [7].

Since the light intensity is directly related to the real elements of Mueller matrix, the polarization properties of a particle can be characterized by its Mueller matrix elements. One can measure the Mueller matrix elements by illuminating a particle and analyzing the scattered light with various combinations of polarizers [15]. For the incident beam, the linear polarizer can have a horizontal, vertical, or 45° orientation. For each of these three options, the scattered beam can have polarizations parallel or perpendicular to the scattering plane. The components of the Mueller matrix are related to the polarization states of the incident and scattered light by equation 2.11. Where the subscripts of I on the numerator indicate the polarization of the incident and scattered

light, respectively. All the incident and scattered intensities can be measured experimentally to determine Mueller matrix elements and the polarization properties of the scatterer [16].

$$\begin{aligned}
\frac{I_{\parallel\parallel}}{I_{inc}} &= \frac{1}{2}((S_{11}+S_{12})+(S_{21}+S_{22})) \\
\frac{I_{\parallel\perp}}{I_{inc}} &= \frac{1}{2}((S_{11}+S_{12})-(S_{21}+S_{22})) \\
\frac{I_{\perp\parallel}}{I_{inc}} &= \frac{1}{2}((S_{11}-S_{12})-(S_{22}-S_{21})) \\
\frac{I_{\perp\perp}}{I_{inc}} &= \frac{1}{2}((S_{11}-S_{12})+(S_{22}-S_{21})) \\
\frac{I_{45^\circ\parallel}}{I_{inc}} &= \frac{1}{2}((S_{11}+S_{21})+(S_{13}+S_{23})) \\
\frac{I_{45^\circ\parallel}}{I_{inc}} &= \frac{1}{2}((S_{11}-S_{21})+(S_{13}+S_{23}))
\end{aligned} \tag{2.11}$$

The spatial intensity distribution of scattered light at a particular point in space can be simulated by solving Maxwell's equations for the scattering problem within the scatter field [7]. Mie theory and extended Mie models offer closed-form solutions of wave equations for a selective group of highly symmetric single scatterers like spheres and cylinders by expanding the fields with series of spherical wave functions. For scatterer of non-spherical shapes, multiple numerical methods including T-matrix, Finite-Difference Time-Domain (FDTD) and Discrete Dipole Approximation (DDA) methods provide practical means to solve Maxwell's equations or wave equations for heterogeneous scatterers. The method used for this dissertation study is DDA that has open-source code available for research and is designed with high flexibility regarding of the importing arbitrary RI distributions for different scatterers of complex morphology as found in biological cells [17].

2.2 Discrete dipole approximation

The DDA algorithm provides a powerful method to simulate light scattering by a finite array of polarizable cubic volumes representing the scatterer's RI distribution. The DDA method divide a dielectric scatterer into N small cubic volumes named dipoles and calculate dipole moments due the incident electric field as well as the electric field by other dipoles. The dipole size need to be

small enough in comparison to both of the scatterer sizes and the incident wavelength for sufficient accuracy, which can be quantified by the parameter "dipoles per wavelength" (dpl) [18]:

$$dpl = \frac{\lambda}{d} \tag{2.12}$$

where d is the elementary cube side length (also called as dipole size) and λ is the wavelength of incident light in the medium hosting the scatterer (water in our study). If the electric field at the site of i^{th} dipole is $\vec{E}_i$, then the polarization vector $\vec{P}_i$ quantifying the oscillating dipole moment distribution is given by [19]:

$$\vec{P}_i = \alpha_i \vec{E}_i \tag{2.13}$$

where α_i is the complex dipole polarizability determined from the input data of RI distribution for the scatterer. Purcell and Pennypacker used the Clausius-Mossotti polarizabilities [19]:

$$\alpha_i^{CM} = V_d \frac{3}{4\pi} \frac{\epsilon_i - 1}{\epsilon_i + 2} \tag{2.14}$$

where $V_d = d^3$ is the volume of a cubic dipole and ϵ_i is the dielectric constant at $\vec{r}_i$ location. However, this formula is exact for an infinte cubic volume in the limit $kd \rightarrow 0$. Various radiative correction to the Clausius-Mossotti polarizability have been suggested [20].

Each dipole moment produces an electric field at the location $\vec{r}_i$ that can be obtained as the difference between the local incident field and the field due to the other $(N-1)$ dipoles [20]:

$$\vec{E}_i = \vec{E}_{inc,i} - \sum_{j \neq i} \vec{A}_{ij} \vec{P}_j \tag{2.15}$$

where $\vec{A}_{ij}\vec{P}_j$ is the electric field at $\vec{r}_i$ that is due to dipole $\vec{P}_j$ at location $\vec{r}_j$, including retardation effects. The 3×3 matrix $\vec{A}_{ij}$ can be written as [20]:

$$\vec{A}_{ij} = \frac{\exp(ikr_{ij})}{r_{ij}} \times \left[k^2(\hat{r}_{ij}\hat{r}_{ij} - I_3) + \frac{ikr_{ij} - 1}{r_{ij}^2}(3\hat{r}_{ij}\hat{r}_{ij} - I_3) \right], i \neq j \tag{2.16}$$

where $\vec{r}_{ij}$ is a displacement vector pointing from $\vec{r}_j$ to $\vec{r}_i$, $\vec{r}_{ij} = |\vec{r}_i - \vec{r}_j|$, $\hat{r}_{ij} = (r_i - r_j)/r_{ij}$, and I_3 is 3×3 unit matrix. By defining $\vec{A}_{ii} = \alpha_i^{-1}$, the scattering problem reduces to finding the polarizations $\vec{P}_j$ that satisfy a system of $3N$ complex linear equations [20]:

$$\vec{E}_{inc,i} = \sum_{i=1}^{N} \vec{A}_{ij}\vec{P}_j \tag{2.17}$$

Once equation 2.17 has been solved for the unknown polarizations $\vec{P}_j$, the extinction and absorption cross sections C_{ext} and C_{abs} are determined directly [21]:

$$C_{abs} = \frac{4\pi k}{Re(m_{in})} \sum Im(\vec{P}_i \vec{E}^*_{inc,i}) \tag{2.18}$$

$$C_{ext} = \frac{4\pi k}{Re(m_{in})} \sum \left[Im(\vec{P}_i \vec{E}^*_{inc,i}) - (2/3)k^3 |\vec{P}_i|^2 \right] \tag{2.19}$$

where m_{in} is the RI of the incoming (host) medium. The absorption cross section C_{abs} represents the area of the energy absorbed by the particle, and the extinction cross section C_{ext} corresponds to the energy removed from the original beam. By conservation of energy, the scattering cross section C_{sca} is [9]:

$$C_{sca} = C_{ext} - C_{abs} \tag{2.20}$$

In this study, we used the open source, parallel computing Amsterdam Discrete Dipole Ap-

proximation (ADDA) implementation of DDA method for all light scattering simulation of small size spherical particles and biological cells. The ADDA is developed in C language by Yurkin and Hoekstra. it is highly portable and It provides direct control over scattering geometry, orientation and incident beam. Compare to other DDA codes, the ADDA has many advantages such as the ability to simulate heterogeneous particles, flexibility regards different scatterer geometry, ability to run on multiple processors in parallel and ease of use with well written documentations and publications [22].

2.3 Cell modeling and light scattering simulations

Hence, biological cells are microscopic and optically heterogeneous particles that composed of different intracellular organelles such as nucleus, mitochondria, and lysosomes embedded in cytoplasm and surrounded by cytoplasm membrane. These organelles scatter light wavefields due to their RI heterogeneity with complex angular distributions resulted from the highly non-spherical shapes inside the cells. The complex patterns of scattered light is a consequence of interference among the coherent light wavefields at the plane of an image sensor by different intracellular organelles. The distribution pattern of the scattered light from biological cells contains a lot of information regarding the cell internal structure and its optical properties, but there are no simple relations can derived on the relationship between the individual speckles and particular component of the cell [16, 23].

Several cells models have been reported to investigate the RI of intracellular organelles and to understand RI distribution within each organelle. These models were improved over the years by comparing to experimental measurement results. Early models treated biological cells as homogeneous or concentric spheres, and these models showed that the side scatter is mainly influenced by the complexity of the cell's internal structure, while the forward scatter is more affected by the size of the cell [24, 25]. Improved models treat cells as spheres or ellipsoids containing off-center and smaller spheres and ellipsoids to quantify light-cell interaction and

understand light scattering by cells. Additionally, a multilayer model of cell containing nucleus and mitochondria has been proposed [26–29].

More recent light scattering studies have used cell models with higher level of complexity. Yurkin et al. have modeled the red blood cells as a biconcave disks to investigate the effect of changes in cell orientation related to the direction of the light incident upon, volume and diameter. They showed a good agreement between the theoretical and experimental analysis [30]. Brock et al. used more realistic technique to construct cell models composed of a cell membrane and nucleus using confocal microscope images acquired from B-cells. Results showed how the details of cell structure affect light scattering properties [10]. Zhang et al. established a realistic optical cell models based on fluorescence confocal imaging of major organelles to generate virtual cells for accurate diffraction image simulation. They showed that the simulated images and extracted parameters can be used to distinguish virtual cells of different nuclear volumes and refractive indices against the orientation variation [28]. Wang et al. used four improved optical cell model to evaluate the effects of internal cell structure and the refractive index distribution of two prostate cell types on simulated diffraction images [31].

The effects of the RI heterogeneity in biological cells have also been examined against experimental results. Some angular depend light scattering studies used the azimuthal angular distribution of the coherent scattered light intensity as a method to profile 3D RI distribution. Such studies offered limited information regarding the cell size and the internal structure of the cell [15, 32–34]. While other studies used interferogram based imaging method to determine RI distribution inside the cell through phase extractions followed by tomographic reconstruction of cell morphology. The tomographic reconstruction depends on computationally expensive filtered back projection algorithm [35, 36].

Since morphological changes in cells affect how light scatters from the cells, accurate modeling of biological cells is important for light scattering simulations. In this dissertation research we used a method developed by Biomedical Laser Laboratory (BLL) to create a realistic 3D model of biological cells for accurate and practical light scattering simulations. The proposed model in this

study is based on the structural information of major organelles extracted from stacks of confocal microscopy images of PPE cells. These images contain the 3D distribution of light intensity of fluorescent dye molecules stained to nucleus, mitochondria and cytoplasm. Then, the model convert the fluorescent light intensity to RI. In addition to that we normally distribute small artificial Gaussian spheres inside the cell cytoplasm to simulate the role of lysosomes in light scattering.

2.4 Flow cytometry

Flow cytometry (FCM) is an advanced method of rapid single cell assay by measuring and analyzing the scattering and fluorescence signals of stained cells or particles, such as intracellular organelles and microorganisms. It is a widely used technique in life science research and clinical applications for investigating cell functions and cancer diagnostics. The FCM method has been developed into an effective multipurpose technique over the last 70 years. The first automatic single cell measurement was done on a red blood cells counting in the 1950's by Wallace H. Coulter [37]. Following the development of fluorescence-based cytometers during the 1960's and 1970's, many types of fluorescent stains were discovered in the late 1980's and became commercially available during 1990's. These stains increased significantly the capability of FCM to measure more functional parameters of cells. To date, the flow cytometer becomes the tool of choice for rapid assay of single cells with the power of detecting multiple parameters simultaneously such as size, granularity by the light scatter signals and various molecular information by fluorescent signals from the cells. These parameters are used to analyze and differentiate many types of cells [38].

Light scattering and fluorescence emission are the primary means of cell probing for conventional FCM [39]. The main components of conventional flow cytometers consist of fluidics system, light source, optical system, light detectors and data analysis system as shown in Figure 2.2.

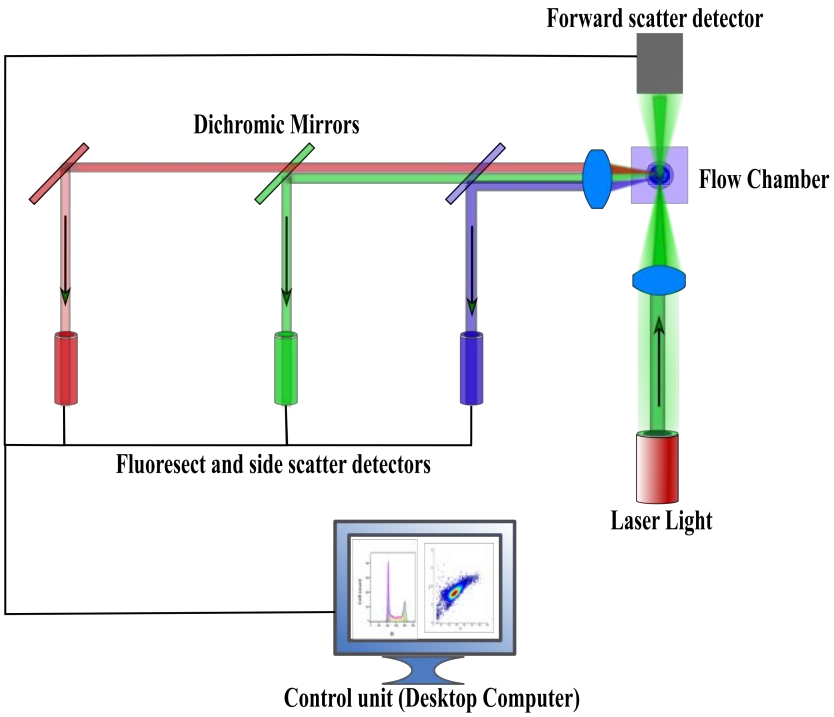

Figure 2.2 Schematic diagram of a conventional FCM. The conventional FCM system was used to obtain a statistical histogram measurements for biological cells based on the scattering light and fluorescence stains.

The fluidic unit is responsible for directing suspended cells in single file through the focus of an incident laser beam placed at the center of the flow chamber and controlling the speed of cells carried by the core fluid. The optical unit is used to focus one or multiple laser beams into the flow chamber for excitation of cells as they flow through the chamber while detectors such as photodiodes or photomultiplier tubes convert the scattered light or fluorescent light into electronic signal to be acquired with multichannel analyzer. The data analysis unit in a host computer acquires these signals for subsequent analysis using histograms and/or scatter plots for cell classification [39, 40]. One can roughly estimate that the forward scattered light (FS) intensity is approximately proportional to the volume of a measured cell while the side scattered light (SS) intensity depends on the gradient of RI within the cell and thus relate to its morphological properties. The fluorescence light intensities at varying wavelengths are derived from various fluorescent stains that bind to a specific cell components or molecules after exciting by laser light.

In order to perform single cell analysis with morphological information, a new type of flow

cytometer was commercially introduced over the last decade merging the power of traditional FCM design with fluorescence microscopy. This type of hybrid instrument called Imaging Flow Cytometry or sometimes called Multispectral Imaging Flow Cytometry. Figure 2.3 shows a simple schematic draw for the imaging flow cytometer system. The system can be used to track the moving cells and acquire multiple images of each cell in different modes such as the forward and side scatter images, along with several fluorescence images of different fluorescence spectral bands that label different cell's components. A powerful analytic software developed and integrated with the imaging flow cytometry to handle the massive amount of images data [41]. It should be noted that this method of imaging flow cytometry acquires mainly the non-coherent light of fluorescence which exhibit no diffraction patterns. In the case of imaging scattered light, it takes no advantage of the image analysis of diffraction patterns for its use of imaging unit at conjugate locations and resulted poor image contrast.

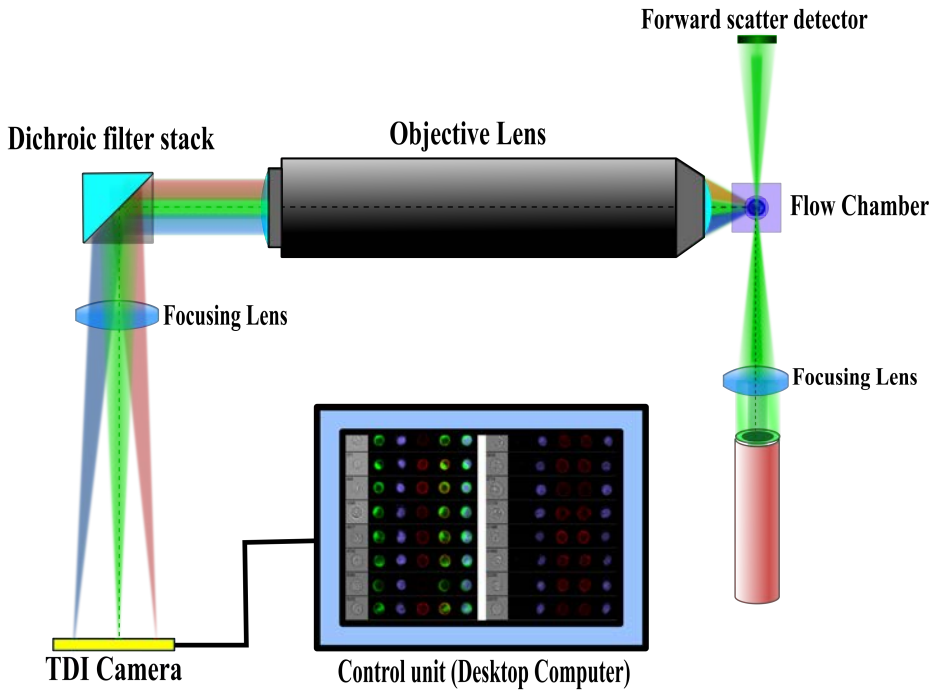

Figure 2.3 Simple 2-D representation diagram of Imaging flow cytometer (IFC). The IFC system was used to obtain a different spectral imaging measurements for biological cells based on the scattering light and fluorescence stains.

Over the past decade, the research group BLL at East Carolina University has developed a method of polarization diffraction imaging flow cytometry (p-DIFC) to acquire a cross-polarized image pair of the coherent light scattered by single cells as high-contrast images excited by a linearly polarized laser beam [42–44]. The p-DIFC method can simultaneously acquire the two cross-polarized diffraction images (p-DI) using two charge coupled device (CCD) cameras or one time-delay-integrated (TDI) camera from each cell at a non-conjugate location by an objective-based imaging unit, which enables for the first time extraction of cell information from the diffraction patterns for detailed cell or particle analysis. The use of TDI camera allows a p-DIFC system to acquire p-DI data at a much higher speed and less blurring by synchronizing the motion of pixel rows in the TDI camera with the flowing cells.

2.5 Evaluation of effusion cells

In the human body, serosa is a thin layer of mesothelial cells lining the inside of the pleural, and peritoneal cavities. It covers the lung parenchyma, the mediastinum and the diaphragm. A small amount of lubrication fluid called pleural and peritoneal fluid fills the pleural and peritoneal spaces, and allows the moving organs such as lungs, and gastrointestinal organs inside the cavities to slide during their movement. The accumulation of excess fluid inside these cavities due to illness leads to pleural or peritoneal effusions. Clinically, the pleural and peritoneal effusions can be classified into two types, the transudative effusions and the exudative effusions. The transudative effusions result from an imbalance of hydrostatic and oncotic pressures associated with congestive heart failure. In comparison, the exudative effusions are caused by injury to the mesothelium which usually occurs with malignancy 44% –77%, infections, or organs leakage [3, 45, 46]. The presence of the malignant cells in the effusion fluid define the malignant effusion (ME) and usually reflects advanced malignancy stages, which is a result of the failure in the protection mechanisms of the pleura and the abnormal mesothelial function [2]. The annual estimated incidence of developing ME in the United States is between 150,000 and 175,000 cases , and in the United Kingdom is

40,000 case [47]. The most common types of malignancies that developing ME are lung cancer, breast cancer, lymphoma, genitourinary, and gastrointestinal carcinomas respectively. The lung and breast are developing ME with about 50% to 65% out of all ME [3, 4].

Sampling of effusion fluid through thoracocentesis allows for further analysis that may help to establish the origin of malignant effusion [2, 48]. In fact, The primary purpose of cytopathologic diagnoses on specimens of effusion fluid is to determine whether or not malignant cells are present. If they are indeed present, then cytologists determine the cancer cell type and tissue or organ of origin. The preparation of effusion cell slides and subsequent diagnosis are very time consuming and depend critically on the level of training and expertise of the pathologists or cytologists. For example, accurate diagnosis requires high quality of preservation and cellular display of the cell slides and the nature of the disease involving the pleural membranes also play a role in the final outcomes [49].

Normal or benign effusions contain mainly three different cell types in varying portions of mesothelial cells, macrophages, and lymphocytes in addition to red and other white blood cells, which are present due to bleeding during the specimen collection processes as common contaminants [45, 49]. Under a microscope, mesothelial cells appear to have round shapes with round shaped nuclei and single nucleoli [50]. Binucleation or multinucleation are infrequent features of normal mesothelial cells. These cells are often dispersed as single cells or aggregate in small groups of cells. Other characteristic features of mesothelial cells are the peripheral lucent zone, the dense perinuclear zone, the occasional binucleation, and the slit-like clear separation between adjacent cells. Unlike mesothelial cells, macrophages cells have small folded nuclei and vacuolated cytoplasm [45]. On the other hand, malignant effusions are mostly identified from patients of cancers with a high tendency to metastasize into the pleural and peritoneal cavities such as lung cancers, breast cancers, ovarian cancers, gastrointestinal cancer, hematological cancers, and genitourinary cancers [3, 48, 51]. In this study, we focus our effort on profiling of cells in effusion samples taken from patients with lung and ovarian cancers because these two cancer types are the most common among the patients going through cytology exams.

The lung cancers are classified into two types of non-small cell lung cancer (NSCLC) and the small cell lung cancer (SCLC). In contrast to SCLC patients who rarely have abnormal pleural effusions, NSCLC patients are likely to develop pleural effusions [52]. The NSCLC cases mainly consist of adenocarcinoma and its sub-type bronchioloalveolar carcinoma, and squamous cell carcinoma. Adenocarcinoma starts from the lung epithelial cells and develops pleural effusions, and bronchioloalveolar carcinoma begins in the cells that produce mucus. Squamous cell carcinoma can be found in the cells that line the airways. Other than these types above is large cell carcinoma [49, 53]. However, the American Joint Commission on Cancer (AJCC) approved a staging system to stage the lung cancer spread in the human body [54]. This system named TNM staging system. In which T, N, M stages represent the size of the tumor, the level of spread within the lymph nodes, and the distance metastatic spread of the tumor cells respectively. The staging process depends on the imaging modalities such as MRI, CT, and PET for identifying lesion site and size[48, 55].

On the other hand, epithelial carcinoma makes up 85% to 90% of the ovarian cancers. The ovarian cancer is divided into four stages based on the tumor cells availability location. The first three stages representing the cancer being limited by the ovaries, by the pelvis and by the peritoneal cavity, respectively. Stage four is more than that i.e. finding the cancer cells in the pleural cavity and in other parts of the body [56].

Chapter 3 Confocal Imaging and Cytology Study

In this chapter, we present a quantitative study of the 3D morphology of 449 cells extracted from fresh human PPE samples. The confocal microscopy method has been employed for acquiring image stacks from single cells with double fluorescent staining. Cell reconstruction has been carried out to obtain 27 morphological parameters for quantitative analysis of cell's 3D structures, which provides essential input data for development of realistic optical cell models in simulation of diffraction imaging process. A clustering algorithm has been developed to analyze the distribution of the 449 effusion cells in the high dimensional space of 3D morphological parameters for study their correlations to the pattern features of diffraction images.

3.1 Cell preparation and image acquisition

We performed a quantitative study of the 3D morphology of live cells extracted from fresh PPE samples. These samples were obtained from patients scheduled for cytology diagnosis. Among a typical patient population undergoing PPE diagnosis, about 20% to 30% of them received positive results for cancer diagnosis [57]. The PPE samples received for this study were small portions of patient samples which otherwise would be disposed. The PPE samples were obtained from the Department of Pathology at the Brody School of Medicine under an ECU-IRB approved protocol for optical study of human tissues.

A PPE sample received for this study was centrifuged twice at 1500 RPM (27°C) for 5 minutes to extract and re-suspend the cells. After the first centrifuge, the supernatant was discarded, and the sediment cells were re-suspended in 10 mL of red blood cell (RBC) lysis buffer to remove the

RBCs. The cell suspension was shaken at room temperature for 10 minutes followed by another centrifuge using the same set of parameters (1500 RPM at 27°C for 5 minutes) to re-suspend the cells in culture medium buffer (Phosphate buffer saline (PBS) pH7.4 and 1% of Bovine serum albumin (BSA)). The suspended cells were counted by taking 15 μL of the prepared cell suspension and adding 15 μL of Trypan blue. The mixed solution was placed into a hemocytometer for cell counting to determine the total cell number and the concentration of the prepared cell suspension. Then, the sample was divided into two parts of equal volume ratio. One part of cell sample was transported on ice for p-DIFC measurements in BLL in the Howell Science Complex building, and the other one was prepared for the confocal microscopy imaging in Brody School of Medicine. The protocols of the cells preparation and counting can be found in Appendix A and B.

The confocal imaging of PPE cells was performed using a laser scanning confocal microscope (LSCM) (LSM510, Zeiss) after adjusting the cells suspension volume and double staining. Two fluorescent reagents were used for staining nuclei by Syto-61 and mitochondria by MitoTracker-orange (S11343, and M7510, Life Technologies). For proper staining, the prepared cells need to be incubated for 40 minutes at 37 °C with 5% CO_2 followed by one washing by culture medium (PBS with 1% BSA) and re-suspension in culture media. The protocol of the fluorescence staining process is provided in Appendix C. At that point, an aliquot of 150 μL of cells suspension will be placed between a dipped and cover glass slides and loaded to the microscope for imaging. Two different laser beams of wavelength 0.633 μm and 0.543 μm were used to excite the fluorescent molecules of Syto-61 and MitoTracker Orange inside the cell organelles. The laser beams are focused inside randomly selected cells by the oil immersed objective lens of 40x or 100x magnification, which is also used to collect the fluorescent light emitted from the reagent molecules. The laser beams are scanned by two mirrors before the objective with their overlapping focal spots moved over the horizontal (x, y) plane. The collected fluorescent light signals at each scanned point are spatially filtered by a pinhole, which prevents fluorescent light coming out of the regions outside of the focal spots to be detected by the photomultiplier tube (PMT) placed after the pinhole. The light signals emitted by the different fluorescent reagents were recorded by different channels of wavelength

filter and PMT and were stored in different color channels in the output color image file at a specific depth or z-axis position. An image slice such as Figure 3.1 is composed of two channels: the red channel for the fluorescence light intensity mainly from nucleus; the green channel for the fluorescence mainly from mitochondria and fluorescence from cytoplasm contributing to both.

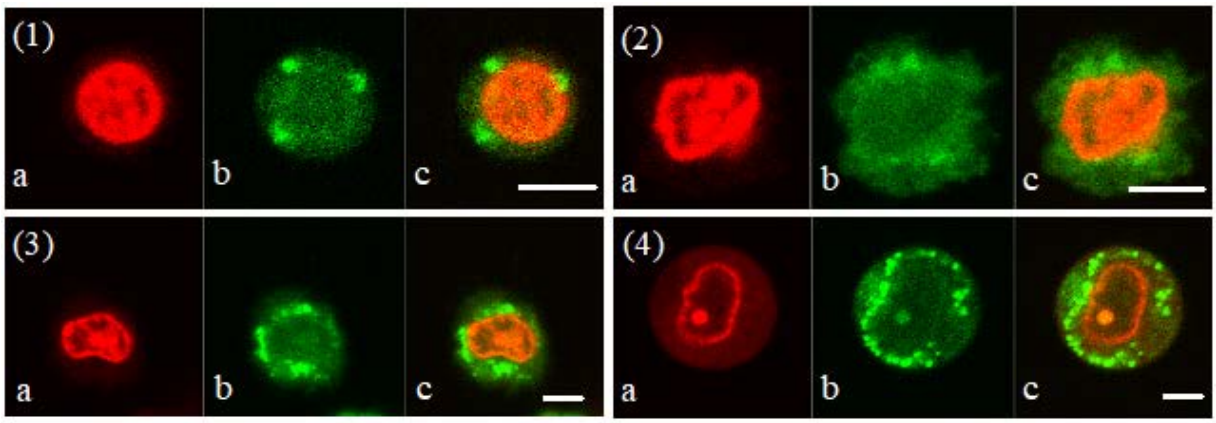

Figure 3.1 Confocal image slices of four different PPE cells double stained with Syto-61 for nucleus and MitoTracker-orange for mitochondria. Nucleus and cytoplasm stained and imaged in the red channel (a), mitochondria and cytoplasm stained in the green channel (b), and the combination of both channels (c). The cells values of cell volume in μm^3, nucleus-to-cell volume ratio, and mitochondria-to-cell volume ratio are: (1) 388.8, 30.9%, 3.5% (2) 955.2, 19.1%, 4.8% (3) 1956.3, 12.5%, 7.1% (4) 3782.2, 18.5%, 8.8%. Bars=5μm

An image stack of multiple image slices can be acquired by translating the imaged cell sample over a sequence of z-axis positions with a stepsize of Δz. Each image stack consists of about 20 to 80 slices with 12-bit pixel depth and the frame size of 512×512 pixel. The z-axis translation step size in air Δz is typically 0.49μm as shown in Figure 3.2. We note that the pixel size in each image slice is about 0.07μm and thus much smaller than the stepsize of z-axis translation.

3.2 Reconstruction and 3D parameter calculation

The confocal image processing includes steps of image segmentation, 3D reconstruction, and morphological parameter extractions. A dedicated code of CIMA (Cell Image and Morphology

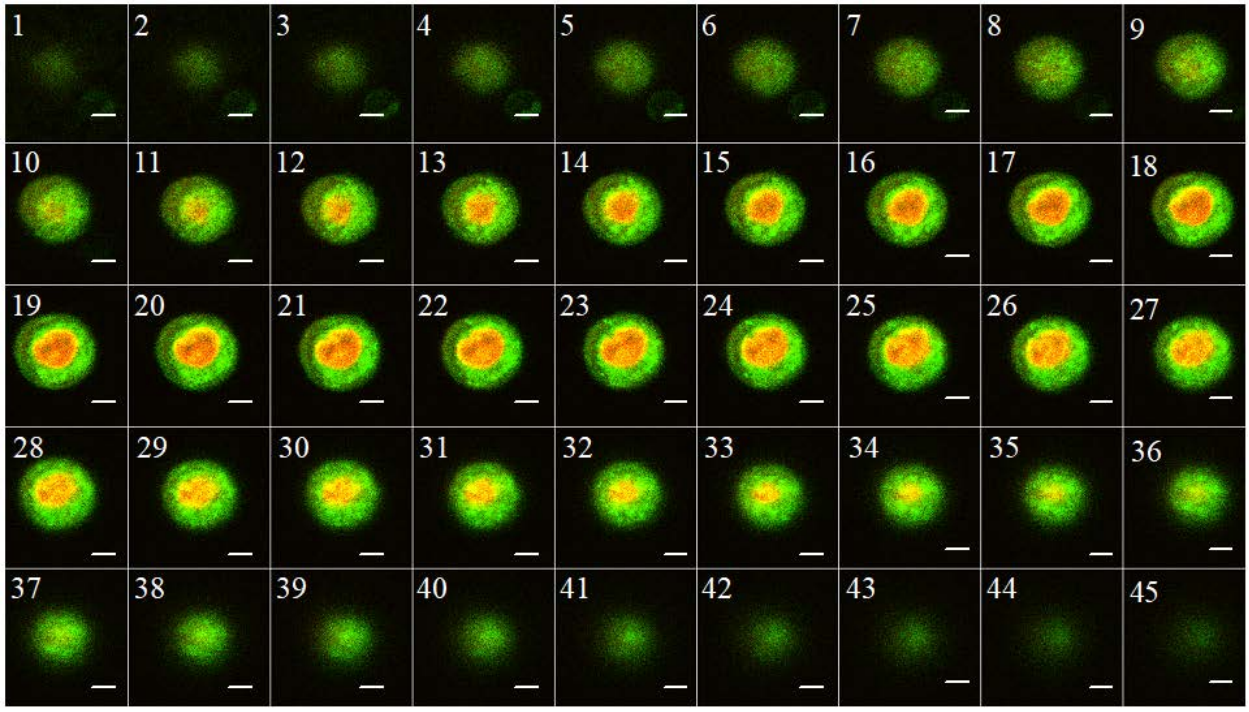

Figure 3.2 Confocal image stack acquired from a PPE cell and marked by the slice index at the upper left corner. Bars= $5\mu m$.

Analysis) has been developed in-house on the platform of Matlab (2013a, Mathworks) [58, 59]. An image stack is first imported into the software which then segments each image slice into the two color channels to produce intensity histogram and obtained appropriate thresholds for segmentation. With these thresholds, the pixels in each slice are separated into four types of extracellular (or background), cytoplasmic, nuclear and mitochondrial pixels. The pixels at the boundaries of cytoplasm and nucleus are also identified as respective membrane pixels.

To make voxels of 3D reconstruction close to cubic, multiple slices need to be interpolated between the segmented raw image slices. For this purpose, a correction factor to the stepsize of z-axis translation was first determined by the reconstruction of a latex microsphere to compensate for the effect of light refraction on z-axis stepsize [10]. Using the corrected z-axis stepsize the number of interpolated slices was obtained. The types of each pixel in an interpolated slice was determined according to the types of the corresponding pixels in and distances to the raw slices neighboring the interpolated slice. After obtaining the 3D reconstruction of the imaged cell, the CIMA software quantitatively analyzes the 3D structures and calculate 27 morphological

parameters of the cytoplasm, nucleus, and mitochondria. The output data files of CIMA provides the essential data for our clustering analysis of cell distribution in the space of morphological space, development of realistic cell models and study of the cell morphology correlation with the p-DIFC data.

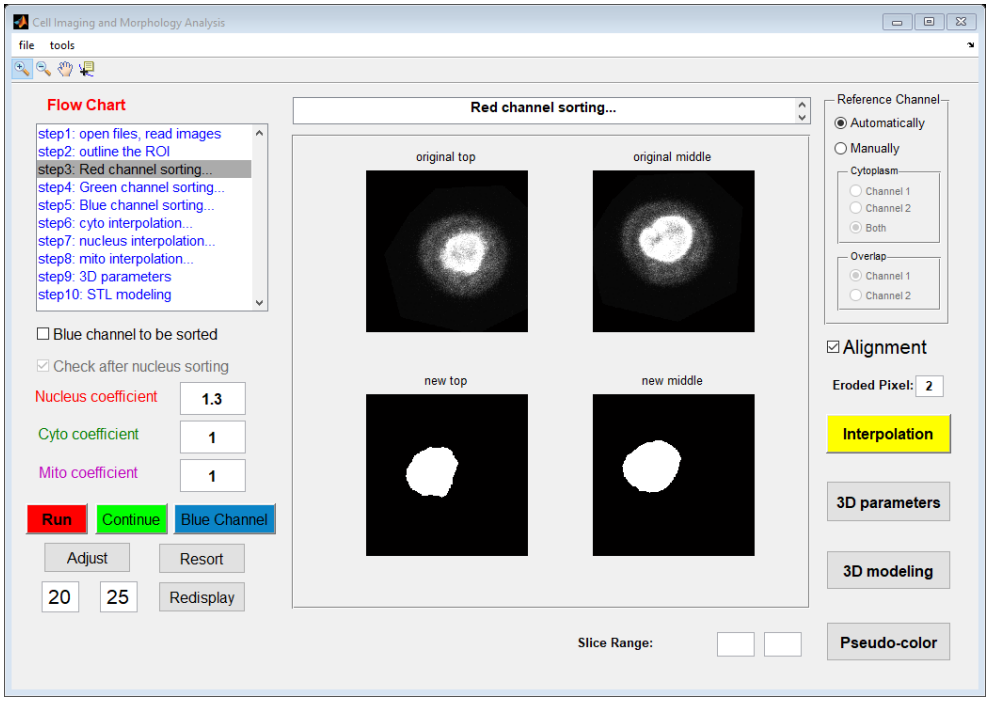

Figure 3.3 Graphical user interface of the 3D reconstruction software (CIMA) showing red channel sorting for the nucleus and cytoplasm.

To process the cell images with the CIMA software, firstly we use Zeiss laser scan microscope (LSM) image browser to open the LSM image format and save it as 12-bit tiff image format because CIMA software is only compatible with tiff image stacks format. The software reads the tiff image stacks according to the input files' path, the total image slice number, the Δz step size, and the pixel size of the image. We outline the region of interest manually to select the pixels for the cell and mask all the other pixels. The system will start sorting the images and calculate the threshold values. A histogram analysis of pixels intensities will be performed for segmentation in both of the red and green channels, and the first minimum value after the first peak will be chosen as the threshold of the cell membrane and extracellular pixels. Meanwhile, the pixel intensity of the second peak will be picked as the threshold of the intracellular, nucleus and mitochondria pixels

respectively. After the histogram analysis was done for each of the two channels, slices in each channel were then processed separately for segmentation of different organelles before they are recombined to output the final 3D structure files.

Sobel operation is performed based on derivative of the spatial gradient for edge detection on all cleaned slices stack, and the threshold of derivatives are selected to generate a binary stack for the segmented cytoplasm and nucleus from the background. Opening and closing fill-point gap operations are implemented to remove the unacceptable pixels and smooth out the cell membranes' border lines. We check the result of the nucleus after sorting. If the size of the nucleus is too large or the generated shape of nucleus does not match the original shape, the threshold we used is too small. Therefore, it is necessary to resort the nucleus image pixels by modifying the threshold. After sorting the nucleus, the green channel the cytoplasm membrane was first detected to form a binary stack. The two cytoplasm stacks from both channels were compared, and the one with the more prominent area in each slice was selected as the final output. For accurate segmentation of mitochondria, other algorithms had to be employed due to their much smaller sizes, which include adaptive median filtering with auto-adjusted window size for smoothing and watershed algorithms to increase the accuracy in separating different mitochondria clusters within aggregated mitochondria pixels. Next, we check the result for the mitochondria sorting to determine if a resorting is required. If we are satisfied with the result, the system will do the final step of shape-based interpolation for the cell, the cytoplasm, the nucleus, and the mitochondria accordingly.

Additional slices are interpolated between the neighboring slices to obtain an array of nearly cubic voxels. Since the refraction of the fluorescence light at various interfaces from the emitting molecules to the microscope objective, the spatial distance between neighboring slides given by the sample translation stepsize needs to be corrected by factor $f = 0.862$ [10]. After the 3D reconstruction, we quantitatively analyze the whole cell structures and obtain 27 morphological parameters of the cytoplasm, the nucleus and the mitochondria, which provide essential cell morphology features. The definitions of these parameters are in Appendix D. The results of the confocal imaging and 3D reconstruction were used to compare the morphological parameters of

different cell types for cell clustering study quantitatively. Also, the fluorescent light intensity distribution in the confocal imaging served as baseline data to develop a more realistic optical cell model and gain insight into cell morphological differences through the simulated light diffraction patterns.

3.3 Results of 3D morphology measurement

We performed multiple confocal imaging measurements on PPE cells samples followed by image segmentation, 3D reconstruction, and morphological parameters extraction. The extracted parameters were used for quantitative investigation of cells classification via clustering technique. While, the 3D reconstructed images of the cells will be recalled to develop realistic optical cell models for p-DI simulation.

Table 3.1 Pathological information of PPE cells samples.

Patient ID [a]	Sample Date [b]	Type of Effusion	Status	Diagnosis	Primary Location	No. of Imaged Cells
P1	Jul.07,17	Pleural	Malignant	Adenocarcinoma	Lung	17
P2	Jul.07,17	Pleural	Malignant	Adenocarcinoma		37
P3	Jul.13,17	Pleural	Malignant	Ovarian serous carcinoma	Ovaries	32
P4	Jan.08,18	Peritoneal	Malignant	Ovarian carcinoma	Ovaries	58
P5	Jan.24,18	Pleural	Malignant	Adenocarcinoma	Lung	48
P6	Feb.26,18	Pleural	Benign	Benign		21
P7	Mar.02,18	Pleural	Benign	Benign		41
P8	Apr.20,18	Pleural	Malignant	NSC Lung cancer	Lung	52
P9	Jun.26,18	Pleural				24
P10	Jun.29,18	Peritoneal	Malignant			46
P11	Jul.19,18	Pleural	Benign	Benign		42
P12	Aug.10,18	Pleural	Benign	Benign		32
Total						499

[a] Fresh PPE fluid samples from twelve different patients diagnosed with abnormal PPE fluid accumulation and identified by P1, P2, ... ,P12

[b] PPE sample collection date is the same date of cells extraction and imaging.

A total of 449 confocal imaged cells, used in our study, were extracted from twelve PPE cell samples for twelve different patients diagnosed with abnormal pleural and peritoneal fluid accumulation. Table 3.1 summarize the pathological information of the studied cells samples with the total number of confocal imaged cells from each sample. In this study, we identify the cell

samples of those patients by patient's ID.

All cells were imaged with the confocal microscope and their 3D morphology were reconstructed from the acquired image stack data with the software described before. The 27 parameters were calculated to characterize 3D morphology of the 449 imaged PPE cells. Table 3.2 provides the mean and standard deviation (STD) values for the morphology parameters of pooled data. It can be seen from these data that the spreads of parameter values are very large with the STD values comparable or even exceeding the mean values that include cell volume, nuclear volume, mitochondrial volumes and their ratios. This indicates well dispersed distributions of cell and intracellular organelle sizes among the PPE cells. The only exception is the cell and nuclear volume sphericity indices VSi_c and VSi_n. Note that by definition VSi is given by the volume ratio of spherical volume calculated by the equivalent radius ER, to actual volume V and thus a value of 1 for VSi represents perfect spherical shape. Taken together, one can see that these PPE cells, large or small, tend to have similar nonspherical shape, which could be attributed to their ability to remain as suspension cells.

Table 3.2 Morphological parameters of PPE cells

Parameter	**Symbol**	**Units**	**Mean ± STD** n=(449)[a]
Cell grid perimeter	GP_c	μm	4110 ± 2596
Cell surface area	S_c	μm^2	539.7 ± 377.8
Cell volume	V_c	μm^3	669.7 ± 649.0
Surface to volume ratio of cell	SVr_c	μm^{-1}	0.970 ± 0.220
Index of surface irregularity of cell	SIi_c	$\mu m^{-1/2}$	167.0 ± 50.68
Cell equivalent spherical radius	ER_c	μm	5.023 ± 1.421
Cell volume sphericity index	VSi_c		0.660 ± 0.066
Average distance from centroid to cell membrane voxels	$< R_c >$	μm	5.068 ± 1.552
Standard deviation of $< R_c >$	ΔR_c	μm	0.421 ± 0.350
Nuclear grid perimeter	GP_n	μm	1905 ± 1040
Nuclear surface area	S_n	μm^2	244.6 ± 136.3
Nuclear volume	V_n	μm^3	185.4 ± 145.9
Nuclear surface to volume ratio	SVr_n	μm^{-1}	1.482 ± 0.331
Index of surface irregularity of nucleus	SIi_n	$\mu m^{-1/2}$	144.8 ± 44.95
Nuclear equivalent spherical radius	ER_n	μm	3.377 ± 0.722
Nuclear volume sphericity index	VSi_n		0.632 ± 0.076
Average distance from centroid to nuclear membrane voxels	$< R_n >$	μm	3.456 ± 0.778
Standard deviation of $< R_n >$	ΔR_n	μm	0.481 ± 0.273
Mitochondrial grid perimeter	GP_m	μm	1402 ± 1657
Mitochondrial surface area	S_m	μm^2	190.1 ± 245.4
Mitochondrial volume	V_m	μm^3	37.00 ± 71.86
Surface to volume ratio of mitochondria	SVr_m	μm^{-1}	8.242 ± 2.941
Index of surface irregularity of mitochondria	SIi_m	$\mu m^{-1/2}$	242.9 ± 131.1
Mitochondrial equivalent spherical radius	ER_m	μm	0.423 ± 0.190
Distance between the centroids of nucleus and cell	CD_m	μm	0.880 ± 0.848
Volume ratio of nucleus to cell	Vr_{nc}		0.362 ± 0.168
Volume ratio of mitochondrion to cell	Vr_{mc}		0.039 ± 0.032

[a] n= number of imaged cells.

While the mean and STD values in Table 3.2 provide an overview of the PPE morphology, they cannot provide information on how these parameters relate to each other. For this purpose, scatter plots of the 27 parameters can yield much detailed descriptions with different combinations of two parameters. Six scatter plots of the imaged PPE cells with different pairs of morphology parameters are presented in Figure 3.4 as examples to compare their distributions. Despite their dispersed distributions, the symbols representing the cells show significant overlap in all of the scatter plots. A more detailed reading of the parameter distributions for PPE demonstrates some interesting observations on the significance of 3D quantification. The PPE cells with small volume

appear to have a less spread in their values of cellular, nuclear, and mitochondrial volume parameters (Figure 3.4 (a) and (b)) than the cells of relatively bigger volume. While, the distribution of the cells in the parameters of volume ratios (Figure 3.4 (c)) shows a more spread in the range of nuclear to cell volume ratio values than the range of the mitochondrion to cell volume ratio values, which also be noted from the standard deviation of these parameters. Since the cell equivalent radius ER_c and the index of surface irregularity SIi_c are proportional to the numbers cell volume voxels, they distribute roughly along three lines representing their clustering mainly dependence on cell volume V_c more than nuclear and mitochondrial volume for majority of the PPE cells (Figure 3.4 (d), (e), and (f)).

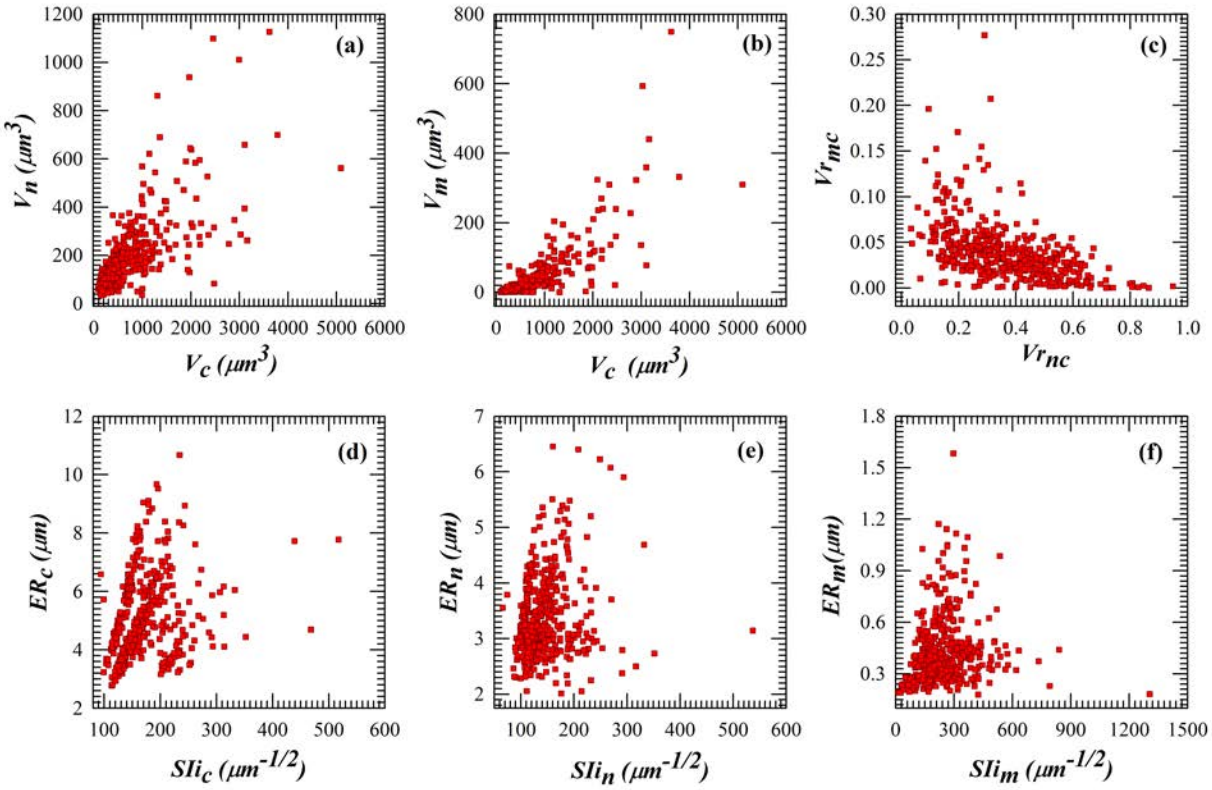

Figure 3.4 The scatter plots of PPE cells with 6 combinations of 3D parameters: (a) V_c vs V_n; (b) V_c vs V_n; (c) Vr_{mc} vs Vr_{nc}; (d) SIi_c vs ER_c; (e) SIi_n vs ER_n; (e) SIi_m vs ER_m.

Therefore, these scatter plots by the 3D parameters are difficult to be used for finding boundary lines in the parameter space to distinguish different types of cells of the effusion samples. To achieve the goal of cells classification, it is imperative to adopt a machine learning algorithm which can be used to objectively and effectively to identify the distribution patterns of cells types. In this study, we employed an unsupervised algorithm of the Gaussian Mixture Model (GMM) for automated cell classification in the 3D parameter space of cell morphology.

3.4 GMM and clustering analysis

GMM is one of the soft unsupervised machine learning algorithms that has been used widely for data mining research such as separating input data into k clusters in high multidimensional parameter space with mixed parameter values [60]. In general, the GMM divides a given collection of samples represented by the input data into k classes by assigning a weighted sum of Gaussian probability density functions (PDF) with model parameters iteratively updated using the Expectation-Maximization (EMax) iteration method [61]. The optimized GMM model parameters that fit the distribution of the input data best will be obtained through the Maximum-Likelihood Parameter Estimation method (MLPE). MLPE parameter estimates can be calculated iteratively with the EMax algorithm starting with an initial set of GMM parameters, which stops as the model parameter values become stabilized [62, 63]. More explained details of the GMM algorithm found in Appendix E.

The GMM clustering outcomes are very sensitive to the initial fitting parameters (values of mean and covariance matrix elements) that are usually obtained from random numbers for each assigned cluster [64]. As a result, the outcomes of GMM clustering fluctuated. To avoid such problem, another algorithm needs to be introduced to produce an initial set of parameters for obtaining stable GMM clustering outcomes [65]. In this study, the hierarchical clustering (HC) has been tested and chosen to output clustering results and used to start the GMM clustering as the initial values of mean and variance for each of k clusters [66, 67]. The HC method starts by assigning each 3D

morphology parameter vector associated with an imaged cell to its own cluster with a total number of clusters equals to the total number of data points available. Then HC measures the distance between each pair of points in the dataset and calculate the similarity between them. The algorithm finds the minimum distance between the nearest pair of clusters (data points) and merge them into a new single cluster. After each cluster merging, the algorithm compute the distance from the average distance of merged cluster to all other clusters and combine the nearest two iteratively into a larger cluster. This process will stop after the total number of survival cluster reaches k as the assigned number of clusters [68]. To take advantage of both algorithms, we developed an unsupervised machine learning code based on a combination of HC method and GMM method to investigate cell classification not only in 3D morphological parameters space but also in GLCM parameter space as shown in the next chapters. In this chapter, we present the GMM based PPE cell classification results using the morphology parameters. The results of PPE cell classification by the GMM based algorithm will be presented in next chapter.

3.4.1 Confocal imaging cluster analysis

All 3D morphological parameters extracted from the reconstructed cells using CIMA are imported to another in-house developed Matlab code for HC and GMM calculations. The imported data is in the form of a 449×27 matrix. The rows (449) correspond to the cell sequence, and the columns (27) correspond to the 3D parameter sequence. A Matlab integrated function "fitgmdst" is used to fit Gaussian mixture distribution to the input data for clustering PPE cells data in 3D parameters space. A clustering task is specified by arguments with pairs of name and value to control the iteration process and achieve stable clustering results. These modeling arguments include the number of mixture components, i.e., assigned number of clusters k, the regularization parameter value, and the EMax algorithm iteration number [69].

As discussed before, The "fitgmdist" function for implementing the EMax algorithm is sensitive to initial values of the mean and elements of the diagonal covariance matrix, and uniform mixing proportion of 3D parameters for the desired Gaussian clusters. By setting the initial parameter

values with random numbers, the GMM algorithm usually end with unstable clustering results. Table 3.3 shows three examples of GMM clustering runs performed on all 3D morphological parameters extracted from the total number of PPE cells separating the data into two clusters denoted as C1 and C2. To avoid fluctuation, the HC algorithm was applied before the GMM clustering to establish the initial parameters.

Table 3.3 Three examples of GMM clustering with a random start for PPE cells samples.

GMM run	**C1**	**C2**
1^{st}	265	184
2^{nd}	241	208
3^{rd}	293	156

3.4.2 Number of clusters

The major problem with any clustering technique is to determine the appropriate number of clusters that best describe the distribution of the input data. The HC and GMM algorithms do not automatically determine the number of clusters. For that reason, clustering algorithms need to be run with assigned values of clusters number k, and the best value for k needs to be determined based on predefined criterion [70]. In this study, we employ the Akaike information criterion (AIC) and the Bayes information criterion (BIC) as statistical estimators to evaluate the optimum number k of clusters in our classification study [71, 72]. Both of these criteria represent a balance between maximizing the degree of fit of the data point to a cluster model and minimizing number of clusters [73].

AIC and BIC are used to find optimized k value by comparing and identifying their minimum values among possible k values. After iterations with the GMM algorithm, the log-likelihood function stabilizes at the maximum value and yields the best model for fitting the input data. The AIC and BIC parameters are defined by [74]:

$$
\begin{aligned}
AIC(k) &= -2\log(ML) + 2D, \\
BIC(k) &= -2\log(ML) + D\log(N),
\end{aligned}
\tag{3.1}
$$

where ML is the maximum of the likelihood function, D is the number of morphology parameters in the model, and N is the number of data points, i.e., sample size. In comparison to the AIC that better suited with models of fewer N than D, the BIC is more appropriate for real-world situations like our study with D much less than N. [74][75].

3.5 Clustering results

We applied HC and GMM clustering technique in high dimensional space to separate all the imaged cells in PPE samples into meaningful groups according to their 3D morphology parameters. We set the software to divide the data into 1, 2, 3, 4, and 5 clusters in the space of 27 morphological parameters. In each step, the AIC and BIC values are calculated with cluster number incremented from 1 to 5. The calculated indices of AIC and BIC are plotted separately against the trial values of cluster number k for comparison purposes in Figure 3.5. The figure shows that the lowest value for the BIC parameter is achieved at k=3, while the AIC starts level off at the same point, i.e., the PPE cells should be separated into k=3 groups based on the 27 morphological parameters. In the following analysis the 3 clusters are named as C1, C2 and C3 which are mainly characterized by their differences in cell volume with C1 includes the cells of smallest volumes while C3 consists of cells of largest volumes.

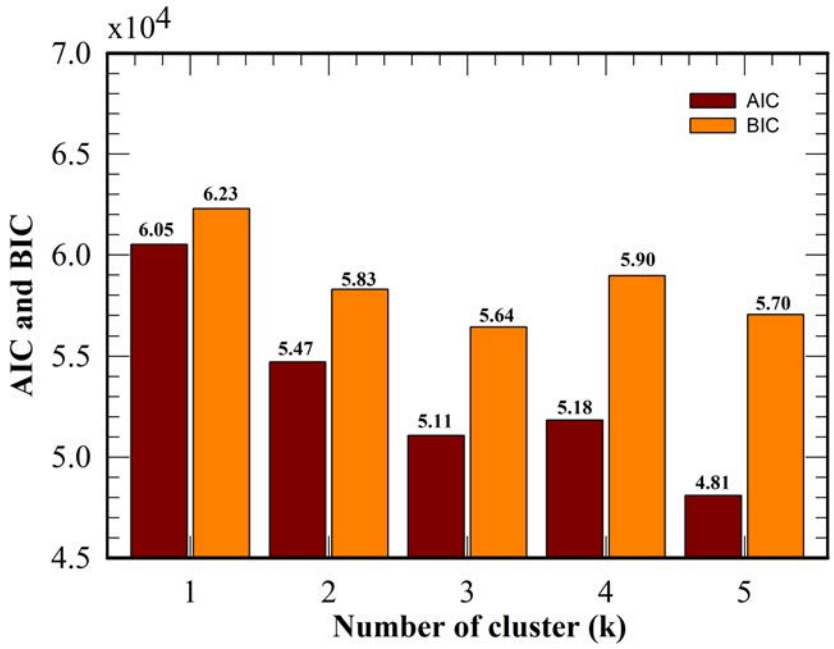

Figure 3.5 AIC and BIC vs. different number of clusters.

Table 3.4, and Table 3.5 present the values of the mean and the standard deviation of 27 morphology parameters with the p-values to examine the statistical significance on the parameter difference among the three cell clusters. The most statistically significant morphological differences are related to cell grid perimeter, cell volume, cell equivalent spherical radius and the average distance of cell membrane voxel to the centroid. It is also clear that the mean volume of the cell, nucleus, and mitochondria tend to have largest values for cells in C3 group and the smallest values for C1 cells, while the cells in the C2 group have the medium values. The results are consistent with our visual examination of the fluorescent images stacks.

Table 3.4 Morphological parameters of partition PPE cells with GMM algorithm (k=3)

Parameter [b]	**Symbol**	**Units**	**Mean ± STD** [a] **C1** n=(250) [c]	**C2** n=(159)	**p-value** [d]
Cell grid perimeter	GP_c	μm	2463 ± 2466	5201 ± 1119	3.08×10^{-76}
Cell surface area	S_c	μm^2	313.8 ± 108.7	700.9 ± 203.0	1.98×10^{-57}
Cell volume	V_c	μm^3	301.7 ± 142.7	889.6 ± 396.4	2.18×10^{-42}
Surface to volume ratio of cell	SVr_c	μm^{-1}	1.090 ± 0.140	0.850 ± 0.200	1.05×10^{-31}
Index of surface irregularity of cell	SIi_c	$\mu m^{-1/2}$	145.8 ± 33.29	183.2 ± 40.34	1.14×10^{-19}
Cell equivalent spherical radius	ER_c	μm	4.070 ± 0.610	5.840 ± 0.880	2.73×10^{-61}
Cell volume sphericity index	VSi_c		0.690 ± 0.050	0.630 ± 0.070	1.41×10^{-19}
Average distance from centroid to cell membrane voxels	$< R_c >$	μm	4.040 ± 0.610	5.920 ± 0.860	1.46×10^{-67}
Standard deviation of $< Rc >$	ΔR_c	μm	0.300 ± 0.160	0.540 ± 0.360	7.22×10^{-14}
Nuclear grid perimeter	GP_n	μm	1462 ± 456.7	2146 ± 789.5	1.83×10^{-19}
Nuclear surface area	S_n	μm^2	182.7 ± 52.21	284.1 ± 103.8	7.32×10^{-24}
Nuclear volume	V_n	μm^3	123.8 ± 53.04	223.1 ± 124.8	7.88×10^{-18}
Nuclear surface to volume ratio	VSr_n	μm^{-1}	1.560 ± 0.290	1.420 ± 0.360	1.13×10^{-4}
Index of surface irregularity of nucleus	SIi_n	$\mu m^{-1/2}$	134.1 ± 33.98	149.9 ± 38.70	3.54×10^{-5}
Nuclear equivalent spherical radius	ER_n	μm	3.040 ± 0.400	3.640 ± 0.680	3.56×10^{-20}
Nuclear volume sphericity index	VSi_n		0.650 ± 0.070	0.610 ± 0.060	3.01×10^{-10}
Average distance from centroid to nuclear membrane voxels	$< R_n >$	μm	3.060 ± 0.390	3.770 ± 0.670	6.73×10^{-26}
Standard deviation of $< R_n >$	ΔR_n	μm	0.350 ± 0.160	0.600 ± 0.240	4.66×10^{-25}
Mitochondrial grid perimeter	GP_m	μm	492.9 ± 332.8	1935 ± 905.4	3.77×10^{-46}
Mitochondrial surface area	S_m	μm^2	64.49 ± 50.58	262.2 ± 137.5	7.90×10^{-41}
Mitochondrial volume	V_m	μm^3	8.930 ± 11.36	48.41 ± 42.38	4.78×10^{-23}
Surface to volume ratio of mitochondria	SVr_m	μm^{-1}	9.210 ± 2.670	7.170 ± 2.650	3.62×10^{-13}
Index of surface irregularity of mitochondria	SIi_m	$\mu m^{-1/2}$	169.7 ± 68.09	306.6 ± 90.99	5.30×10^{-42}
Mitochondrial equivalent spherical radius	ER_m	μm	0.360 ± 0.120	0.480 ± 0.190	4.50×10^{-12}
Distance between the centroids of nucleus and cell	CD_m	μm	0.520 ± 0.450	1.130 ± 0.720	3.87×10^{-18}
Volume ratio of nucleus to cell	Vr_{nc}		0.450 ± 0.150	0.260 ± 0.110	3.15×10^{-39}
Volume ratio of mitochondrion to cell	Vr_{mc}		0.030 ± 0.020	0.050 ± 0.030	3.00×10^{-13}

[a] STD = standard deviation.

[b] parameters in bold font present the most statistically significant difference between the two clusters.

[c] n = number of imaged cells clustered together.

[d] p-values were obtained by a two-sample t-test method. The parameters with p-values ≤ 0.05 are regarded as the morphology features with statically significant differences between the two clusters off PPE cells.

Table 3.5 Morphological parameters of partition PPE cells with GMM algorithm (k=3)

Parameter[b]	**Symbol**	**Units**	**Mean ± STD** [a]		**p-value**[d]
			C2 n=(159) [c]	**C3** n=(40)	
Cell grid perimeter	GP_c	μm	5201 ± 1119	10058 ± 3233	9.00×10^{-12}
Cell surface area	S_c	μm^2	700.9 ± 203.0	1311 ± 594.9	1.14×10^{-7}
Cell volume	V_c	μm^3	889.6 ± 396.4	2095 ± 960.62	1.09×10^{-9}
Surface to volume ratio of cell	SVr_c	μm^{-1}	0.85 ± 0.20	0.66 ± 0.16	2.04×10^{-8}
Index of surface irregularity of cell	SIi_c	$\mu m^{-1/2}$	183.2 ± 40.34	235.5 ± 84.04	4.07×10^{-4}
Cell equivalent spherical radius	ER_c	μm	5.840 ± 0.880	7.740 ± 1.270	6.56×10^{-12}
Cell volume sphericity index	VSi_c		0.630 ± 0.070	0.610 ± 0.060	3.29×10^{-1}
Average distance from centroid to cell membrane voxels	$< R_c >$	μm	5.920 ± 0.860	8.070 ± 1.820	4.93×10^{-9}
Standard deviation of $< R_c >$	ΔR_c	μm	0.540 ± 0.360	0.720 ± 0.670	1.16×10^{-1}
Nuclear grid perimeter	GP_n	μm	2146 ± 789.5	3715 ± 1930	9.25×10^{-6}
Nuclear surface area	S_n	μm^2	284.1 ± 103.8	474.1 ± 259.6	4.65×10^{-5}
Nuclear volume	V_n	μm^3	223.1 ± 124.8	420.6 ± 277.0	7.06×10^{-5}
Nuclear surface to volume ratio	VSr_n	μm^{-1}	1.420 ± 0.360	1.240 ± 0.330	2.52×10^{-3}
Index of surface irregularity of nucleus	SIi_n	$\mu m^{-1/2}$	149.9 ± 38.70	191.6 ± 81.79	3.06×10^{-3}
Nuclear equivalent spherical radius	ER_n	μm	3.640 ± 0.680	4.440 ± 0.980	1.11×10^{-5}
Nuclear volume sphericity index	VSi_n		0.610 ± 0.060	0.580 ± 0.100	1.46×10^{-1}
Average distance from cetroid to nuclear membrane voxels	$< R_n >$	μm	3.770 ± 0.670	4.690 ± 1.080	5.47×10^{-6}
Standard deviation of $< R_n >$	ΔR_n	μm	0.600 ± 0.240	0.790 ± 0.440	1.33×10^{-2}
Mitochondrial grid perimeter	GP_m	μm	1934 ± 905.4	4961 ± 2838	4.99×10^{-8}
Mitochondrial surface area	S_m	μm^2	262.2 ± 137.5	689.0 ± 470.4	1.27×10^{-6}
Mitochondrial volume	V_m	μm^3	48.41 ± 42.38	167.0 ± 168.2	7.15×10^{-5}
Surface to volume ratio of mitochondria	SV_m	μm^{-1}	7.170 ± 2.650	6.480 ± 3.380	2.35×10^{-1}
Index of surface irregularity of mitochondria	SIi_m	$\mu m^{-1/2}$	306.60 ± 90.99	447.60 ± 200.70	8.66×10^{-5}
Mitochondrial equivalent spherical radius	ER_m	μm	0.480 ± 0.190	0.600 ± 0.310	3.31×10^{-2}
Distance between the centroids of nucleus and cell	CD_m	μm	1.130 ± 0.720	2.130 ± 1.480	1.47×10^{-4}
Volume ratio of nucleus to cell	Vr_{nc}		0.200 ± 0.110	0.220 ± 0.130	6.65×10^{-2}
Volume ratio of mitochondrion to cell	Vr_{mc}		0.050 ± 0.030	0.070 ± 0.050	2.10×10^{-2}

[a] STD = standard deviation.

[b] parameters in bold font present the most statistically significant difference between the two clusters.

[c] n = number of imaged cells clustered together.

[d] p-values were obtained by a two-sample t-test method. The parameters with p-values ≤ 0.05 are regarded as the morphology features with statically significant differences between the two clusters off PPE cells.

Figure 3.6 (a), (b), and (c) present charts for the mean and standard deviation values of cell, nucleus, and mitochondrial volume of the three clusters of PPE cells. Figure 3.6 (d) shows the mean and standard deviation values of the volume ratio of nucleus to cell and the volume ratio of mitochondria to cell. These data show monotonous changes of organelles' volumes for cytoplasm, nucleus and mitochondria. We note in Figure 3.6 (d) that the large cell cluster C3 exhibit opposite trend of volume ratio change between nucleus and mitochondria.

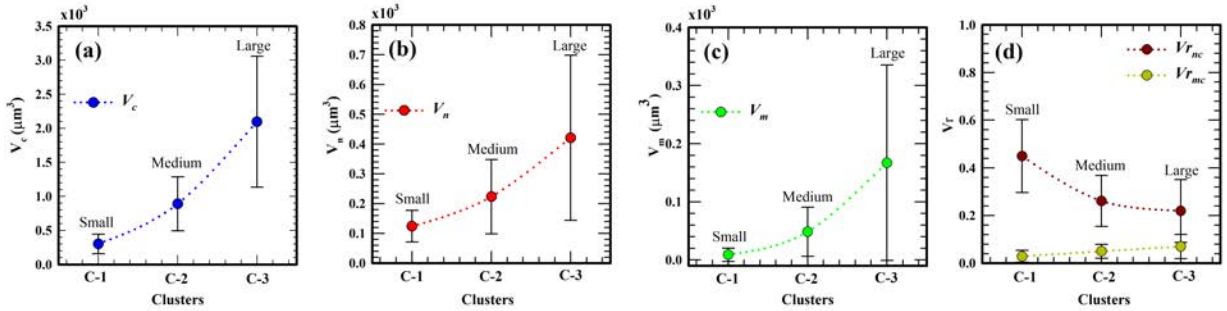

Figure 3.6 This figure shows the mean values of cell, nucleus, and mitochondria volume data and the volume ratios with the standard error bars of three clusters resulted from GMM clustering process for the confocal image data.

In Figure 3.7, we present the perspective views of the 3D structures for each of the three cells of the PPE cells labeled as C1 (small), C2 (medium), and C3 (large). Three parameters at the description of each cell are cell volume V_c, nucleus-to-cell volume ratio Vr_{nc}, and mitochondria-to-cell volume ratio Vr_{mc}. It can be observed directly from the image data that the significant differences among the three labeled clusters of PPE cells are in the cell volume. The C3 labeled cells are almost 2 and 3 times more prominent than C2 and C1 labeled cells respectively in the cell volume, while the C3 labeled cells have the largest volume ratio of nucleus-to-cell compare to other two clusters.

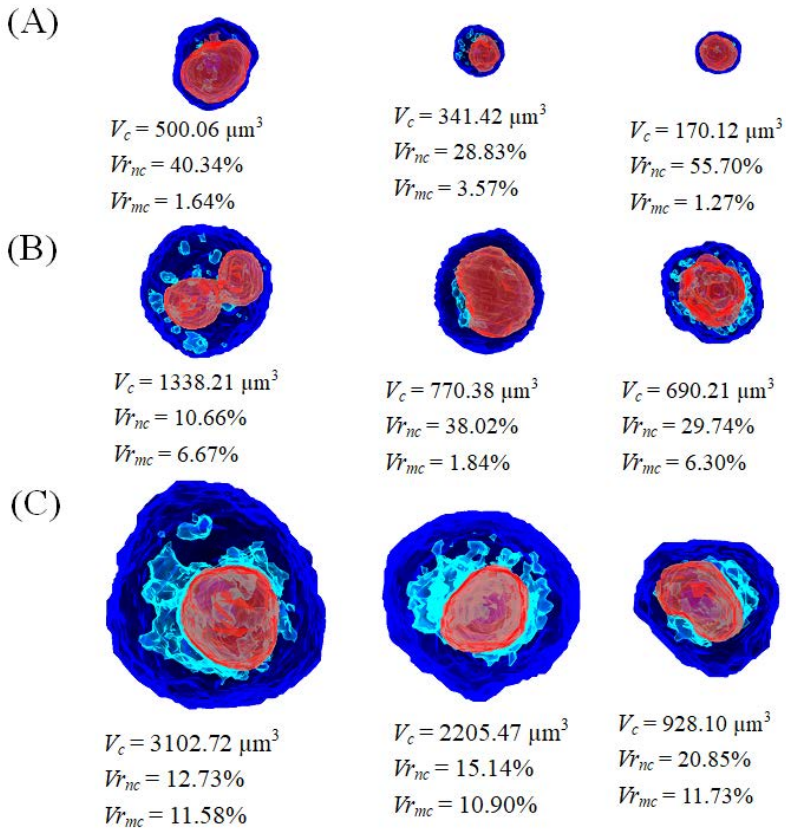

Figure 3.7 Perspective views of reconstructed 3D structures of PPE cells (A) C1 (B) C2 (C) C3. Three parameter at the bottom for each cell volume V_c, nucleus to cell volume ratio Vr_{nc}, and mitochondria to cell volume ratio Vr_{mc}.

Scatter plots of the imaged PPE cells with the morphology parameters of cell grid perimeter, cell volume, and surface to volume ratio of the cell are provided in Figure 3.8 to visualize and compare their distributions. Although some of the cells of the three types overlap each other in the scatter plots, they are well-separated. The C3 labeled cells cluster showed smaller values of the cell surface to volume ratio and higher grid perimeter and equivalent radius than those of the C2 and C1 labeled cells, which is consistent with the standard deviations of most parameters in Tables 3.4 and 3.5. In addition the C3 labeled cells have significantly larger cellular and nuclear volumes than C2 labeled cells, which is larger than C1 labeled cells. One can also observe similar trends from the perspective views of the 3D structures of the PPE cells in Figure 3.7. These results provide

insights on how to understand clustering PPE cells using the feature parameters of p-DI data as discussed in next Chapter. In Figure 3.8 we plot the distribution of the PPE cells as labeled by C1, C2 and C3 in the space of selected morphological parameters.

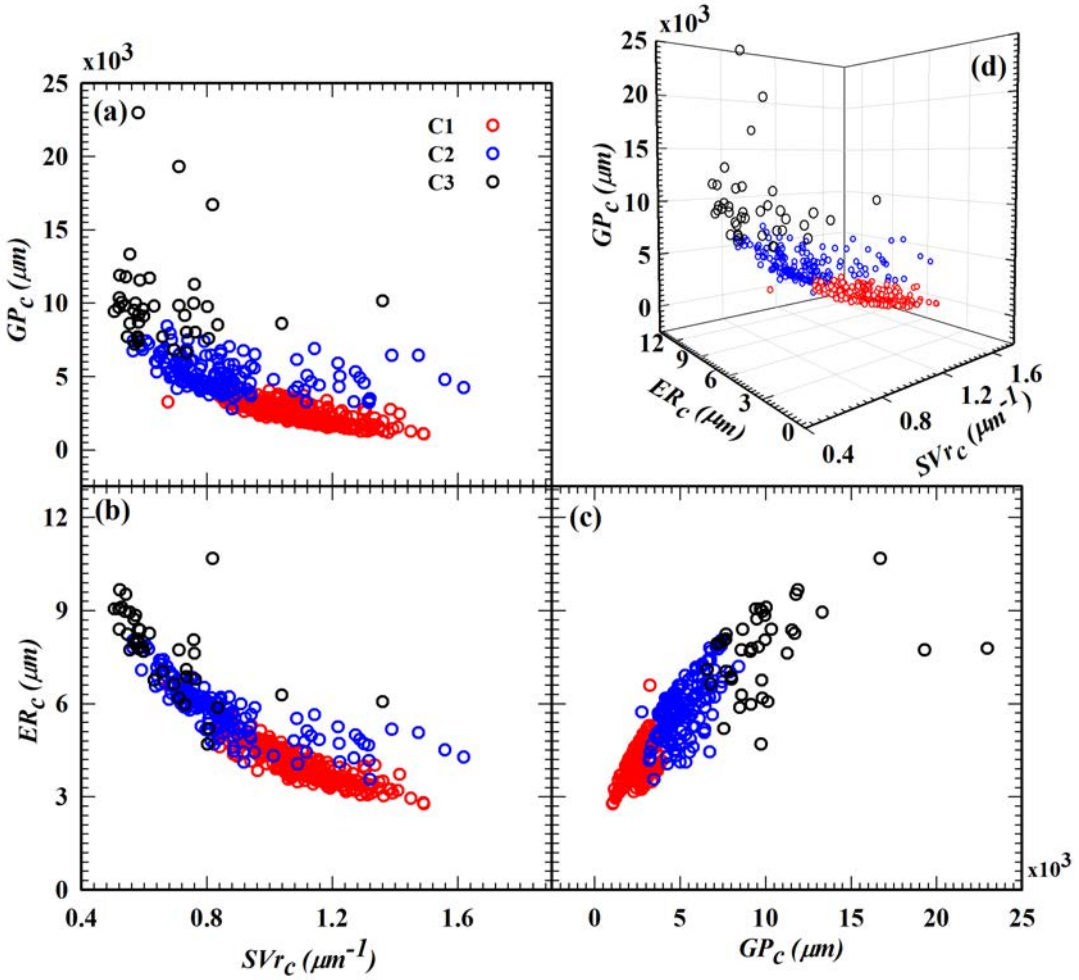

Figure 3.8 The distribution of the PPE cells as labeled by C1, C2 and C3 in the space of selected morphological parameters.

3.6 Cytopathological imaging and analysis

Morphological features in PPE cells may provide useful markers for detecting malignant cells in conventional cytology as the gold standard of cancer diagnosis. Among these features, the cell size, nuclear size and nuclear to cell size ratio have been regarded as the essential markers. In

this section, we present the results of a quantitative analysis of 2D morphology of PPE cells by cytopathological image slides for comparison to the previous 3D parameters. Specimens prepared from five PPE patients were performed by Drs. Heng Hong and Diana Dai of Department of Pathology of ECU and identified as P1 for patient 1, P2, P3, P4, and P8 in this study.

Cytological slides of PPE cells were stained by the conventional cytospin Diff-Quik (DQ) and Papanicolaou (Papa) liquid-based technique. The slides were evaluated by Drs. Hong and Dai for the known marker features associated with cancer cells. The slides were imaged under a conventional microscope with objective of 40x magnification and saved as tiff image files with examples shown in Figure 3.9. One can notice from the images that there are different types of cells present in each sample. These cells can be divided into two major groups, normal cells (NC) and cancer cells (CC). These cytology images were evaluated by Dr. Dai to mark numerous normal and cancer cells, which were quantitatively measured in terms of cell and nuclear areas for this study.

In addition to the aid of cytopathological staining, cytologists also utilize numerous cytological features such as single and cluster cells population, cell size, nuclear to cell ratio, vacuolated cytoplasm, multinucleation, increased nuclear size, nuclear shape, and presence of prominent nucleoli to distinguish between normal and cancer cells. Since the image slides are 2D, the above features can thus be measured as 2D quantities as well. In contrast to the normal cells, the cancer cells have a bigger size than normal cell and presence in two or more cells population with a relatively high nuclear to cell ratio due to increased nuclear size. Multinucleation, vacuolated cytoplasm, nuclear irregularity, and presence of nucleoli are also common features of cancer cells.

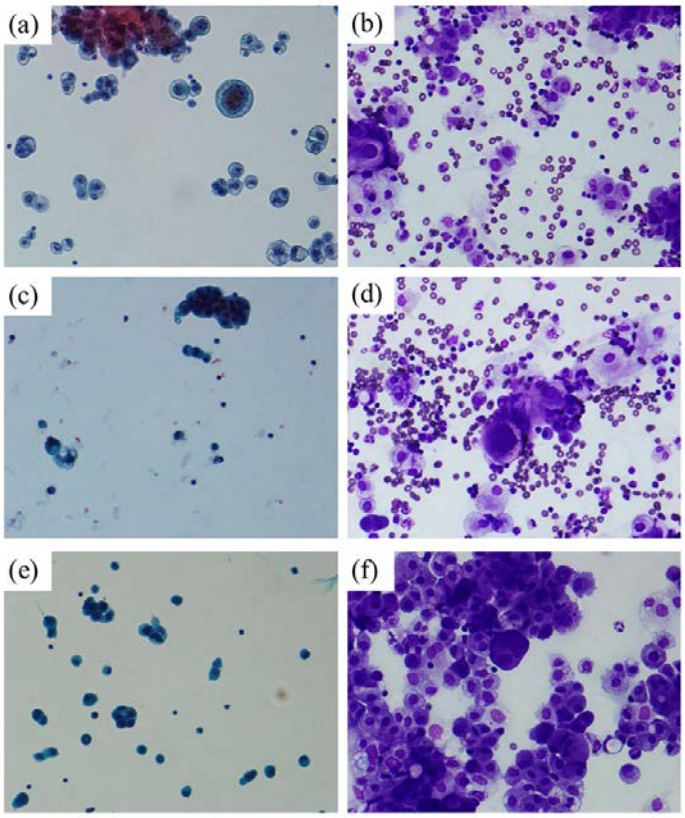

Figure 3.9 Samples of cytological Images 40x shows clusters of tumor cells and, (a) Diff-Quik stain, and (b) Papanicolaou stain P1, (c) Diff-Quik stain, and (d) Papanicolaou stain P2, (e) Diff-Quik stain, and (f) Papanicolaou stain P5.

3.7 2D feature extraction

In this study, we used an image analysis software named (Fiji). Fiji is an ImageJ distribution focused on biological image applications [76], to calculate selected cells' 2D morphological features (cell area (A_c) and associated nuclear area (A_n)) from these images. ImageJ is an open-source image processing and analysis software developed in Java source code [77]. It can be used to visualize, process, edit, and analyze many supported types of single and multiple (stack) of image files in addition to all image processing functions such as contrast manipulation and sharpening. Also, ImageJ has the ability to calculate area and pixel value statistics of user-defined selections. Spatial dimension calibration is available to provide measure distances in units such as micrometers using

micrometer rule line slides [78].

To measure structures in the unit of μm, an image of micrometer line slide was acquired and loaded into ImageJ by either drag the image file icon to drop it into ImageJ window or select "File-Open" drop-down menu and navigate to the file location to open the image. Then we used the straight line selection tool in ImageJ to draw a selection line on stage micrometer image for known distance. After that, we used the "Analyze-Set Scale" menu to assign the number of line pixel to a known distance and save the scale with appropriate unit. To use the same scale with all opened images, check the "Global" option box. Figure 3.10 shows the ImageJ interface with the image of the stage micrometer magnified 40 times under the microscope. By that, all the images are spatially calibrated. The info bar provides image information with the distance unit.

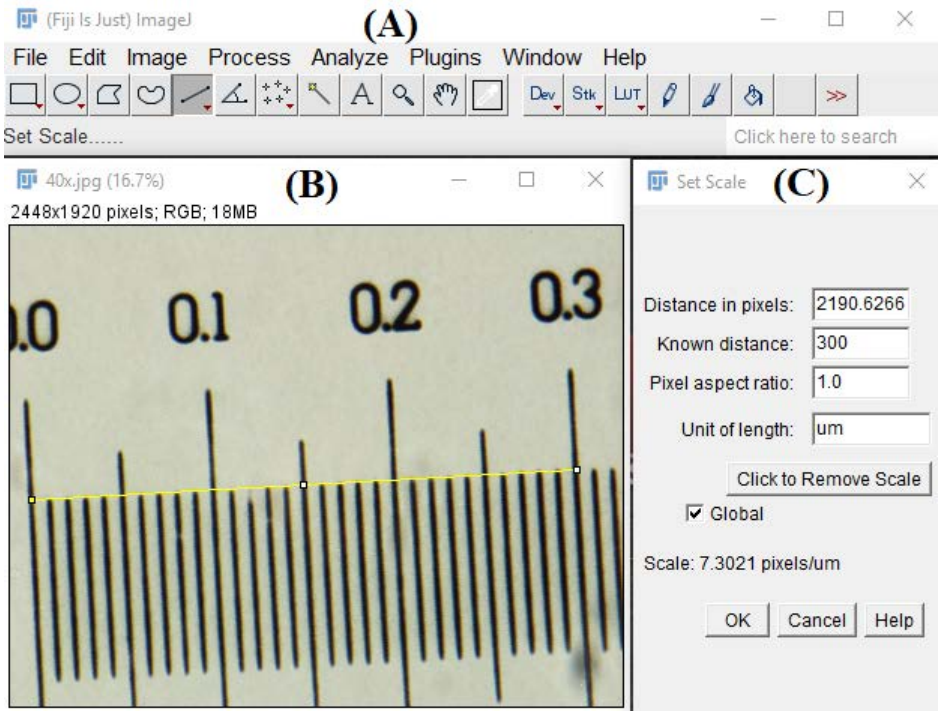

Figure 3.10 User interface of the ImageJ software (Fiji distribution), (A) main options menu, (B) loaded micrometer image, (C) and the set scale menu.

The next step is to load cytopathological images and use the rectangular selection tool to select the cell need to be analyzed. A freehand selection tool can be used to manually create a user-defined region of interest (ROI) around the single cell or nucleus and select "Analyze ⇒ Measure"

drop-down menu to take measurements under the manually defined region of interest, Figure 3.11. Images of single and multiple cells can be segmented into different ROIs using automated threshold method as described below. The thresholds can also be set to segment sub-cellular organelles such as nucleus or cells. Objects are detected by applying a threshold to the image, separating different regions based on pixel intensity histograms. The process starts by loading an image, then use the rectangular tool to select the ROI. Duplicate the image by running "Image ⇒ Duplicate" option and leave the original image untouched. Afterward, change the image format through "Image ⇒ Type" option to a monochromatic and set pixel value. Then, use "Image ⇒ Threshold" option to choose a suitable threshold range for transform the image into a binary one with ROI "cut out" from the rest of the image. Finally, select the "Analyze ⇒ Measure" drop-down menu to take spatial measurements given by the total pixel number in the automatically defined ROI (Figure 3.12).

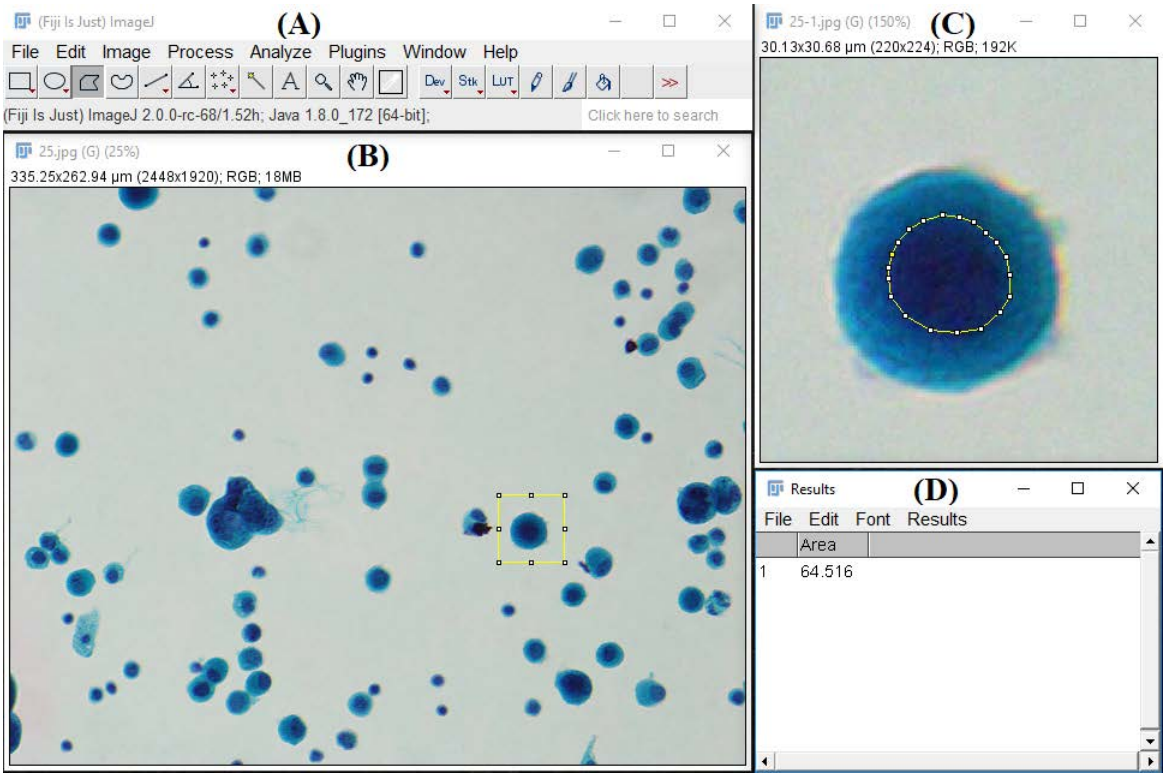

Figure 3.11 User interface of the ImageJ software (Fiji distribution), (A) main options menu, (B) loaded cytological slides images, (C) selected ROI using freehand selection tool on duplicate image, (D) measurement results.

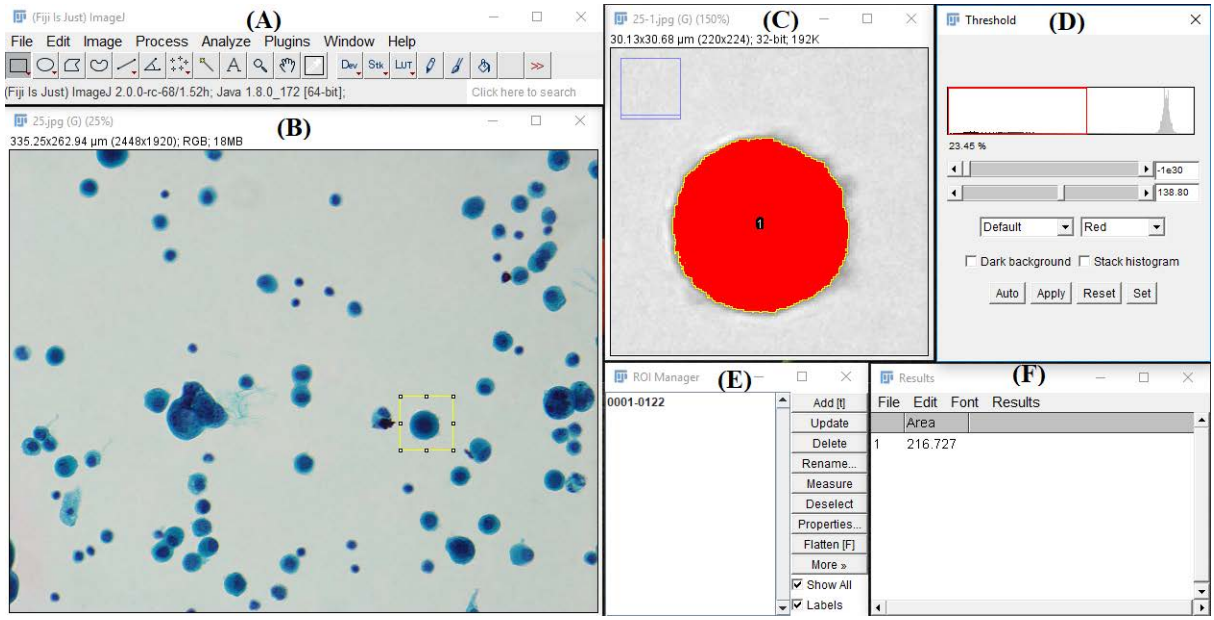

Figure 3.12 User interface of the ImageJ software (Fiji distribution), (A) main options menu, (B) loaded cytological slides images, (C) selected ROI using automated threshold method on duplicate image, (D) threshold options menu with pixel histogram, (E) ROI manager menu, (F) measurement results.

3.7.1 Results

We have performed quantitative measurement of cell and nuclear area measurement on 560 PPE cells selected from the cytology slides acquired from PPE samples of 6 patients. The distribution of normal cells (NCs) and cancer cells (CCs) of the measured cells are listed in Table 3.6. The labels of NC or CC for each measured PPE cell were given by a trained cytologist (Dr. Dai) and reviewed by another expert cytologist (Dr. Hong). The results of area measurements are summarized in Table 3.7 according to patients. Again, we can see the relatively large fluctuations in the measured values of cell and nuclear areas. Furthermore, it can be observed that cells of CC labels tend to have large cell and nuclear areas which indicate that cell size does correlate to cell type and morphology. The NCs are mostly include different portions of cells line the body's cavities, immune system cells, and white and red blood cells. Where, the CCs are likely cells falling off solid tumor tissues and diffuse in the body fluid. To gain insight, the GMM based clustering algorithm was employed for separating the measured cells into k=3 groups.

Table 3.6 The total number of both cancer and normal cells extracted from the cytology images of 6 patients

Patient ID	Cancer Cells (CC)	Normal Cells (NC)	Total cells/Patient
P1	65 (45.8%)	77 (54.2%)	142
P2	27 (28.4%)	68 (71.6%)	95
P3	24 (47.1%)	27 (52.9%)	51
P4	12 (50.0%)	12 (50.0%)	24
P5	48 (46.6%)	55 (53.4%)	103
P8	66 (45.5%)	79 (54.5%)	145
Total cells/Cluster	242 (43.2%)	318 (56.7%)	560

Table 3.7 Summary of area measurements of PPE cells imaged in cytology slides[a]

	NC			CC		
Patient ID	$A_c(\mu m^2)$ $mean \pm STD$	$A_n(\mu m^2)$ $mean \pm STD$	Ar_{nc} $mean \pm STD$	$A_c(\mu m^2)$ $mean \pm STD$	$A_n(\mu m^2)$ $mean \pm STD$	Ar_{nc} $mean \pm STD$
P1	44.21 ± 25.37	14.33 ± 13.38	0.31 ± 0.19	373.0 ± 341.0	157.4 ± 212.3	0.39 ± 0.16
P2	54.27 ± 70.00	23.13 ± 20.71	0.48 ± 0.12	298.3 ± 293.8	116.3 ± 123.9	0.41 ± 0.15
P3	36.08 ± 6.80	18.50 ± 3.81	0.52 ± 0.07	179.3 ± 61.75	89.76 ± 36.97	0.51 ± 0.15
P4	29.10 ± 4.45	16.26 ± 2.80	0.57 ± 0.10	146.2 ± 60.61	84.73 ± 44.07	0.57 ± 0.14
P5	35.36 ± 22.85	14.97 ± 6.29	0.46 ± 0.10	185.0 ± 173.8	80.92 ± 115.5	0.42 ± 0.12
P8	30.16 ± 4.55	17.96 ± 3.06	0.60 ± 0.07	125.8 ± 69.20	47.11 ± 32.97	0.40 ± 0.13

[a] NC= Normal cells, CC= Cancer cells. A_c and A_n are the cell and nucleus area in (μm^2). Ar_{nc} nucleus to cell area ratio.

3.7.2 GMM based clustering analysis

The 2D morphology parameters calculated from the ImageJ software are analyzed for cell classification by a developed MATLAB code uses the HC and GMM clustering techniques. These two techniques are used previously in the study of the 3D morphology. First, we loaded the calculated parameters; then we set k=3 to divide the data into three clusters based on the results we got from the clustering in 3D parameters space. After that, the algorithm starts with the HC method to prepare the data point for the GMM by assigning each point to its cluster based on the distances among them. The GMM clustering calculates the values for the mean and covariance matrix elements iteratively with MLPE and EMax to find the best Gaussian model fit the data. Table 3.8 lists the results of clustering with GMM method and corresponding numbers of NC and

CC cells. From this Table, one can observe that GMM perform well for clustering the cells based on their 2D morphology. Among the three clusters about 98% of the cells clustered in C1 are NC and 100 of the cells clustered in C3 are CC. C2 has a mixed portions of 19% NC and 81% CC.

We analyzed the clustering results using two parameters (A_c and Ar_{nc}) in 2D parameter space for data of measurement as shown in Figure 3.13. Figure 3.13 (A), shows the distribution of NCs and CCs in the two parameters space for each patient. One can find from this figure that most of NCs form all patients are found in the left side of the graph specifically in the region of small cell area and while the CCs are spread in bigger region of large cell area. Certain amount of NCs and CCs are overlapped in the middle region. In contrast, both types of cells span a wide range of nuclear to cell area ratio. Figure 3.13 (B), presents the distribution of both types of cells in term of clustering results, and clearly demonstrates the satisfactory results of the GMM clustering method. In comparison with measurements summarized in Table 3.7, these results provide educated guess on the nature of cells based the parameters of cell and nuclear area value.

Table 3.8 The total number of both cancer and normal cells extracted from the cytology images of three patients

	C1		C2		C3		
Patient ID	CC	NC	CC	NC	CC	NC	Total cells
P1	0	58 (100%)	53 (73.6%)	19 (26.4%)	12 (100%)	0	142
P2	3 (5.80%)	49 (94.2%)	18 (48.6%)	19 (51.4%)	6 (100%)	0	95
P3	0	22 (100%)	24 (82.8%)	5 (17.2%)	0	0	51
P4	0	12 (100%)	12 (100%)	0	0	0	24
P5	0	47 (100%)	46 (85.2%)	8 (14.8%)	2 (100%)	0	103
P8	2 (2.50%)	78 (97.5%)	64 (98.5%)	1 (1.5%)	0	0	145
Total cells/Cluster	271		269		20		560

[a] C1, C2, and C3 refer to three clusters.

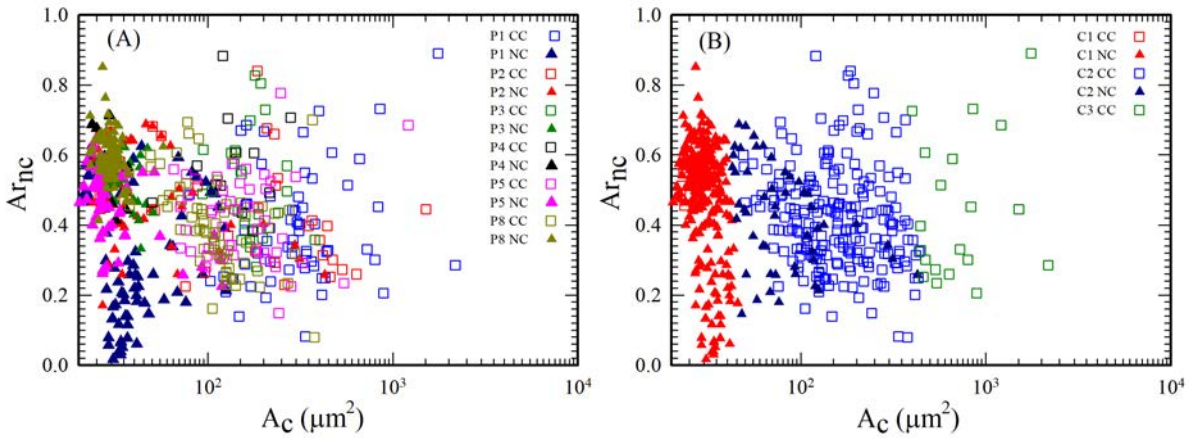

Figure 3.13 Scatter plots of cell area vs. nucleus-to-cell area ratio of selected cells from cytology images of three patients: (a) normal cells (NC) and cancer cells (CC) of patient marked in colors; (b) results of all analyzed cells partitioned by GMM clustering with k=3.

Chapter 4 Simulation and Analysis of p-DIs

In this chapter we will discuss the method of diffraction imaging simulation and results of p-DI calculations using PPE cell structures. Virtual PPE cells in terms of optical cell models (OCMs) were obtained from the 3D morphological structures by converting fluorescent light intensity values into refractive indices. Light scattering simulations were carried out by a validated ADDA method to obtain angle-resolved Mueller matrix elements. The elements were then projected in an angular range of choice onto a plane as an input plane image of light intensity distribution. We then applied a commercial software of ray-tracing to model the imaging unit for calculation of p-DI pairs by tracing the scattered light as rays from the input plane to an image plane, which provides a markedly fast approach for simulation of an objective based diffraction imaging process. The simulation results were used to study the correlation between cell morphological features and classification of calculated p-DI data by a clustering algorithm and understand cell profiling and classification by the measured p-DI data to be presented in the next chapter. Implication on profiling of the PPE cells by the p-DI based cell assay method will be discussed.

4.1 Construction of optical cell models

To simulate accurately the spatial distribution of scattered light, we need to develop an optical model of for a realistic cell structure or OCM. An OCM is essentially a 3D array of voxels that carry values of refractive index (RI) characterizing the molecular dipoles induced by the incident light wavefields. For this dissertation research, we employed a previously developed method to build OCMs from the CIMA output files by converting fluorescent light intensity into RI [59, 79].

As described in previous chapter, CIMA is a MATLAB code that can read a confocal image stack of a double stained cell and segment all pixels into four organelle groups of background (pixels outside the cell), cytoplasm, nucleus and mitochondria. After segmentation, multiple slices are interpolated between raw image slice to make cubic voxels. Then the 3D structure of the cell is established to store the intensities of fluorescent molecules tagging various biomolecules in the organelles into the output data files. Three of them will be used to build OCM. The first file was named "cell.mat" that contains 16-bit integers as the voxel values in a 3D data array. The 16-bit integers are coded for the organelle type in different ranges and the 12-bit pixel values in three color channels of the fluorescence intensity in each voxel. For example, background voxel values outside the cell membrane are set to 0 while the mitochondrial voxel values as indicated by the green channel fluorescence are set in a range from 5000 to 9095. The membrane voxel values of cell and nucleus are set to integer 20000 and 1. The second and the third data files store the fluorescence intensity in green channel and red channel for cytoplasm voxels after interpolation.

Construction of an OCM requires calculation of RI values for each pixel using the fluorescence intensity and output the 3D distribution of RI values as two input data files for ADDA simulation that are named as "Geometry_files.dat" and "refractive_index.txt". These calculations were carried out by another MATLAB script that reads the three files output from CIMA and requires the user to define RI parameters for host medium, cytoplasm, mitochondria, nucleus, nuclear membrane, cytoplasm membrane, n_{host}, n_{c0}, $n_{m,av}$, $n_{n,av}$, $n_{c,mem}$, and $n_{n,mem}$ respectively. Then the script code generates the relative RI values by dividing with the host medium RI and related domain values as 8-bit integers in the two output RI data files. Moreover, the voxel size used in the RI calculations is used as the dipole size used for execution of ADDA. The following equations are used to obtain RI values for each voxel at $\mathbf{r}$ location in a cellular organelle from the fluorescent intensity values

$$n_\alpha(\mathbf{r}, \lambda) = n_{c0} + b_r F_r(\mathbf{r}) + b_g F_g(\mathbf{r}) \quad \forall \mathbf{r} \in \Omega_\alpha \tag{4.1}$$

where $F_r(\mathbf{r})$ or $F_g(\mathbf{r})$ are intensity of the fluorescent dyes tagged to the biomolecules inside nucleus or mitochondria. The subscript $\alpha = c$ for cytoplasm voxels, $\alpha = m$ for mitochondria voxels and

$\alpha = n$ for nucleus voxels. The constant n_{c0} is the base value of RI all cellular organelles and b_r or b_g is respectively the coefficients to convert fluorescent intensity into RI for the voxel. Note that $F_g(\mathbf{r}) = 0$ inside nuclear volume or Ω_n while $F_r(\mathbf{r}) = 0$ inside mitochondrial volume or Ω_m. We should note here that in the above equation, we have utilized the widely accepted assumption that RI values of intracellular organelles depend linearly on the dry-mass or concentration of biomolecules [80, 81].

Based on the above equation, the mean values of RI for mitochondrial and nuclear voxels can be calculated as follows. First, let's denote the mean value of RI for mitochondria or over the mitochondrial voxels as $n_{m,av}$ which can be written as

$$n_{m,av} = n_{c0} + b_g \left\{ \frac{1}{N_m} \sum_{i=1}^{N_m} F_g(\mathbf{r}_i) \right\} = n_{c0} + b_g F_{gm,av} \tag{4.2}$$

where $\mathbf{r}_i$ refers to *i-th* mitochondrial voxel with total number of N_m, $F_{gm,av}$ is the mean value of green fluorescence intensity over all mitochondrial voxels. In the above equation, we utilized the fact that F_r is set to 0 for all mitochondrial voxels. This leads to the equation for the coefficient b_g expressed as

$$b_g = \frac{n_{m,av} - n_{c0}}{F_{gm,av}} \tag{4.3}$$

Similarly, we can obtain the mean value of RI for nucleus as

$$n_{n,av} = n_{c0} + b_r \left\{ \frac{1}{N_n} \sum_{j=1}^{N_n} F_r(\mathbf{r}_j) \right\} = n_{c0} + b_g F_{rn,av} \tag{4.4}$$

or

$$b_r = \frac{n_{n,av} - n_{c0}}{F_{rn,av}} \tag{4.5}$$

where N_n is the total number of nuclear voxels and $F_{rn,av}$ is the mean value of red fluorescence intensity over the nuclear voxels. Once the two model parameters of b_g and b_r are determined from

the given values of $n_{m,av}$, $n_{n,av}$ and n_{c0}, equation 4.1 can be rewritten as

$$n_\alpha(\mathbf{r},\lambda) = n_{c0} + (n_{n,av} - n_{c0})\frac{F_r(\mathbf{r})}{F_{rn,av}} + (n_{m,av} - n_{c0})\frac{F_g(\mathbf{r})}{F_{gm,av}} \quad \forall \mathbf{r} \in \Omega_\alpha \tag{4.6}$$

For nucleus, RI of voxels becomes

$$n_n(\mathbf{r},\lambda) = n_{c0} + b_r F_{rn}(\mathbf{r}) \quad = n_{c0} + (n_{n,av} - n_{c0})\frac{F_{rn}(\mathbf{r})}{F_{rn,av}} \tag{4.7}$$

where F_{rn} refers to the red fluorescence intensity in nuclear voxels. For mitochondrial voxels, we find

$$n_m(\mathbf{r},\lambda) = n_{c0} + b_g F_{gm}(\mathbf{r}) \quad = n_{c0} + (n_{m,av} - n_{c0})\frac{F_{bm}(\mathbf{r})}{F_{gm,av}} \tag{4.8}$$

where F_{gm} refers to the green fluorescence intensity in mitochondrial voxels. For cytoplasm, RI of a voxel at $\mathbf{r}$ becomes

$$n_c(\mathbf{r},\lambda) = n_{c0} + (n_{n,av} - n_{c0})\frac{F_{rc}(\mathbf{r})}{F_{rn,av}} + (n_{m,av} - n_{c0})\frac{F_{gc}(\mathbf{r})}{F_{gm,av}} \tag{4.9}$$

where F_{rc} and F_{gc} refer to the red and green fluorescence intensities in cytoplasm voxels. The mean RI value over all cytoplasm voxels are given by

$$\begin{aligned} n_{n,av} &= n_{c0} + b_r \left\{ \frac{1}{N_n} \sum_{k=1}^{N_n} F_r(\mathbf{r}_k) \right\} + b_g \left\{ \frac{1}{N_n} \sum_{k=1}^{N_n} F_r(\mathbf{r}_k) \right\} \\ &= n_{c0} + (n_{n,av} - n_{c0})\frac{F_{rc,av}}{F_{rn,av}} + (n_{m,av} - n_{c0})\frac{F_{gc,av}}{F_{gm,av}} \end{aligned} \tag{4.10}$$

where N_c is the total number of cytoplasm voxels and $F_{rc,av}$ ($F_{gc,av}$) is the mean value of red (green) fluorescence intensity over the nuclear voxels. It should be noted that all RI values in the above equations can be regarded either as absolute value or relative values against that of the host medium (n_h) by dividing both side of equations with n_h. Examples of confocal image and segmented slices are shown in Figure 4.1 (a), and (b) for one PPE cell. In addition to the three intracellular organelles

of cytoplasm, mitochondria and nucleus, We have also investigated OCM improvement by adding artificially an organelle of lysosome that is also important to light scattering [82]. This has been achieved by adding a distribution of small spheres inside the cell to study the effect of the lysosomes. The radii of these spheres are distributed according to a Gaussian function and embedded in the cytoplasm as shown in Figure 4.1 (c) for cell with lysosomes.

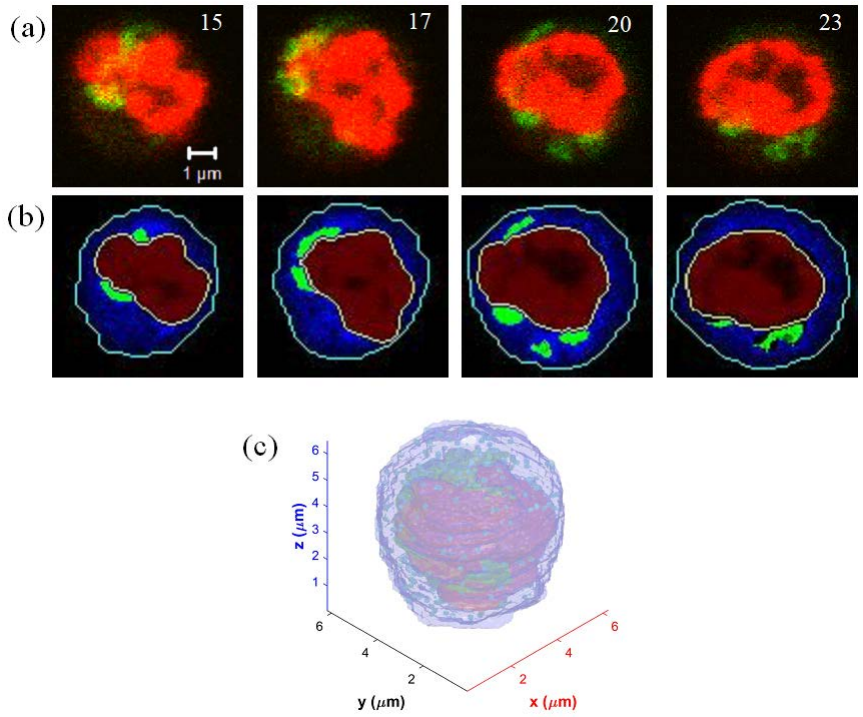

Figure 4.1 (a) Selected confocal image slices acquired from PPE cell. The red and green channels store Syto-61 and MitoTracker Orange intensities respectively. (b) present segmented slices with nuclear region in red pixels of intensity F_r, mitochondria in green pixel of intensity F_r and cytoplasm in blue. Each slice is labeled by its sequence number in the image stack and bar = 1 μm. (c) is perspective view of the 3D reconstruction on the same cell with nuclues colored in red, mitochondria in green, and artificial lysosomes normally distributed in the cytoplasm with light blue.

4.2 Diffraction imaging simulation

With an OCM, the process of light scattering and recording spatial distribution of scattered light is simulated in two steps. First, we apply a rigorous wave-optic model of discrete dipole

approximation (DDA) [22, 83] to obtain the angular distribution of light scattering by a biological cell represented by its OCM in space. Then the distribution of scattering light is converted into calculated p-DI pairs by projecting and ray-tracing the ADDA output data to simulate the propagation of coherent light through the imaging unit used in a p-DIFC system as described in next chapter. Despite the fact that ray-tracing method is based on the geometric optical model, it has been shown that this method of diffraction imaging simulation is valid in producing calculated p-DI data with diffraction patterns identical to the measured p-DI data in the case of single and aggregated spheres [84, 85].

4.2.1 ADDA simulation

ADDA is a flexible method with regards to the definition of the scatterer through many different options. Some of these options are related to the geometry of the scattering problems like reference frame, and orientation of scatterer and incident beam. Three different reference frames are offered by ADDA: laboratory, particle, and incident wave reference frame. The laboratory frame is the default one, and other reference frames are specified relative to it along with all input parameters. Also, the scatterer can be oriented in any direction relative to the laboratory frame. To minimize the size of the computational grid, ADDA simulates light scattering in the particle reference frame. This frame naturally corresponds to the particle geometry and symmetries. The incident wave reference frame is defined by setting the propagation direction along the z-axis, and it can be used to define the scattering plane and angles. ADDA embeds the scatterer in a rectangular computational box called grid. The grid is divided into small identical cubes named a dipole; its size should be very small compared to the incident wavelength. The size and the number of the dipole in the grid can be initialized manually using a command line option or automatically when the scatterer geometry is read from a file. Each dipole in the grid should be assigned a refractive index, and a dipole with refractive index equal to 1 is a void dipole [21].

The input "geometry_files.dat" of ADDA defines the RI morphology of scatterer whose effect on light scattering is to be simulated by ADDA. For a scatterer of multi-domain (coded) RI values,

this file contains a section of dipoles that defines the RI domains and the computational grid contains the scatterer. ADDA automatically places the minimum rectangular computational box or grid around all dipoles and centers it as described in ADDA manual [21]. A void dipole is the one with RI domain value of 0. Any lines in the dipoles section with RI domain values equal to 0 is ignored by ADDA. In ADDA execution, one first determines the wavelength of incident light in host medium, then extracts the pixel size of confocal image slices in x-y plane and number of voxels along the x-axis to enclose the cell from CIMA to define the grid size and number of grid cube or dipoles. The other two dimensions will be scaled according to the voxel array defined in the "Geometry_files.dat". To pass accurate information on scatterer's structure in terms of RI distribution to ADDA, one has to ensure that the dipole or cube size in ADDA execution must equal to the voxel size.

ADDA output the results of scattering simulation in terms of 4×4 Mueller-matrices $S_{ij}(\theta_s, \phi_s)$ for $i, j = 1, 2, 3, 4$ with θ_s and ϕ_s as the scattering polar and azimuthal angles. This requires to use scattering parameters file and to store scattering grid information for ADDA. The parameter file defines the range of angles the Mueller matrix will be computed for, and how many angle steps will be used sweeping across the θ_s and ϕ_s angles. As such, it is a very good idea to keep these two angles ranges as close to the desired input plane field-of-view (FOV) area as possible as calculating this grid is extremely intensive computationally. The output data of a successful ADDA simulation is saved into a file in the results folder. This file is the Mueller matrix scattering grid for the defined range of angles and serves as the sole input to create diffraction images. The element $S_{11}(\theta_s, \phi_s)$ was projected onto a plane intercepting with the x-axis as shown in Figure 4.4 as a simulated diffraction image of the side scatters within a half-cone angle of about 30°. In projecting the S_{11} element towards a pixel located at $\mathbf{r}_p$=(-x_0,y,z) both effects of incident angle of $\mathbf{r}_p$ and distance to the projected plane are considered. The pixel intensity I of the scattered light can be written as

$$I(y, z) = \frac{|\cos\phi_s \sin\theta_s|}{x_0^2(1 + \tan^2\phi_s + \dfrac{1 + \tan^2\phi_s}{tan^2\theta_s})} S_{11}(\theta_s, \phi_s). \tag{4.11}$$

where x_0 is the distance of the projection plane to the origin which was eliminated after image normalization [44].

4.2.2 ADDA performance

To validate the ADDA code, several simulations have been carried out for homogeneous spherical particles at various radii with properties similar to biological cells. These particles were generated via two different methods. The first method is using the ADDA predefined shape of sphere, where the second method is using externally generated shape of sphere using MATLAB algorithm developed for this purpose. We simulated the light scattering by spheres with radii of 1, 2, 3, 5 μm and RI of 1.588 + 0.00035 using Mie theory [86] and ADDA. The incident light assumed to a plane wave with $\lambda = 0.533\mu$m. With ADDA, the simulations used a value of dpl equal to 20. The ADDA results for the normalized Mueller matrix element of S_{11} have also been compared against the Mie theory results for spheres of the same radii. These results are shown in Figures 4.2 and 4.3. Notice that the relative errors in the ADDA results are more pronounced in the larger scattering angles, but overall the result is satisfactory. In addition to that, we believe that the range of errors seen in the figures are mainly come from the shape errors due to the representing particles with cubical dipoles. The angular range that our studies are concerned with, mostly between $\theta = 65^\circ$ and $\theta = 115^\circ$. The scatterer's size, orientation, incident wavelength, all of these parameters must be set when run ADDA simulation. An example script for running ADDA across single orientation has been included in Appendix(G) of this study.

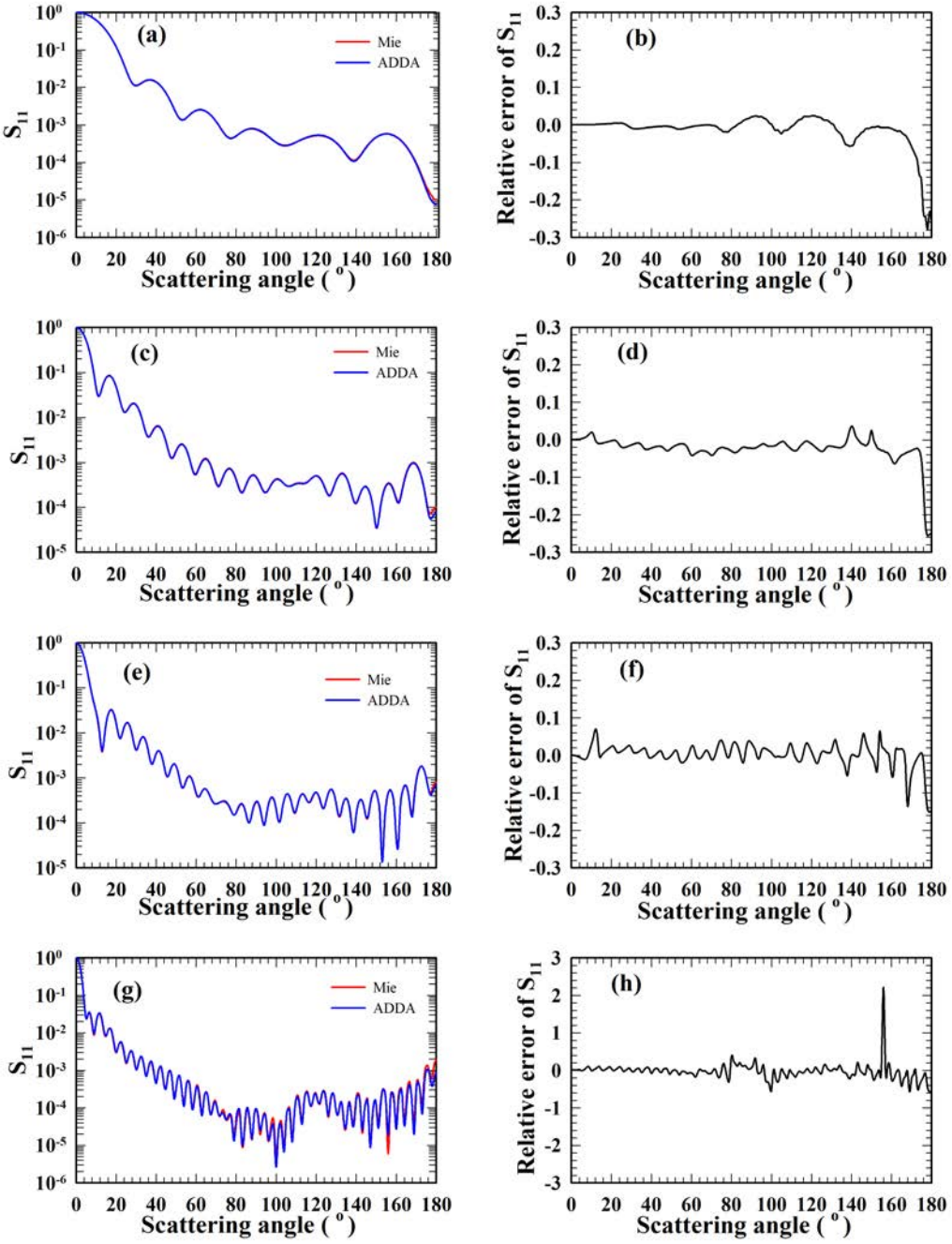

Figure 4.2 Comparison of S_{11} calculated by Mie theory and ADDA (pre-defined shape) for spheres of various radii and RI of 1.588+0.00035. (a) r=1 μm (c) r=2 μm (e) r=3 μm (g) r=5 μm. The relative errors are shown in (b), (d), (f), and (h), respectively.

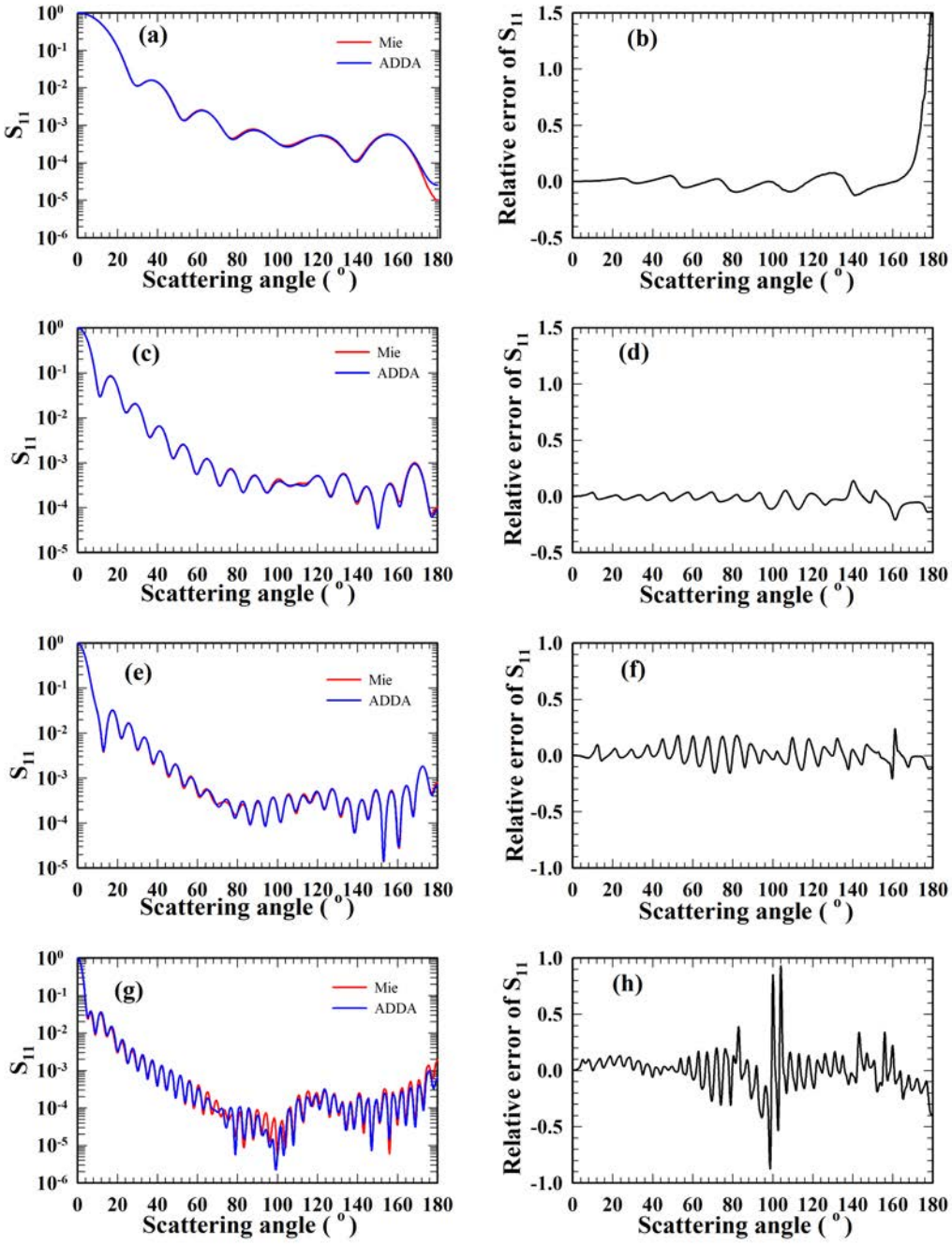

Figure 4.3 Comparison of S_{11} calculated by Mie theory and ADDA (externally generated shape) for spheres of various radii and RI of 1.588+0.00035. (a) r=1 μm (c) r=2 μm (e) r=3 μm (g) r=5 μm. The relative errors are shown in (b), (d), (f), and (h), respectively.

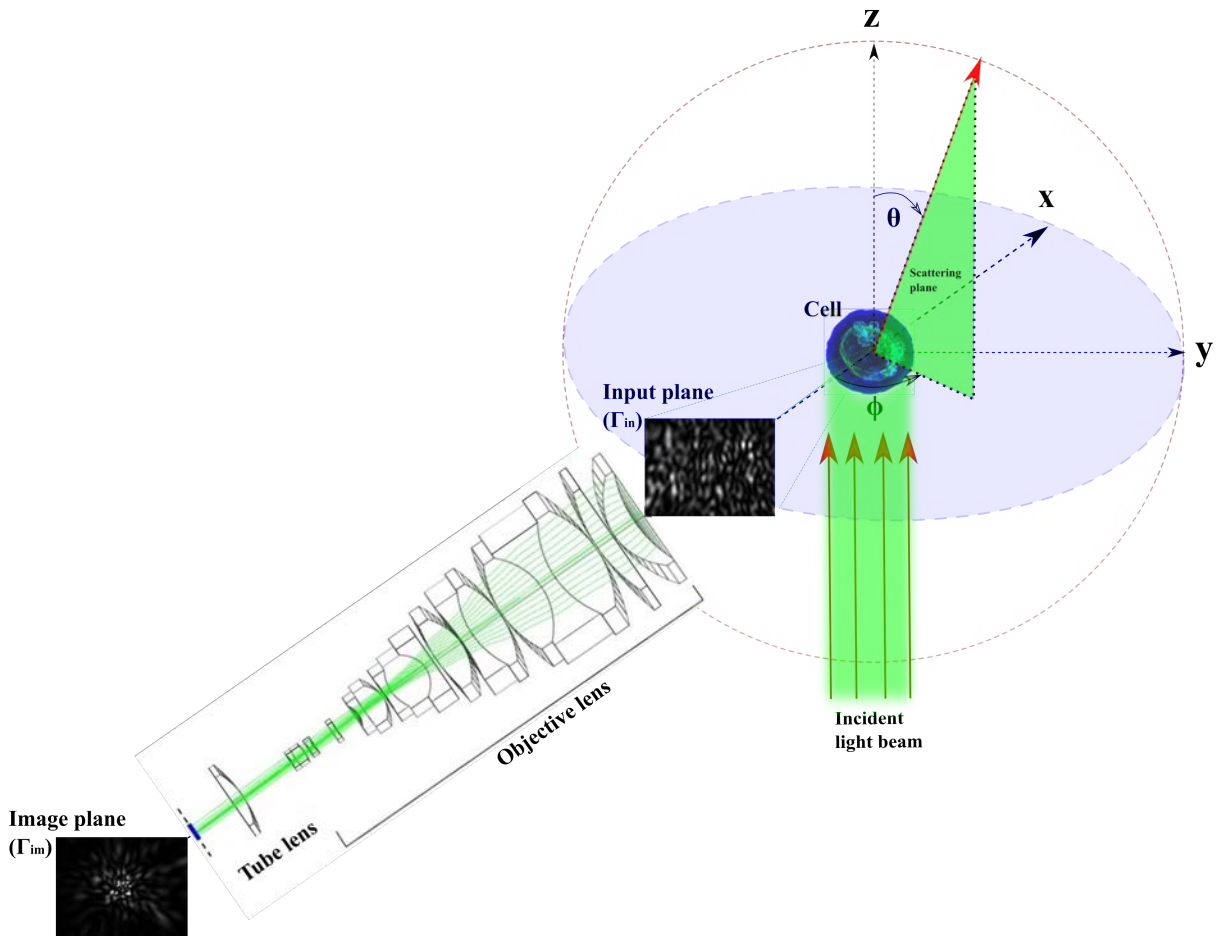

Figure 4.4 Diffraction imaging configuration.

4.2.3 p-DI Calculation and texture analysis

Two planes are defined in our simulation process to obtain a diffraction image (DI) comparable to the experimental configuration as shown in Figure 4.4. First, input plane Γ_{in}: this is a plane perpendicular to the x-axis (center axis of scattered light cone measured by a DI imaging unit) corresponding to a "virtual" plane at 150 μ m away from the origin (center of scatterer) of coordinate system in which z-axis points to the incident beam direction and y-axis points to the direction of fluid flow. Second, image plane Γ_{im}: this is the plane corresponding to the focal plane of tube lens in the imaging unit where the DI is acquired by an imaging sensor. Note here that the imaging unit including the Γ_{im} or imaging sensor can be translated along the x-axis from the focusing position

with $\Delta x > 0$ indicating translation toward the scatterer or the origin. The last part of DI simulation is to first project the Mueller matrix elements produced by ADDA onto Γ_{in} followed by ray tracing using an optical design software (ZEMAX) to obtain DI on Γ_{im}. This process is carried out by a MATLAB code developed for this purpose.

Both incident light and scattered light are polarized in our diffraction imaging flow cytometer system. The polarizations directions are defined relative to the incident plane of $x-z$ with $z-axis$ as the propagating direction of incident light and $x-axis$ as that for the center of the scattering light cone. We denote the incident light beam's polarization as vertical (ver), horizontal (hor) and 45° relative to the horizontal $x-z$ plane. There are 6 combination of s-polarized DI and p-polarized DI with the 3 incident beam polarization. They are defined by the scattered light intensity with 2 subscripts: the first denotes the polarization of scattered light as s or p and second denotes the polarization of incident beam as ver, hor, or 45° between ver and hor. The followings equation 4.12 provide the calculation of p-DI from the Mueller matrix elements with I_0 as unpolarized incident light intensity.

$$\begin{aligned}
I_{p,hor} &= I_0((S_{11}+S_{12})+(S_{21}+S_{22})) \\
I_{s,hor} &= I_0((S_{11}+S_{12})-(S_{21}+S_{22})) \\
I_{p,ver} &= I_0((S_{11}-S_{12})+(S_{22}-S_{21})) \\
I_{s,ver} &= I_0((S_{11}-S_{12})+(S_{22}-S_{21})) \\
I_{p,45^\circ} &= I_0((S_{11}+S_{21})+(S_{13}+S_{23})) \\
I_{s,45^\circ} &= I_0((S_{11}-S_{21})+(S_{13}+S_{23}))
\end{aligned} \tag{4.12}$$

To quantify the amount of energy transferred between the co-polarized and cross-polarized components of light, we used the linear depolarization ratio measured at normal scattering angle ($\theta_s = 90^\circ$). The linear depolarization ratio δ defined as the ratio of the average pixel intensity of the calculated image in perpendicular polarization with respect to the incident polarization to the average pixel intensity of the calculated images in parallel polarization with respect to the incident polarization axis as defined below [87]

$$\delta_{L,ver} = \frac{\overline{I}_{p,ver}}{\overline{I}_{s,ver}}, \quad \delta_{L,hor} = \frac{\overline{I}_{s,hor}}{\overline{I}_{p,hor}} \quad \delta_{L,45^\circ} = \frac{\overline{I}_{p,45^\circ}}{\overline{I}_{s,45^\circ}} \tag{4.13}$$

where $\overline{I}$ is the average pixel intensity of an image with its subscripts indicating the polarizations of incident and scattered light.

The texture represents the statistical (spatial) arrangement of pixels or pictorial patterns [88], and gives meaningful information about the surfaces structural distribution [89]. Therefore, The textural features are used to quantify the properties of an image region by using space relations underlying the gray-level distribution of a given image and have general use in image classification [90]. One of the most common technique in texture feature extraction includes the computation GLCM as a second order texture measure. GLCM describes how many times of one gray-level appearing in a specified statistical linear relationship with other gray-level within the area of study. Several statistical parameters can be extracted from the GLCM. Some of these parameters are directly related to first order gray-level statistics concepts such as gray-level mean and variance while other parameters contain more complex textural information associated with multiple texture meanings [91]. In this study we employed up to 15 GLCM parameters (texture features) to define the texture of the DI's, Appendix F.

4.3 Effect of intracellular RI distribution among organelles

Five different OCM are introduced in this study to investigate the effect of intracellular organelles and their RI heterogeneity on the DIs using the same set of input parameters. These models are: (1) OCM$_{fl}$ is a fluorescence intensity based model as given by equation 4.1 and the parameters b_r and b_g were adjusted by the average RI of nuclear and mitochondrial voxels to examine the effects of organelle molecules on light intensity distribution in p-DI pairs. (2) OCM$_{pfn}$ is a partial fluorescence based model in which voxels of only nucleus have its RI values obtained by equation 4.1 while the voxels of other organelle types are set to have constants RIs. (3) OCM$_{pfm}$ is a partial fluorescence based model in which voxels of only mitochondria have its RI values obtained by

equation 4.1 while the voxels of other organelle types are set to have constants RIs. (4) OCM_{nf} is a no fluorescence based model in which the RI values are set to different constants as $n_c(\mathbf{r}) = n_{c,ave}$, $n_n(\mathbf{r}) = n_{n,ave}$, and $n_m(\mathbf{r}) = n_{m,ave}$. (5) $\mathrm{OCM}_{fl,lyso}$ is this model we artificially add one more important cell organelle which is the lysosomes. The lysosomes have major role in light scattering since they have RI that increase the heterogeneity of the cell.

We performed p-DI simulations with the first four types of OCMs built from confocal image based PPE cell structures of small, medium, and large for each of three incident beam polarizations. The results presented here are mainly on the variation of the average RI value $n_{n,av}$ for nuclear voxels between 1.390 and 1.51 in steps of 0.060 steps and average RI value $n_{m,av}$ for mitochondrial voxels between 1.490 and 1.550 while n_{c0} was fixed at 1.360 as shown in Table 4.1. The choice of these RI value ranges was based on results published in literature [36, 92, 93]. The off-focus distance Δx of the imaging unit in ray-tracing calculations was fixed to 150 μm, which is the same value used for p-DI measurement to facilitate the comparison with measured data. The wall-clock time T of light scattering simulation depends on the size of cell structures (small, medium, or large) and choices of RI values for parallel execution of ADDA code on 120 CPU cores on several nodes of a computing cluster at the Department of computer Science, ECU. For PPE of small cell size and same values of n_{c0} at 1.360 and $n_{m,av}$ at 1.490, T ranges from about 41 for $n_{n,av} = 1.39$ to 63 minutes for $n_{n,av} = 1.51$ while T ranges from 212 to 320 minutes for PPE of medium size and around 645 minutes for PPE of large size with the same RI variations. Compared to ADDA simulation, projection of the Mueller matrix elements to the input plane Γ_{in} and ray-tracing to the image plane Γ_{im} takes only around 10 minutes on a computer with one i7-870 CPU of 2.93GHz.

Table 4.1 Cell morphology and RI for simulated DIs using four different OCMs without the lysosomes.

Structure[a]	Number[b]	$V_c(\mu m^3)$[c]	$Vr_{nc}(\%)$[c]	$Vr_{mc}(\%)$[c]	n_0	$n_{n,av}$	n_{mav}
Small	3	[215.71, 407.07]	[27.2, 41.4]	[3.3, 4.7]	1.3607	[1.390, 1.510]	[1.490, 1.550]
Medium	3	[502.53, 770.39]	[29.7, 38.0]	[5.7, 13.9]	1.3607	[1.390, 1.510]	[1.490, 1.550]
Large	3	[502.5, 770.4]	[29.7, 38.0]	[1.8, 6.3]	1.3607	[1.390, 1.510]	[1.490, 1.550]

[a] Cell structure group based on classification results of 3D morphology parameters in chapter three.
[b] Number of cell structures used for each group.
[c] Cell volume, nuclear to cell volume ration, and mitochondrial to cell volume ratio range.

Table 4.2 Cell morphology and RI for simulated DIs $OCM_{fl,lyso}$.

Structure	# of p-DI pairs	$V_c(\mu m^3)$	$Vr_{nc}(\%)$	$Vr_{mc}(\%)$	$V_{lyso}(\mu m^3)$[a]	$Vr_{lysocyto}(\%)$	n_0	$n_{n,av}$	n_{mav}	n_{lyso}[b]
Small	72	(89.3, 543.0)	(33.1, 94.8)	(0.009, 5.7)	(0.11, 3.05)	(0.32, 0.47)	1.3607	1.390	1.53	1.45 ± 0.02
Medium	60	(290.2, 1560.9)	(11.9, 57.3)	(1.0, 3.5)	(0.11, 3.05)	(0.38, 0.45)	1.3607	1.390	1.53	1.45 ± 0.02
Large	39	(432.7, 5092.9)	(3.4, 65.9)	(0.002, 8.2)	(0.11, 3.05)	(0.39, 0.44)	1.3607	1.390	1.53	1.45 ± 0.02

[a] Lysosomes volume range.
[b] Lysosomes RI, mean ± STD.

Figures 4.5 to 4.8 show examples of calculated p-DI pairs with 6 combinations of nuclear and mitochondrial RI for ver, hor, 45° incident beam polarizations that are analyzed in detail below. By comparing the calculated p-DI pairs in these four figures to the measured data presented in next chapter, certain similarity in image textures can be identified to validate the realistic OCMs as defined in this study for simulation of diffraction imaging. We have performed clustering analysis of these p-DI data to obtain insight for analysis of measured p-DI data to be discussed in next chapter.

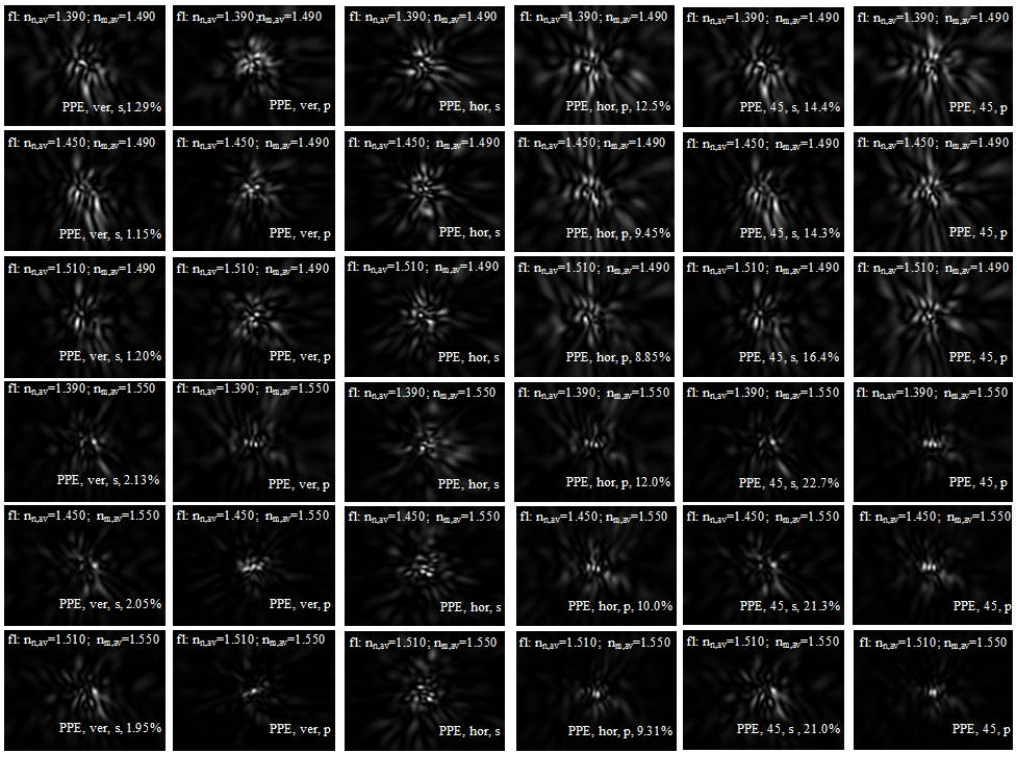

Figure 4.5 Normalized cross-polarized diffraction image (p-DI) pairs calculated by optical cell model OCM_{fl} with three cell structures with vertical, horizontal, and 45° incident polarization, $\lambda = 532$ nm and $\Delta x = 150\mu$m. Each pair is marked with averaged nuclear and mitochondrial RI, cell structure, incident and scattered polarizations and value of δ_L.

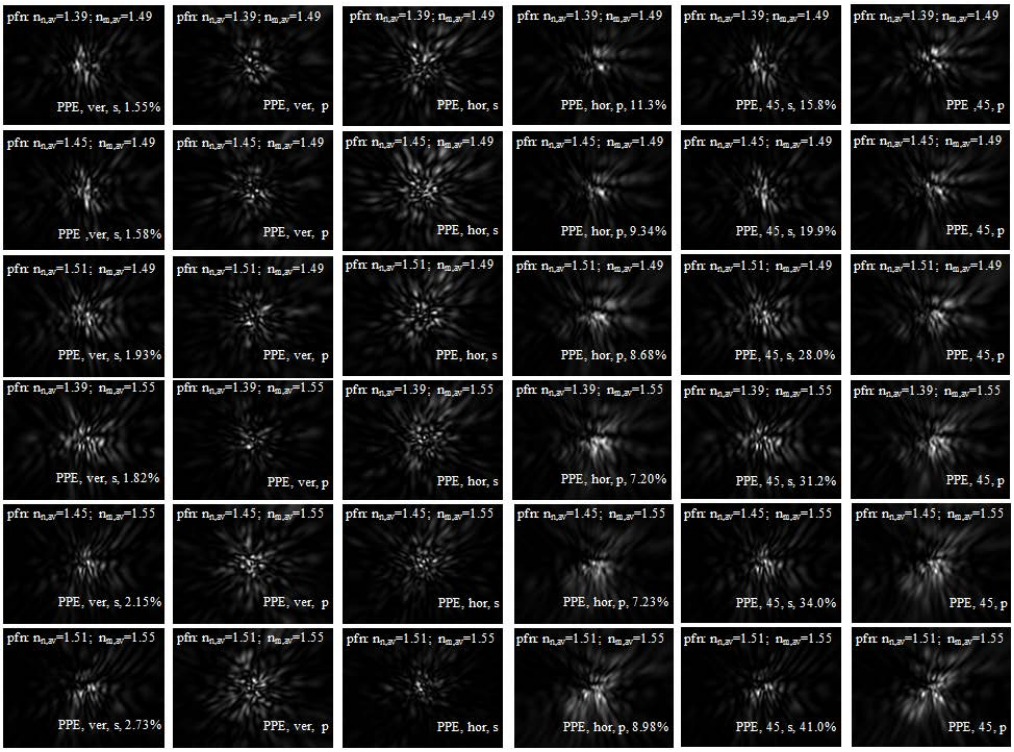

Figure 4.6 Same as Figure 4.5 except with different model (OCM_{pfm}).

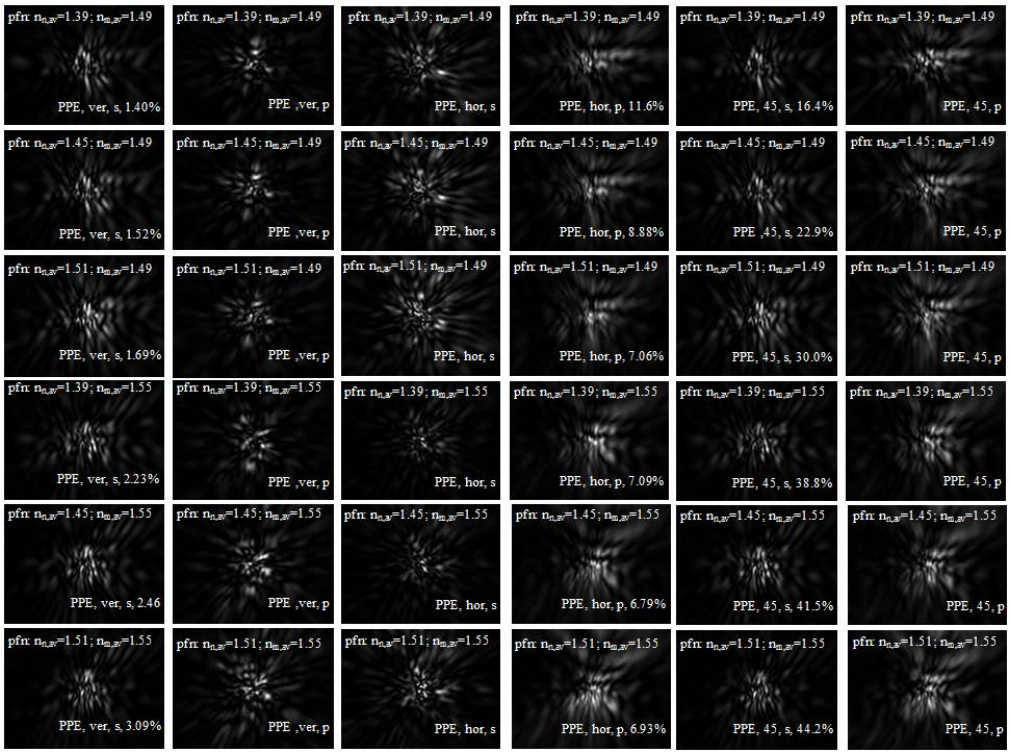

Figure 4.7 Same as Figure 4.5 except with different model (OCM_{pfn}).

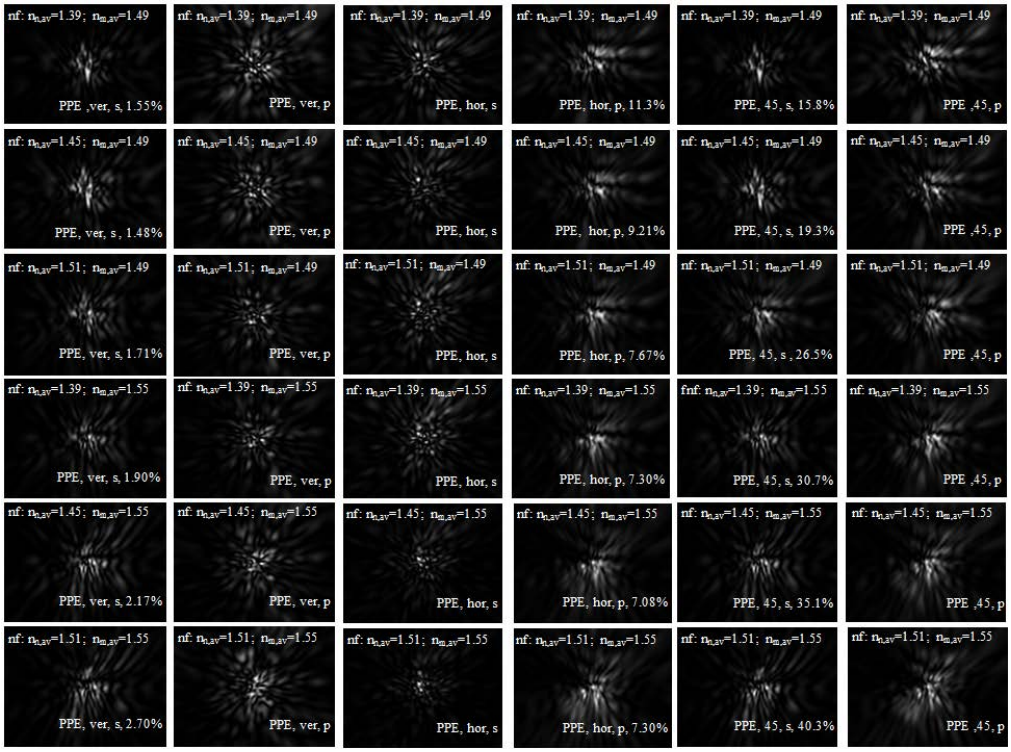

Figure 4.8 Same as Figure 4.5 except with different model (OCM_{nf}).

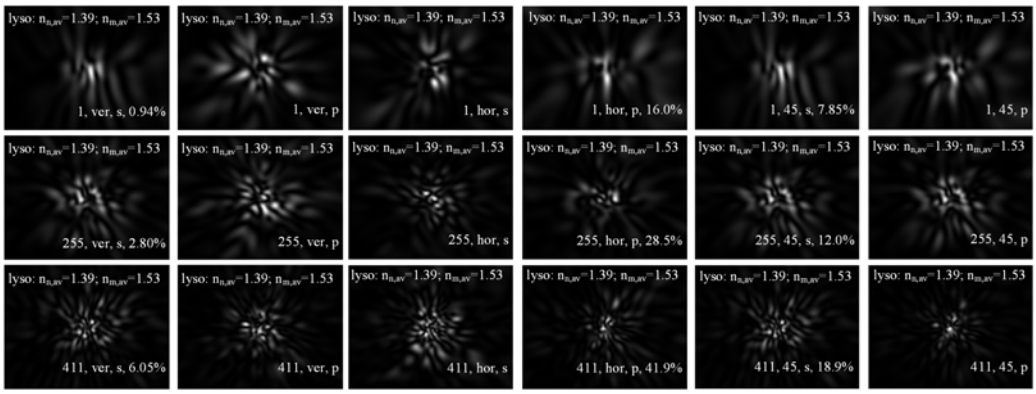

Figure 4.9 Same as Figure 4.5 except with different model ($OCM_{fl,lyso}$) with artificial lysosomes added. The lysosomes are Gaussian spheres with parameters mean size of $0.6 \pm 0.3\mu$m. The lysosomes refractive indices are 1.45 ± 0.02, and the lysosomes to cytoplasm ratio are around 0.4%

For quantitative comparison between the measured and calculated GLCM parameters, we applied the GLCM algorithm for characterization of image textures with four parameters selected for their capacities to quantify image textures and high performance to classify PPE cells by the measured p-DI data. We plot in Figures 4.10 to 4.15 the dependence of GLCM parameters on average nuclear and mitochondrial voxel RI values $n_{n,av}$ and $n_{m,av}$ in the three cell types of small, medium, and large for four OCM types with ver, hor, 45° incident polarizations. Each bar represents the mean value of a GLCM parameter obtained from calculated p-DI by the same OCM in three incident beam polarizations, which allow the estimation of fluctuations in p-DI data due to variation of cell structure size and the detection of scattered light.

It is clearly noticed in the case of all cell structures with all incident beam polarization that the OCM models yield GLCM parameters of p-DIs significantly different from each other when $n_{n,av}$ ranging from 1.390 to 1.510 and for $n_{m,av}$ increases from 1.490 to 1.550. The deviation of GLCM parameters of p-DIs by the OCM models indicate that the heterogeneity in RI values and the size of the cell structure can significantly modify the textures of DIs and these results show that the structures shape irregularity and RI heterogeneity inside cell organelles have more important roles than the average RI values in the spatial distribution of scattered light.

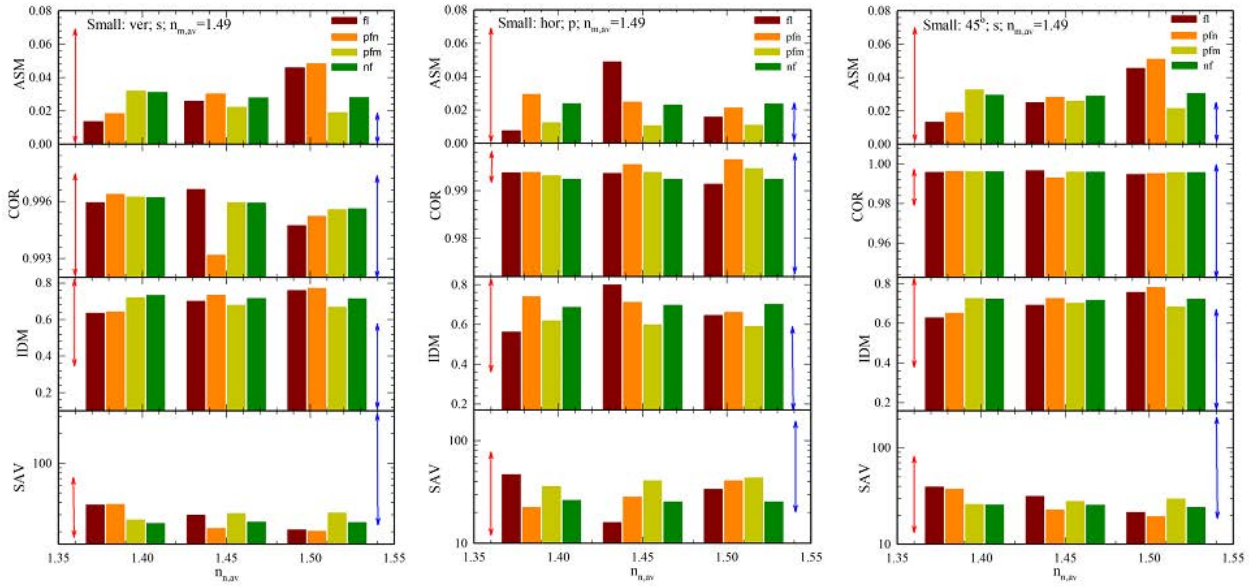

Figure 4.10 Selected four gray-level co-occurrence matrix (GLCM) parameter of s-polarized and p-polarized diffraction images (DIs) and vertical, horizontal, and 45° incident polarization vs $n_{n,av}$ in different optical cell models (OCMs) of small PPE and $n_{m,av} = 1.49$. The arrowed vertical lines in blue on the right and in red on the left indicate the parameter ranges of the measured data and calculated data with $OCM_{fl,lyso}$ for the same cell type respectively. The Bar colors are for visual guide

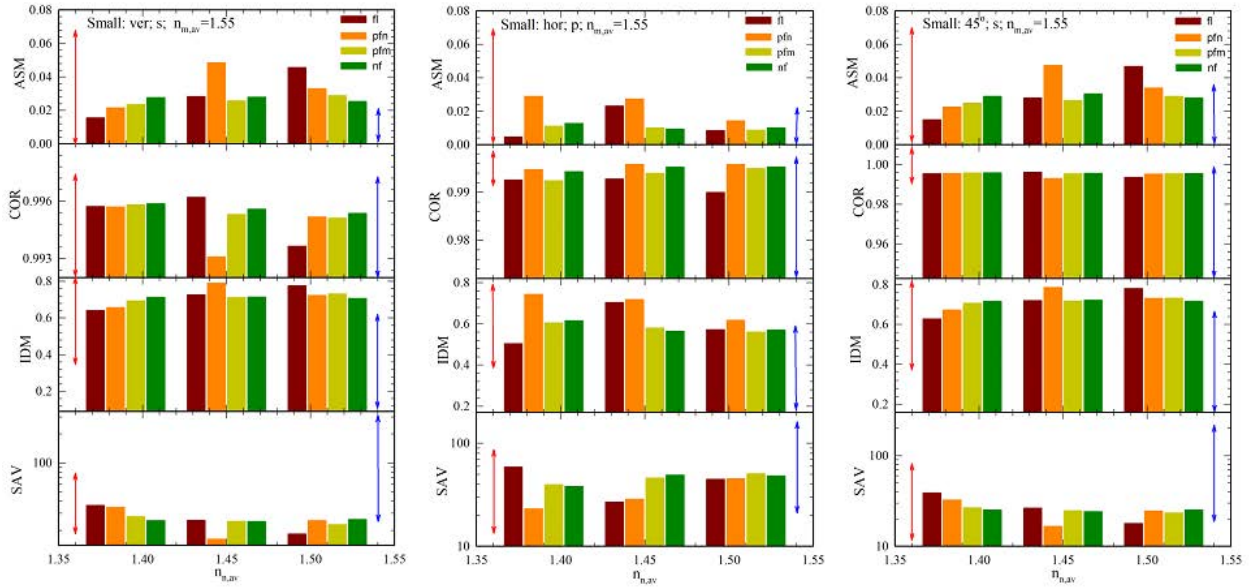

Figure 4.11 Same as Figure 4.10 except gray-level co-occurrence matrix (GLCM) parameters of diffraction images (DIs) calculated with $n_{m,av} = 1.55$.

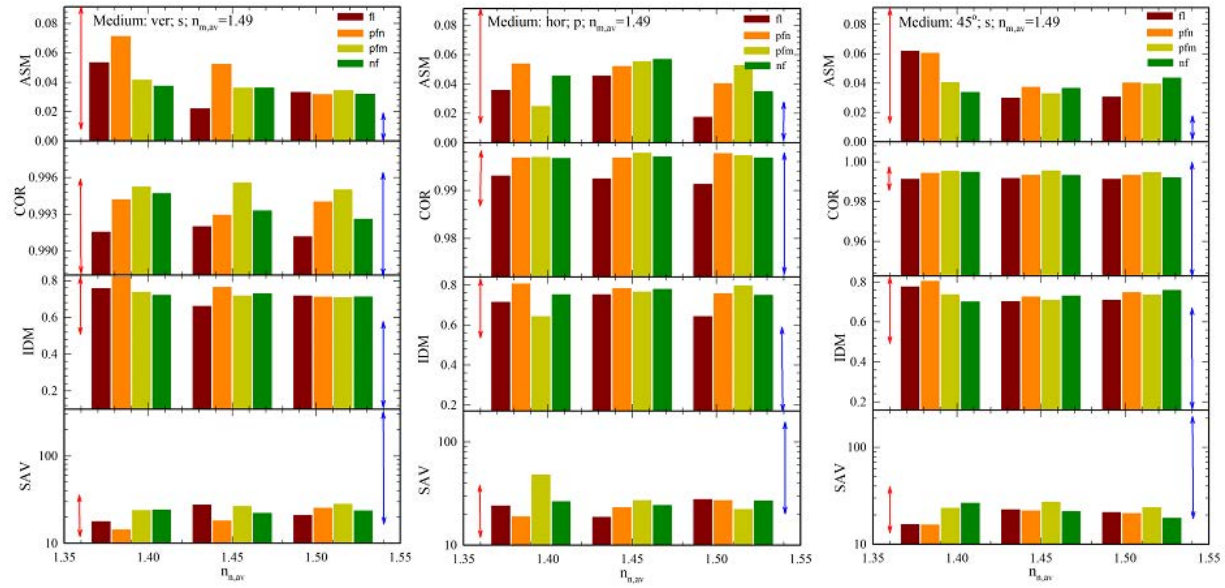

Figure 4.12 Same as Figure 4.10 except gray-level co-occurrence matrix (GLCM) parameters of diffraction images (DIs) calculated for medium size PPE and $n_{m,av} = 1.49$.

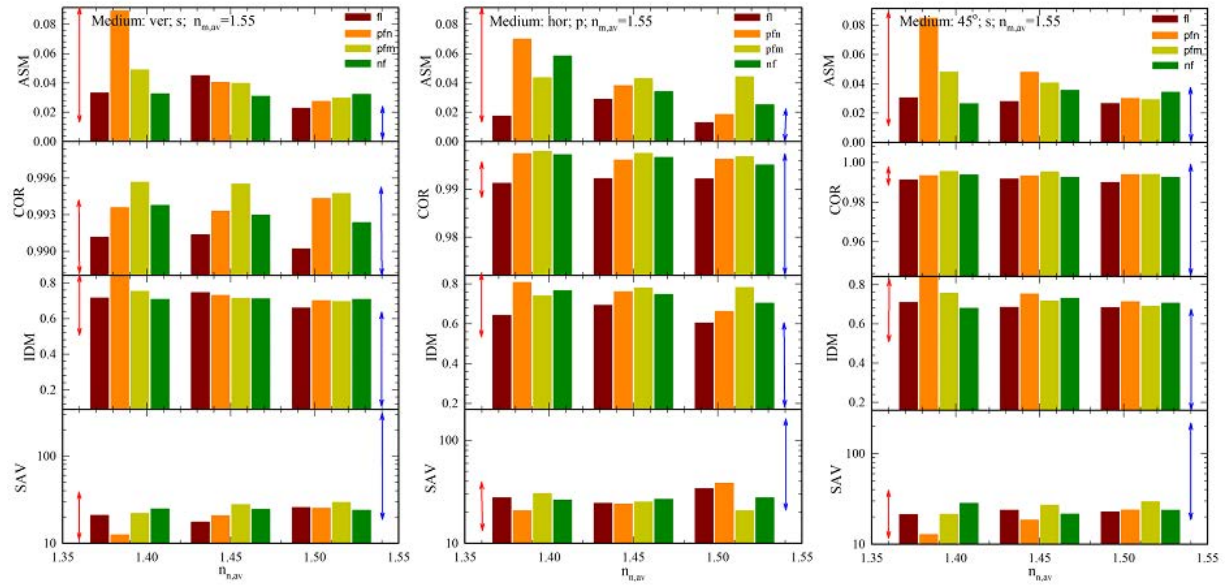

Figure 4.13 Same as Figure 4.10 except gray-level co-occurrence matrix (GLCM) parameters of diffraction images (DIs) calculated for medium size PPE and $n_{m,av} = 1.55$.

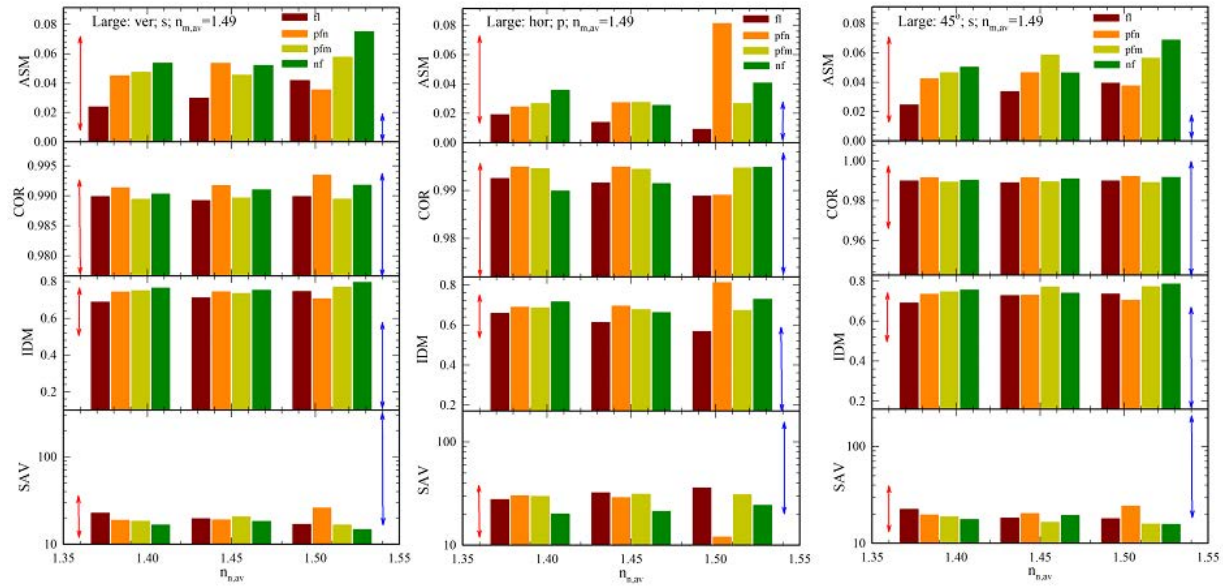

Figure 4.14 Same as Figure 4.10 except gray-level co-occurrence matrix (GLCM) parameters of diffraction images (DIs) calculated for large size PPE and $n_{m,av}$ = 1.49.

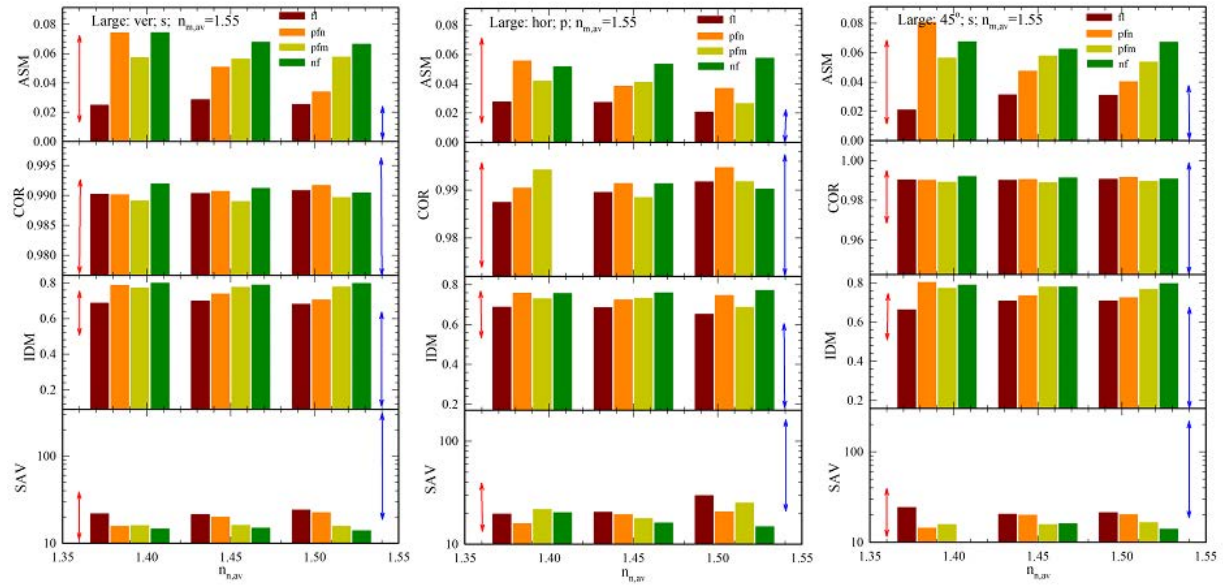

Figure 4.15 Same as Figure 4.10 except gray-level co-occurrence matrix (GLCM) parameters of diffraction images (DIs) calculated for large size PPE and $n_{m,av}$ = 1.55.

4.4 Clustering results

We used the combination of hierarchical and GMM clustering algorithms to classify the calculated p-DI data , or associated cells into three clusters named as C1 , C2 and C3. Each cluster was assigned to one cell types defined in Table 4.3 small, medium, and large based on the dominant type in a confusion matrix. To measure the performance of a classifier, a classification accuracy A is defined by the number ratio of correctly identified cell type to the total number of cells. To explore the possibility of cell classification according to the diffraction patterns only, we have investigated different combination of GLCM parameters with different combinations of GLCM parameters. The values of each parameter were normalized to the range of [0, 1] and this normalization scheme is denoted as "parameter norm.". The results in Table 4.3 present the confusion matrices of classifiers to divide cell structures related to the calculated p-DI data into three clusters with the value of μA defined as the mean value of A using 15 or 5 GLCM parameters by the parameter normalization scheme. These results show clearly that the value of μA in all 15 parameters classifier is higher than the 5 parameters classifier. But, the classification process in 15 parameters classifier is not good enough to separate the groups of medium and large cell types within C2 cluster as in 5 parameters classifier. Because the overfitting problem that usually happen with high dimensional GMM classification of relatively small amount of data.

Table 4.3 Confusion matrices of GLCM clustering for all cells.

Clusters[a]	C1	C2	C3	normalization scheme; P_i; $\mu_A \pm \sigma_A$ [b]
Small	20 (28.2%)	16 (22.5%)	**35** (49.3%)	parameter norm. (all 15 parameters) 60.7% ± 20.7%
Medium	3 (5.0%)	**29** (48.3%)	28 (46.7%)	
Large	0 (0.0%)	**33** (84.6%)	6 (15.4%)	
Small	**36** (50.7%)	35 (49.3%)	0 (0.00%)	parameter norm. (ASM, CON, COR, IDM, SAV) 57.8% ± 10.0%
Medium	19 (15.0%)	**32** (53.3%)	9 (31.7%)	
Large	6 (15.4%)	6 (15.4%)	**27** (69.2%)	

[a] Rows represent ground truth. Bold numbers represent assigned cell types for three clusters derived by the hierarchical + GMM classifier.

[b] P_i represents the set of normalized GLCM parameters used by the classifier, μ_A and σ_A are the mean value and standard deviation of clustering accuracy for 3 cell types.

Figure 4.16 presents examples of calculated p-DI pairs for different cell structure classified as C1, C2, and C3 where C1, C2, and C3 are mainly dominated by small, medium, and large cells

respectively. The first row in each cluster type presented in the figure show the p-DI for correctly classified cell types while the second row show the p-DI for misclassified cell types. By visual comparison of the p-DI, we found the texture and the pixel intensity changes among the three clusters are relatively large. The p-DIs of the small cell in C1 present less number of speckles than the medium and large cells in C2, and C3. Also, the cells in C1 showed DIs with largest speckles size than the DIs of the cells in C2, and the cells in C3 where the finest speckles are displayed.

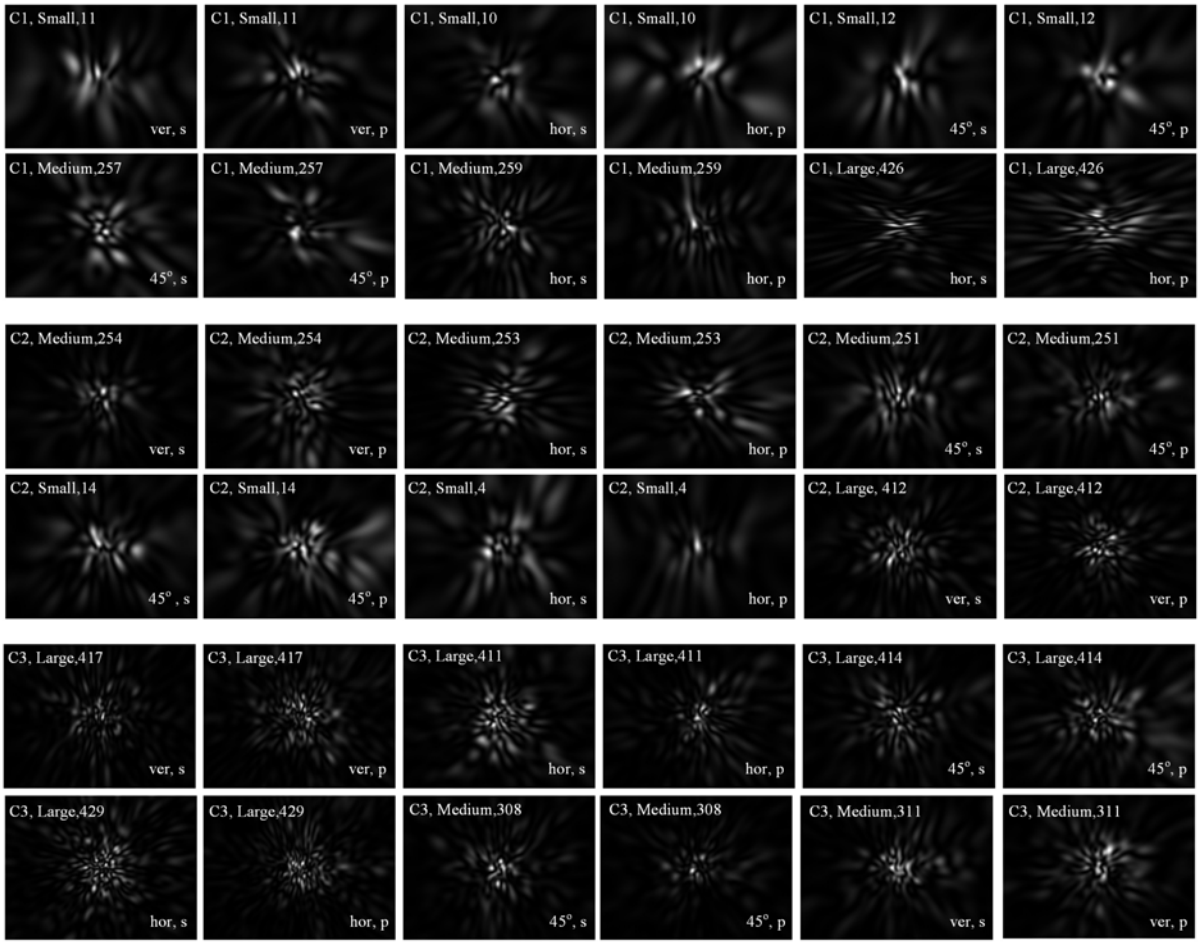

Figure 4.16 Clustering results of normalized p-DI pairs calculated by $\mathrm{OCM}_{fl,lyso}$ with different cell structures. Each images was labeled by cluster number, cell type, and reference number in the upper left corner, and the polarization of the incident and scatter light in the lower left corner

In Figure 4.17 , we present the normalized scatter plots of five GLCM parameter ASM, CON, COR, IDM, and SAV for cells having the second HC and GMM classifier in Table 4.3. One can see obvious differences in the distribution of the cells clustered in different groups. while some of the cells in each cluster overlap each other in the scatter plots, the cells in C1, C2 and C3 appear to have significant degree of spreads in their values of the selected GLCM parameters, which can be also noted from the box plot of these parameters in Figure 4.18. Taken together, the quantitative characterization of the DIs provides insight into the morphologic differences among the three types of cells. In Figure 4.18 we create a box plot for each cluster and for each of the selected GLCM parameter. The horizontal line inside the box is the sample median, the tops and bottoms of each box are the 25^{th} and 75^{th} percentiles of the samples, respectively. The distances between the tops and bottoms are the interquartile ranges. The whiskers are lines extending above and below each box. The whiskers are drawn to show the furthest observations within the whisker length, adjacent values. Two sample points beyond the whisker length are marked as 5^{th} and 95^{th} percentiles of the samples.

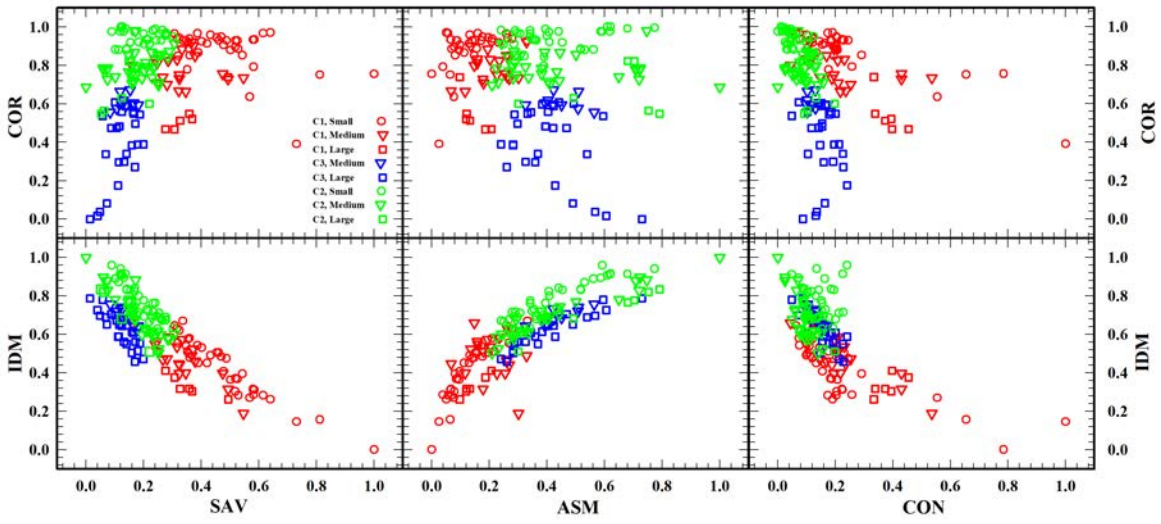

Figure 4.17 The distribution of the PPE cells as labeled by the cell type and cluster number in the space of selected GLCM parameters.

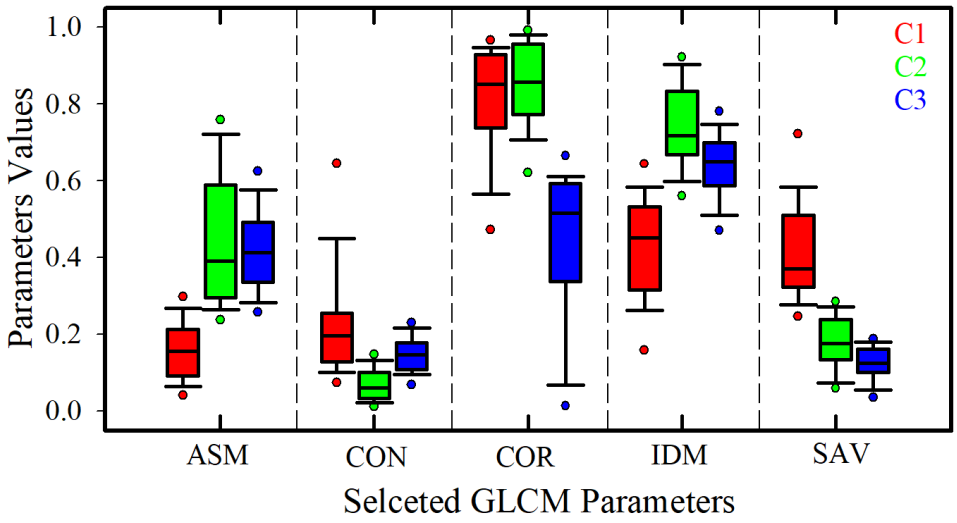

Figure 4.18 Comparison of five GLCM parameters extracted form the calculated p-DI and classified to three clustered HC and GMM

DIs in C1, mainly come from small cells, are dominating the lower end of the ASM parameter with less spread compared to DIs in C2 and C3 which clearly can be correlated to the low level of uniformity or less similarity of gray-level pixels of these images. For the CON parameter the DIs of small cells showed a wide range of local variation among the image pixels than DIs in C2, and C3, while the COR parameter have more consistent gray-level values for their pixels that is correlated to the large speckles areas in these images. Also, the DI of small cells showed less homogeneity in gray-level pixels in comparison to the DIs in C3.

On other hand, the DIs in C3, mostly are coming from the large cells show an increase in the level of ASM parameter than in C1. This means that these images are more uniform in term of the gray-level pixels values which correlate to the large number of small size speckles. Low level of CON and SAV values showed in these DIs are also due to the distribution of small size speckles. The DIs of C3 spread over the lower end of the COR parameters indicate the less consistent of gray-levels.

Chapter 5 Measurement and Analysis of p-DIs

The measurement of cross-polarized diffraction image (p-DI) pairs and results are presented in this chapter together with the method of polarization diffraction imaging flow cytometry (p-DIFC). The measured p-DI data were preprocessed to remove those by cellular debris and aggregated non-cellular particles followed by extraction of the image texture parameters using GLCM algorithm. We first evaluated the correlations among the GLCM parameters to identify the most independent parameters. These GLCM parameters were further assessed on their ability to classify the PPE cells with the GMM based clustering algorithm. We obtained an optimized set of GLCM parameters that has sufficient information capacity for texture characterization of the p-DI data and investigated their relation to cell size and other parameters based on the clustering analysis.

5.1 Diffraction imaging flow cytometry

Figure 5.1 shows a schematic diagram of an improved version of experimental p-DIFC system used to obtain high-contrast data of p-DI pairs of the PPE cells. A fluidic control unit is used to drive the cell suspension as the core fluid into the flow chamber (FC) through a round glass nozzle of the inner diameter of 100 μm. At room temperature, the cells are carried by the core fluid through a concentric sheath fluid at a higher pressure upon entering the chamber at a speed of 4 mm/s. The cells move in single file through an incident beam produced by a continuous-wave solid-state green laser (MGL-III-532-100, CNI), and each emits scattered light at the same wavelength of 0.532 μm. An infinity-corrected 50x microscope objective (378-805-3, Mitutoyo) (OB) of 0.55 in numerical aperture used to collect the light scatter from flowing cells. The objective is aligned along the

direction perpendicular to the flowing direction along the y-axis and the incident beam along the z-axis. A polarizing beam splitter (PBS) and a half-wave plate (WP) are used to adjust the incident beam's power. The incident beam's direction of polarization is changed with another WP into one of three directions of vertical (ver), horizontal (hor), and 45° between different runs of cell measurements. The polarization direction of the incident beam determines the average orientation of the induced molecular dipoles inside the illuminated cell.

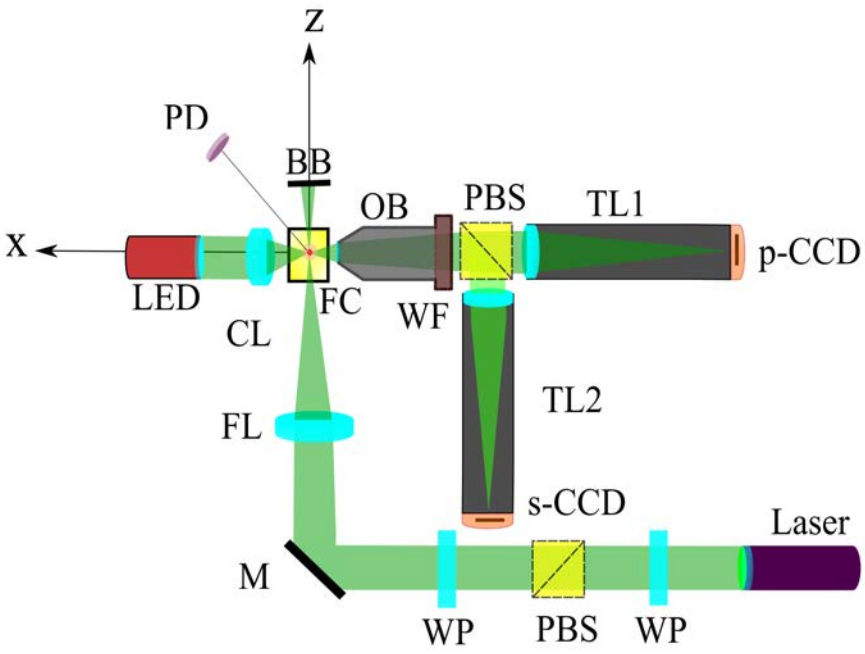

Figure 5.1 Top view diagram of an experimental p-DIFC system for acquisition of s- and p- polarized diffraction images. WP: half-wave plate; PBS: polarizing beam splitter; M: mirror; FL: focusing lens; FC: flow chamber; CL: condenser lens; PD: photodiode; OB: objective; WF: 532 nm wavelength filter; TL: tube lenses; CCD: camera. The x-axis and z-axis are labeled by black lines.

As a result, rotation of incident beam polarization allows different ways to "tune" the directions of the intracellular molecular dipoles and provide an opportunity to optimize the performance of classification. An interference wavelength filter (WF) of 0.532 μm is used in front of PBS to remove any nonelastic scattered and ambient light and the PBS divides the collected side scatters light into two components of horizontal and vertical polarizations, labeled as p-polarized and s-

polarized scatter, separated by 90°, and focused to the two CCD cameras (p-CCD and s-CCD). Pair of polarized diffraction images (p-DI) of 640×480 pixels and 12-bit pixel depth are acquired from each flowing cell with one of three incident beam polarizations. A beam splitter and photomultiplier (not shown) are used to detect a smaller portion of s-polarized scatter to trigger the two CCD cameras (LM075, Lumera). The exposure time will be set to range from 0.3 to 1.0 ms to reduce blurring of the cellular diffraction images. The throughput of the p-DIFC system will be maintained at about 1 to 4 cells/s and is mainly limited to the frame rate of the CCD cameras triggered externally. More technical details of the p-DIFC fluidic design, which explains the cell positioning through hydrodynamic focusing in, a square flow channel and the imaging of scattering light have been published elsewhere [42, 43, 94, 95].

A graphical user interface (GUI) software has been previously developed to control the p-DIFC system and two CCD cameras. This software designed to connect the cameras to the system and to set basic parameters such as exposure time, image signal gain for both cameras. When an external trigger signal is received from the photomultiplier, the system starts to acquire images, save them in the hard drive of the computer and display selected ones with their intensity parameters on the computer monitor. The image counter will add up one and compare with the set number of measure cells. When the counter number equals to the set number, the software terminates the acquisition process. With the aid of a multithreading mechanism, the image acquisition software can process multiple pairs of p-DIs per second. Also, the software can obtain image pixel intensity parameters and present the detail of this information in real time. This functionality help with adjustment of incident laser beam power to reduce the probabilities of acquiring underexposed or overexposed image data in subsequent acquisition because the dynamic ranges of the cameras are quite limited. For this purpose, we first acquire ten to twenty pairs of diffraction images for adjustment of the incident beam power. As each image pair is continuously acquired, the image pixel intensity parameters are calculated and displayed as maximum pixel intensity, the minimum pixel intensity, the average pixel intensity, and the total number of saturated pixels. The real-time feedbacks of image pixel intensities are not only helpful to adjust the beam power before starting

the data acquisition but also useful to monitor the image quality during data acquisition.

The design of p-DIFC system was improved in January 2018. The improvements included the modification of illumination unit and use of one time-delay-integration CCD (TDI-CCD) camera to increase the throughput rate and image quality [96, 97]. As shown in Figure 5.2 two cylindrical lenses (CyL1 and CyL2) of 500 mm and 60 mm in focal lengths are used instead of spherical lenses to focus the incident laser beam on the core fluid carrying the cells. The profile of linearly polarized incident beam propagating along the z-axis is close to Gaussian with an elliptical cross-section of major and minor axes of about 300 μm along the flow direction the y-axis and about 50 μm along the x-axis respectively. The power and polarizing direction of the incident beam are adjusted with a WP and the Glan-Thompson polarizing prism (GP). The light scattered by the cell passes through OB. A Wollaston prism (WSP), (LSP-3A14, Laser Institute, QFNU), employed to separate the scattered light into s- and p-polarized beams. These two beams are separated by an angle of 20°, and focused on the TDI-CCD camera by a tube lens of 75 mm in focal length.

Unlike the regular CCD sensor, the TDI camera requires synchronization between the linear transfer frequency of the imaging sensor and the speed of moving object for blur-free imaging [97]. The TDI camera acquire one 12-bit image of 2014×512 pixels that consists of two equal sections recording the s-polarized scattered light intensity on the left and p-scattered light on the right. To obtain a p-DI pair, preprocessing software crops the acquired image to two regions of 400×300 pixels based on maximum total pixel intensity. The preprocessing software is incorporated with GUI software develop to control the p-DIFC system and TDI camera system [97].

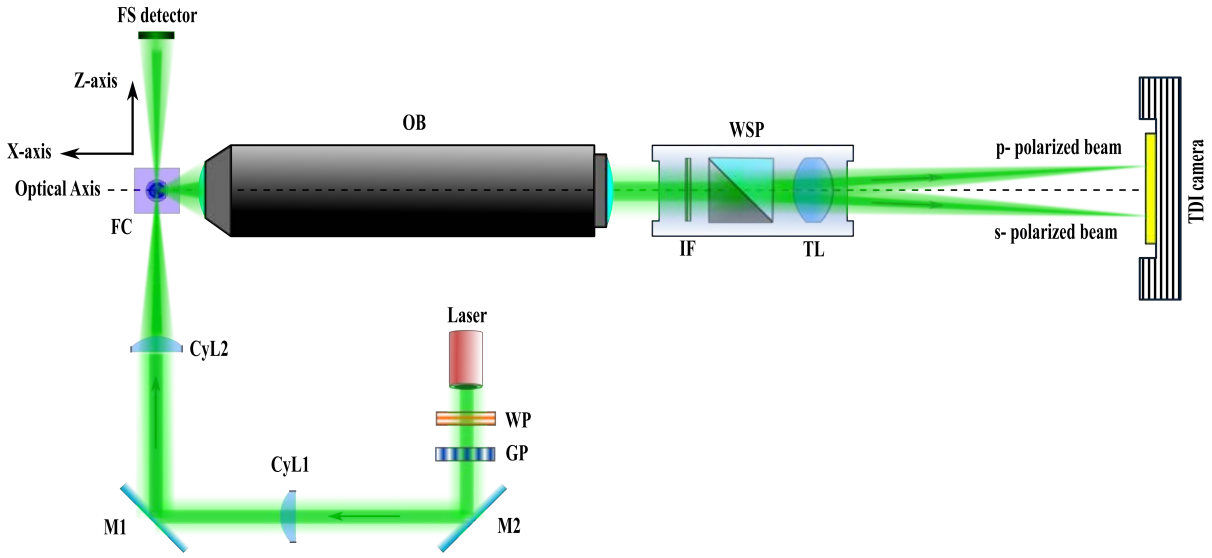

Figure 5.2 Top view diagram of an improved experimental p-DIFC system of s- and p- polarized diffraction images acquisition. Laser: laser beam source; WP: half-wave plate; GP: Glan-Thompson prism; M: mirror; CyL1 and CyL2: two cylindrical lenses; FC: flow chamber; FS: forward scatter; OB: objective; IF: interference filter; WSP: Wollaston prism; TL: tube lenses; TDI camera: camera. The x-axis and z-axis are labeled by black lines.

5.2 Preprocessing of p-DI data

Cell suspension sample contains various types of particles other than complete cells such as cellular debris and aggregated homogeneous particles formed inside the cell suspension medium. Furthermore, due to the variations of experimental conditions such as the core fluid position relative to the focus of the incident laser beam, the raw diffraction images acquired by our p-DIFC system can become underexposed or overexposed. Thus the p-DI pairs obtained by the p-DIFC system require preprocessing before analysis of the image textures for cell assay. The p-DI pairs filtered with an in-house developed image preprocessing software using MATLAB. The process started by obtaining the minimum, maximum, and the mean pixel value of an original 12-bit image pair data. Then, all the overexposed and underexposed p-DI pairs were removed based on the number of saturated pixels and mean pixel intensity respectively. The saturation value of 12-bit pixel is 4095. If one image has more than 8% of the total pixels are saturated, then the p-DI pair

considered overexposed. If the mean pixel intensities of both images are less than 2% of the pixel saturation value, then the p-DI pair was considered underexposed. The rest of p-DI pairs are linearly normalized to the 8-bit images then manually pre-screened [98]. The p-DI pairs with large speckles or highly symmetric strip patterns were removed since these have been shown to associate with cellular fragments or aggregates of non-cellular particles as shown in Figure 5.3 [23]. After the pre-screening, the remaining p-DI pairs were imported into another in-house software developed with MATLAB that uses the GLCM algorithm to extract image feature parameters that characterize the texture and the pixel intensity of the normalized image pair [44, 91].

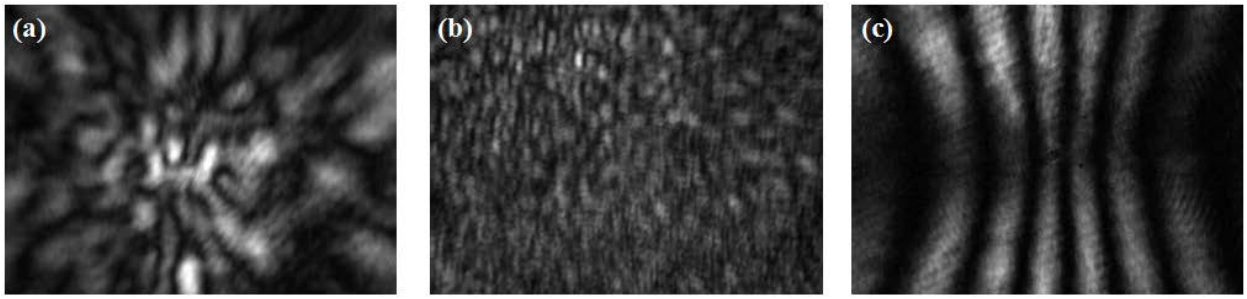

Figure 5.3 Examples of raw DIs of (a) a cell (b) cellular fragments, and (c) non-cellular particles.

5.3 The measured p-DI data and GLCM analysis

We have performed multiple diffraction imaging measurements on the PPE cells to investigate cell classification by the p-DIFC method. Table 5.1 presents the numbers of the diffraction image pairs for three incident laser beam polarization acquired from viable PPE cells extracted from the same twelve patient samples used for confocal imaging and denoted as P1, P2, ..., and P12. The p-DI pairs investigated for this study were acquired from P1, P2, and P3 cell samples to demonstrate the results of HC and GMM clustering technique used on the extracted feature parameters for the diffraction images texture of different types of cells in three samples. we also compared the results with results of simulated DI in previous chapter. Figure 5.4 shows examples selected 12-bit image pairs of the p-DI images acquired from single PPE cells of P1, P2, and P3 with different incident

beam polarizations. The size and the distribution of speckles in PPE images can be seen to be different among all images, and It is difficult to tell the difference among the diffraction image pairs by visual examination. To quantify the diffraction image pair features, we employed the GLCM based image processing to extract 34 texture parameters from each pair of images for each imaged cell and use them in different combinations to analyze the cells in a multidimensional feature space.

Table 5.1 p-DI measured data

				Pre-screening				
Patient ID	Camera	Polarization	p-DI pairs	Cell	Debris	Strips	Over Exp.	Under Exp.
P1	CCD	vertical	2911	1451	1180	126	108	46
		horizontal	600	342	73	185	0	0
		45°	3000	1166	403	311	28	1092
P2	CCD	vertical	4698	1446	1738	384	80	1050
		horizontal	6000	588	140	106	0	5166
		45°	3051	1006	427	185	1	1432
P3	CCD	vertical	4000	2603	269	366	1	761
		horizontal	4000	291	82	96	5	3526
		45°	6001	2816	892	1072	127	1094
P4	TDI	0	0	0	0	0	0	0
P5	TDI	vertical	4171	1649	1997	42	288	0
		horizontal	1000	440	489	2	69	0
		45°	1426	636	698	30	54	0
P6	TDI	vertical	5126	2396	2440	180	100	1953
		horizontal	3555	1622	1751	104	27	0
		45°	2487	1628	610	45	17	0
P7	TDI	vertical	7400	6107	1011	142	95	0
		horizontal	6482	4148	1831	418	41	0
		45°	10293	8678	1133	265	168	0
P8	TDI	0	0	0	0	0	0	0
P9	TDI	vertical	4006	2363	1044	567	32	0
		horizontal	4061	2014	1830	211	6	0
		45°	4945	2496	968	1449	32	0
P10	TDI	vertical	5315	4129	606	272	308	0
		horizontal	3761	3086	527	118	30	0
		45°	5556	4769	396	163	228	0
P11	TDI	vertical	5840	5149	528	99	64	0
		horizontal	5929	5022	861	35	11	0
		45°	6000	5222	669	66	43	0
P12	TDI	0	0	0	0	0	0	0

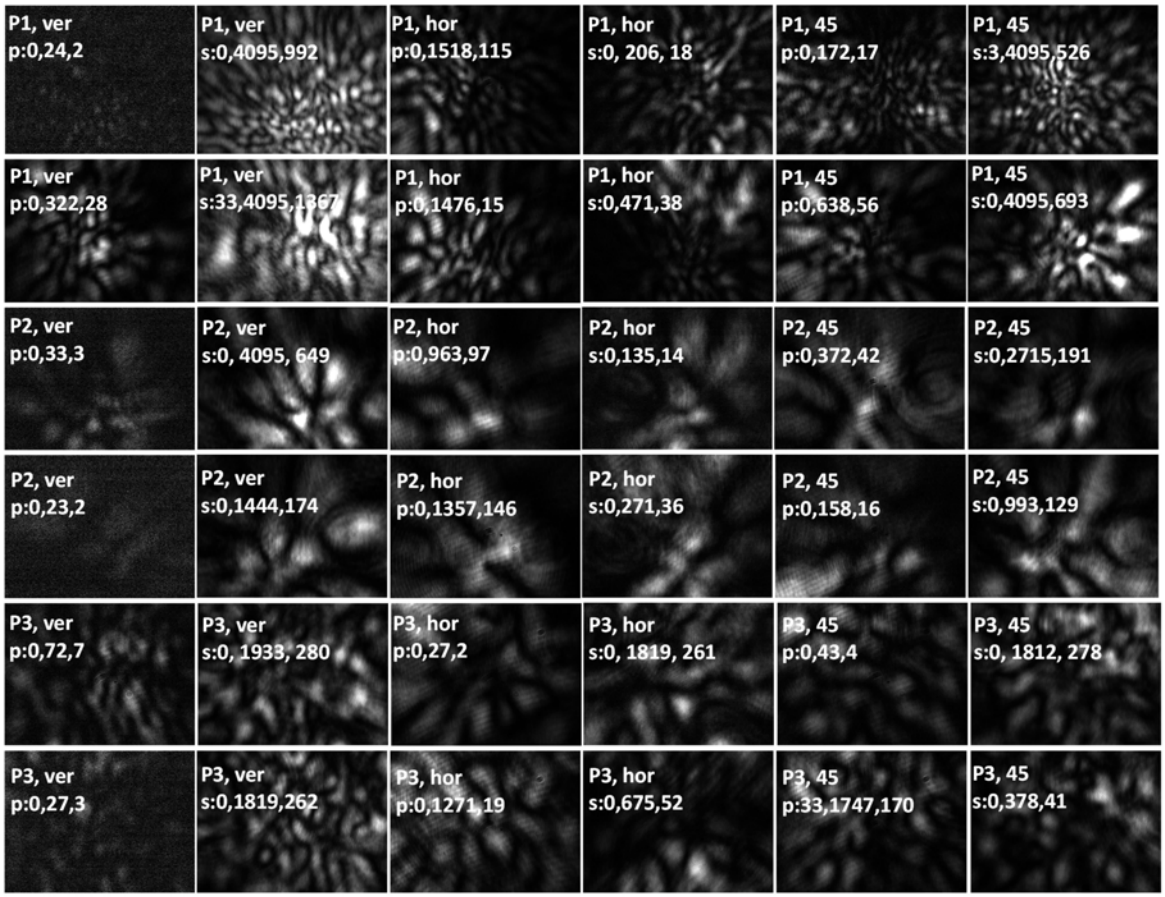

Figure 5.4 Examples of normalized p-DI images pairs of the PPE cells for the measurements in the vertical, horizontal, and 45° polarization of incident beam. Each image was labels with patient ID, the polarization direction of incident beam, the polarization direction of the scattered light, and minimum, maximum, and mean pixel intensities of the acquired 12-bit images.

Also, one can very clearly notice from these images that the PPE cells present non-even light intensity distributed between the image pairs. The cells show stronger s-polarized scattered light for a vertical scattered incident beam than the p-polarized light. While much more p-polarized scattered light observed for a horizontal scattered incident beam than the s-polarized light. This is because s-polarized incident beam induces molecular dipoles of the imaged cell oriented along the vertical axis and p-polarized incident beam induces molecular dipoles along the horizontal axis. For side scattering direction along the horizontal axis (θ = 90°), those dipoles induced by the s-polarized beam have a higher contribution to light signals than those induced by the

p-polarized beam for the transverse nature of light wavefields [97]. For that reason, we used the linear depolarization ratio δ_L measured at normal scattering angle to quantify the amount of energy transferred between the co-polarized and cross-polarized components of light. Table 5.2 provides the values of linear depolarization ratio and other parameters for the three sets of measured data (P1, P2, and P3). The mean values of δ_L show that these cells exhibit strong ability to transfer energy in the side scatter from co-polarized to cross-polarized component relative to the incident light. In comparison, single particles of spherical symmetry have $\delta_{L,ver} = \delta_{L,hor} = 0$.

Table 5.2 The values of linear depolarization ratio δ_L and other parameters

Paramters	**P1**			**P2**			**P3**		
	ver	hor	45°	ver	hor	45°	ver	hor	45°
incident power $P_\circ$ (mW)	168	168	168	100	100	100	31	167	167
number of p-DI pairs	1451	342	1166	1446	588	1006	2603	291	2816
pixel intensity $\overline{I}_s, \overline{I}_p$	781.4, 6.0	40.22, 178.8	337.8, 33.00	401.5, 7.17	24.38, 166.9	208.1, 28.42	307.4, 4.76	36.88, 178.1	369.8, 20.50
maximum δ (%)	10.80	54.89	117.12	24.95	54.7	56.55	36.53	48.60	42.33
minimum δ (%)	0.17	3.44	1.86	0.38	3.90	2.21	0.59	5.70	1.65
mean δ (%)	0.91	23.52	7.33	1.95	15.40	12.49	1.65	20.3	6.05

After the acquisition and preprocessing of the p-DI pair data, the images of the cells were further processed by the GLCM based MATLAB code to extract 30 image texture parameters for classification study. Three p-DIFC measurements of the PPE cells have been carried out to examine the clustering of the data. Tables 5.3, 5.4, and 5.5 present all minimum, maximum, mean, and STD values of 15 GLCM parameters calculated for p-DIs of stronger polarized scattered light with respect to incident polarized light (i.e. vertical, horizontal, and 45°) acquired for P1, P2, and P3.

Table 5.3 Distribution range of GLCM parameters for p-DIs of P1.

GLCM parameters	vertical, s				horizontal, p				45°, s			
	minimum	maximum	mean	STD	minimum	maximum	mean	STD	minimum	maximum	mean	STD
ASM	3.17×10^{-4}	1.03×10^{-2}	1.67×10^{-3}	1.38×10^{-3}	7.59×10^{-4}	1.63×10^{-2}	4.25×10^{-3}	2.65×10^{-3}	6.64×10^{-4}	1.47×10^{-2}	4.06×10^{-3}	2.26×10^{-3}
CON	4.11	2.31×10^{2}	29.8	19.4	4.48	67.00	15.50	7.75	3.94	99.50	16.00	10.60
COR	0.92	0.99	0.99	6.81×10^{-3}	0.97	0.99	0.99	4.02×10^{-3}	0.96	0.99	0.99	4.58×10^{-3}
VAR	1.69×10^{2}	5.74×10^{3}	1.77×10^{3}	1.19×10^{3}	2.71×10^{2}	5.83×10^{3}	8.66×10^{2}	5.28×10^{2}	1.42×10^{2}	5.81×10^{3}	1.06×10^{3}	8.75×10^{2}
IDM	9.64×10^{-2}	0.53	0.27	6.70×10^{-02}	0.20	0.57	0.36	7.47×10^{-2}	0.17	0.56	0.38	7.64×10^{-2}
SAV	29.8	239	109	41.1	27.5	142	51.8	14.5	25.0	168	58.8	21.9
SEN	4.14	5.98	5.37	0.36	4.12	5.59	4.81	0.26	4.01	5.80	4.88	0.31
SVA	6.71×10^{2}	2.29×10^{4}	7.04×10^{3}	4.75×10^{3}	1.08×10^{3}	2.32×10^{4}	3.45×10^{3}	2.11×10^{3}	5.65×10^{2}	2.32×10^{4}	4.22×10^{3}	3.49×10^{3}
ENT	5.38	8.76	7.48	0.60	5.23	7.75	6.55	0.50	5.19	8.39	6.57	0.54
DEN	1.49	3.36	2.34	0.26	1.46	2.62	2.03	0.23	1.46	2.94	2.01	0.25
DVA	2.36	133	14.3	10.0	2.80	43.4	8.16	4.01	2.18	54.5	8.74	5.89
DIS	1.32	10.8	3.74	1.10	1.24	4.78	2.60	0.70	1.25	6.62	2.56	0.80
CLS	6.23×10^{4}	4.13×10^{6}	9.85×10^{5}	7.99×10^{5}	1.28×10^{5}	4.72×10^{6}	5.22×10^{5}	4.59×10^{5}	5.69×10^{4}	4.48×10^{6}	7.25×10^{5}	7.50×10^{5}
CLP	1.14×10^{7}	1.67×10^{9}	3.46×10^{8}	3.33×10^{8}	2.71×10^{7}	1.93×10^{9}	1.47×10^{8}	1.83×10^{8}	1.47×10^{7}	1.81×10^{9}	2.35×10^{8}	3.04×10^{8}
MAP	1.41×10^{-3}	7.60×10^{-2}	1.17×10^{-2}	1.37×10^{-2}	2.52×10^{-3}	6.90×10^{-2}	1.75×10^{-2}	1.09×10^{-2}	1.99×10^{-3}	6.75×10^{-2}	1.66×10^{-2}	9.95×10^{-3}

Table 5.4 Distribution range of GLCM parameters for p-DIs of P2.

GLCM parameters	vertical, s				horizontal, p				45°, s			
	minimum	maximum	mean	STD	minimum	maximum	mean	STD	minimum	maximum	mean	STD
ASM	4.92×10^{-4}	1.72×10^{-2}	3.95×10^{-3}	2.35×10^{-3}	5.40×10^{-4}	1.95×10^{-2}	4.57×10^{-3}	3.50×10^{-3}	4.40×10^{-4}	2.90×10^{-2}	3.85×10^{-3}	2.68×10^{-3}
CON	3.63	1.71×10^{2}	16.5	12.4	3.54	59.6	15.6	8.58	2.67	63.3	15.1	8.38
COR	0.97	0.99	0.99	2.94×10^{-3}	0.97	0.99	0.99	2.87×10^{-3}	0.97	0.99	0.99	3.44×10^{-3}
VAR	1.93×10^{2}	6.14×10^{3}	1.50×10^{3}	1.15×10^{3}	3.46×10^{2}	2.95×10^{3}	1.10×10^{3}	4.00×10^{2}	2.74×10^{2}	5.14×10^{3}	1.00×10^{3}	5.50×10^{2}
IDM	0.16	0.58	0.39	7.62×10^{-02}	0.17	0.59	0.37	8.88×10^{-2}	0.16	0.62	0.36	8.60×10^{-2}
SAV	26.4	176	67.6	27.5	29.0	109	57.5	15.3	19.9	168	58.8	21.9
SEN	4.11	5.85	5.02	0.33	4.06	5.63	4.90	0.26	4.01	5.80	4.88	0.31
SVA	7.69×10^{2}	2.45×10^{4}	5.97×10^{3}	4.57×10^{3}	1.38×10^{3}	1.18×10^{4}	4.38×10^{3}	1.60×10^{3}	1.09×10^{3}	2.05×10^{4}	4.00×10^{3}	2.20×10^{3}
ENT	5.20	8.62	6.70	0.57	5.23	8.10	6.63	0.60	4.55	8.25	6.62	0.56
DEN	1.41	3.16	2.02	0.26	1.42	2.79	2.03	0.27	1.26	2.82	2.03	0.26
DVA	2.13	94.1	9.02	6.99	2.22	26.0	8.09	3.78	1.72	27.1	7.50	3.63
DIS	1.18	8.52	2.57	0.84	1.15	5.79	2.61	0.83	0.97	6.00	2.63	0.82
CLS	9.70×10^{4}	5.35×10^{6}	10.8×10^{5}	9.81×10^{5}	1.76×10^{5}	2.80×10^{6}	6.58×10^{5}	3.31×10^{5}	16.5×10^{4}	4.56×10^{6}	6.41×10^{5}	4.84×10^{5}
CLP	2.52×10^{7}	2.17×10^{9}	3.73×10^{8}	4.09×10^{8}	3.47×10^{7}	1.02×10^{9}	1.83×10^{8}	1.20×10^{8}	3.63×10^{7}	1.86×10^{9}	1.89×10^{8}	1.95×10^{8}
MAP	1.80×10^{-3}	7.59×10^{-2}	1.90×10^{-2}	1.28×10^{-2}	1.69×10^{-3}	8.74×10^{-2}	1.96×10^{-2}	1.47×10^{-2}	1.72×10^{-3}	10.3×10^{-2}	1.52×10^{-2}	11.0×10^{-3}

Table 5.5 Distribution range of GLCM parameters for p-DIs of P3.

GLCM parameters.	vertical, s				horizontal, p				45°, s			
	minimum	maximum	mean	STD	minimum	maximum	mean	STD	minimum	maximum	mean	STD
ASM	5.99×10^{-4}	1.50×10^{-2}	2.67×10^{-3}	1.34×10^{-3}	7.01×10^{-4}	10.8×10^{-2}	2.57×10^{-3}	1.57×10^{-3}	4.85×10^{-4}	1.15×10^{-2}	2.53×10^{-3}	1.33×10^{-3}
CON	4.00	170	18.7	9.02	4.15	100	19.7	11.0	4.97	176	22.4	13.5
COR	0.94	0.96	0.99	3.06×10^{-3}	0.98	0.99	0.98	3.49×10^{-3}	0.94	0.99	0.99	3.42×10^{-3}
VAR	2.73×10^{2}	5.23×10^{3}	9.79×10^{3}	4.21×10^{3}	2.77×10^{2}	3.52×10^{3}	9.48×10^{2}	4.23×10^{2}	2.72×10^{2}	5.67×10^{3}	1.21×10^{3}	8.39×10^{2}
IDM	0.17	0.57	0.34	6.17×10^{-2}	0.19	0.54	0.32	6.94×10^{-2}	0.16	0.54	0.33	6.36×10^{-2}
SAV	25.3	188	63.1	14.4	29.2	147	61.0	15.6	28.9	207	69.6	23.1
SEN	4.09	5.82	5.03	0.22	4.24	5.80	5.00	0.25	4.22	5.88	5.09	0.28
SVA	10.9×10^{2}	2.07×10^{4}	3.90×10^{3}	1.68×10^{3}	1.10×10^{3}	1.40×10^{4}	3.77×10^{3}	1.68×10^{3}	10.8×10^{2}	2.26×10^{4}	4.82×10^{3}	3.35×10^{3}
ENT	5.10	8.57	6.86	0.42	5.38	8.45	6.91	0.48	5.35	8.73	6.97	0.49
DEN	1.40	3.18	2.12	0.20	1.50	2.96	2.17	0.23	1.52	3.26	2.18	0.23
DVA	2.60	91.3	9.78	4.52	2.38	51.1	9.81	5.30	3.02	89.5	11.9	7.18
DIS	1.17	8.64	2.87	0.69	1.32	6.86	3.03	0.82	1.36	9.18	3.08	0.85
CLS	15.4×10^{4}	3.55×10^{6}	5.43×10^{5}	3.26×10^{5}	1.14×10^{5}	2.63×10^{6}	5.16×10^{5}	3.16×10^{5}	12.8×10^{4}	3.97×10^{6}	7.25×10^{5}	6.49×10^{5}
CLP	3.38×10^{7}	1.44×10^{9}	1.53×10^{8}	1.25×10^{8}	2.21×10^{7}	9.66×10^{9}	1.43×10^{8}	1.20×10^{8}	2.64×10^{7}	1.58×10^{9}	2.26×10^{8}	2.64×10^{8}
MAP	2.23×10^{-3}	6.41×10^{-2}	1.07×10^{-2}	5.35×10^{-3}	3.10×10^{-3}	3.72×10^{-2}	1.07×10^{-2}	6.08×10^{-3}	1.91×10^{-3}	7.23×10^{-2}	1.14×10^{-2}	7.37×10^{-3}

In this research, bivariate Pearson and Spearman correlation coefficients employed to quantify the correlation between the GLCM parameters extracted from the p-DI data in search of a small set that has sufficient information capacity for texture characterization. The values of these coefficients measure the strength of associations, and non-causal relationships between a selected pair of GLCM parameters [99]. The correlation coefficients are unitless, this makes an evaluation of relationships easier through a correlation approach, and allows us to decide whether one relationship is stronger than another without worrying about units inducing that assessment. Correlation coefficients of null give an indicator of no associate relationship, while the values of $+1$ or -1 give an indicator of the linear relationship between two parameters, which is what Pearson's correlation determines, or the monotonic relationship between your two parameters, which is what Spearman's correlation determines [100]. Whenever the points in the plot become more spread around that straight line,

the correlation reduces, and any likeness of a linear relationship vanishes [101]. The Pearson correlation coefficient is given by

$$r_p = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{[\sum_{i=1}^{n}(x_i - \overline{x})]^2[\sum_{i=1}^{n}(x_i - \overline{x})]^2}}, \tag{5.1}$$

where x_i and y_i are the two GLCM parameter values of image i, $\overline{x}$ and $\overline{y}$ are the mean values of x and y of n images [99]. The numerator of equation 5.1 is the sample covariance of x and y, and the denominator is the square root of the sample variance of x; and the sample variance of y. If the covariance is positive, it expresses a positive linear relationship, and if it is negative, then it expresses a negative linear relationship [102]. Correlation coefficients of null give an indicator of no linear relationship, while the values of 1 give an indicator of the points all align on a straight line, and this line must have a non-zero slope [100].

Unlike Pearson correlation, the Spearman correlation determines the strength and direction of the monotonic relationship between the two parameters and dose not affected by the shape of the distributions and the outliers because it depends on the distance between the ranked parameters. Spearman's correlation coefficient r_s is given by

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}, \tag{5.2}$$

where n is the sample size and d_i is the difference between first parameter and second parameter ranks [99]. To evaluate d_i^2 we first need to sort the data of the two parameters from the minimum to the maximum and assign ranking values for each one of them, then square difference between two ranked values. Tables 5.6 and 5.7 present r_p and r_s values obtained between paired GLCM parameters extracted form 22828 DIs from three sets of p-DI pairs measured form three PPE cell samples of PPE cells. The r_p values range from less than 0.1 for very weak correlations to larger than 0.9 for very strong correlations for a given pair of GLCM parameters. In addition to r_p and r_s, we also obtained the multiple correlation coefficient R^2 using the same combined DI data set to assess each GLCM parameter's predictability of the other 14 parameters in their values. The

list of GLCM parameters is given in Table 5.8. We chose a set of four parameters of the top four parameters of low R^2 values to perform classification of PPE cells by the HC and GMM method.

Table 5.6 Pearson's correlation coefficient r_p

GLCM Parameter	ASM	CON	COR	VAR	IDM	SAV	SEN	SVA	ENT	DEN	DVA	DIS	CLS	CLP	MAP
ASM															
CON	0.93														
COR	-0.93	-0.96													
VAR	-0.12	-0.08	0.16												
IDM	-0.10	-0.28	0.34	0.09											
SAV	-0.28	-0.21	0.28	0.86	-0.02										
SEN	-0.88	-0.87	0.93	0.41	0.25	0.56									
SVA	-0.30	-0.28	0.35	0.98	0.14	0.87	0.56								
ENT	-0.88	-0.84	0.89	0.36	0.03	0.55	0.97	0.52							
DEN	0.10	0.27	-0.33	-0.02	-0.97	0.02	-0.24	-0.08	-0.04						
DVA	0.94	0.99	-0.95	-0.07	-0.25	-0.20	-0.86	-0.27	-0.84	0.24					
DIS	0.85	0.95	-0.97	-0.13	-0.51	-0.23	-0.89	-0.31	-0.81	0.50	0.94				
CLS	-0.33	-0.34	0.40	0.92	0.25	0.73	0.55	0.95	0.48	-0.17	-0.33	-0.38			
CLP	-0.27	-0.29	0.34	0.92	0.23	0.72	0.49	0.95	0.43	-0.15	-0.28	-0.33	0.99		
MAP	0.97	0.94	-0.97	-0.12	-0.26	-0.29	-0.93	-0.30	-0.91	0.27	0.94	0.92	-0.33	-0.27	

Table 5.7 Spearman correlation coefficient r_s

GLCM Parameter	ASM	CON	COR	VAR	IDM	SAV	SEN	SVA	ENT	DEN	DVA	DIS	CLS	CLP	MAP
ASM															
CON	0.46														
COR	-0.55	-0.91													
VAR	-0.33	0.04	0.29												
IDM	-0.05	-0.77	0.72	0.09											
SAV	-0.64	-0.09	0.36	0.82	-0.01										
SEN	-0.90	-0.56	0.74	0.54	0.27	0.77									
SVA	-0.65	-0.34	0.62	0.86	0.19	0.86	0.83								
ENT	-0.96	-0.38	0.54	0.42	-0.02	0.68	0.92	0.72							
DEN	0.09	0.80	-0.71	-0.04	-0.98	0.02	-0.26	-0.16	0.01						
DVA	0.48	0.99	-0.89	0.06	-0.75	-0.08	-0.56	-0.32	-0.39	0.79					
DIS	0.45	0.99	-0.92	0.01	-0.79	-0.10	-0.57	-0.36	-0.38	0.81	0.99				
CLS	-0.65	-0.57	0.80	0.67	0.35	0.64	0.83	0.92	0.71	-0.32	-0.55	-0.59			
CLP	-0.64	-0.61	0.82	0.63	0.39	0.61	0.82	0.89	0.70	-0.36	-0.58	-0.63	0.99		
MAP	0.90	0.67	-0.67	-0.24	-0.34	-0.54	-0.84	-0.58	-0.80	0.41	0.69	0.66	-0.63	-0.64	

Table 5.8 R^2 correlation coefficient.

Parameters	ASM	CON	COR	VAR	IDM	SAV	SEN	SVA	ENT	DEN	DVA	DIS	CLS	CLP	MAP
R^2 (%)	99.63	100.00	**98.94**	100.00	**98.55**	**96.47**	99.60	100.00	99.29	**98.58**	99.98	99.80	99.57	99.51	99.55

5.4 Results of clustering analysis

The selected set of the four GLCM parameters are imported to the Matlab for clustering analysis. Each parameter is normalized between 0 and 1 by its own values. Based on the results we got from 3D morphological analysis in previous chapter, we used k=3 as the estimated number of clusters within the data extracted from 11414 cell. The HC analysis gives us a fixed initial starting point to classify the PPE cells with GMM method by separating the data points to three clusters depending on the distance among the points within the space of the selected parameters. The mean and the covariance matrix are calculated to obtain the Gaussian PDF for each cluster. The clustering is optimized by iteratively over all cells assigned to different clusters. The results of the clustering are denoted by C1, C2, and C3 with total number of cells 1825, 8672, and 917 respectively. Figure 5.5 presents the scattered plots of three selected GLCM parameters. It can be seen from these plots that the DIs of C1 cells have its parameter values spread in a cluster toward the high ends of the ranges of the three parameters in contrast to C3 which have its parameter values spread in a cluster toward the lower ends of the ranges of the three parameters. while the C2 DIs cluster near the low end of one parameters and the near high end of the other with less spread compare to the other clusters. Furthermore, both of the calculated and the measured DI data presented in this study exhibit highly symmetirical scatter plot distribution within the same selected parameters space as shown in Figures 4.17 and 5.5.

Selected example of clustered measured PPE DIs pair are presented in Figure 5.7. We randomly select three pairs from each cluster and from each patient to illustrate their diffraction patterns. By examining these DIs, one can observe that the diffraction of pattern of C1 exhibit relative high degrees of non-similarity to those of C2, while the patterns of both of them show significant differences from those of C3. These differences can be found among all cells extracted from the three patients and also within the same patient sample. The size of the speckles in C1 images can be seen to be slightly larger than those in C2 and C3 images and the total amounts of speckles in C1 and C3 images are less than those in C2 images. With the naked eyes, one is unable to quantify

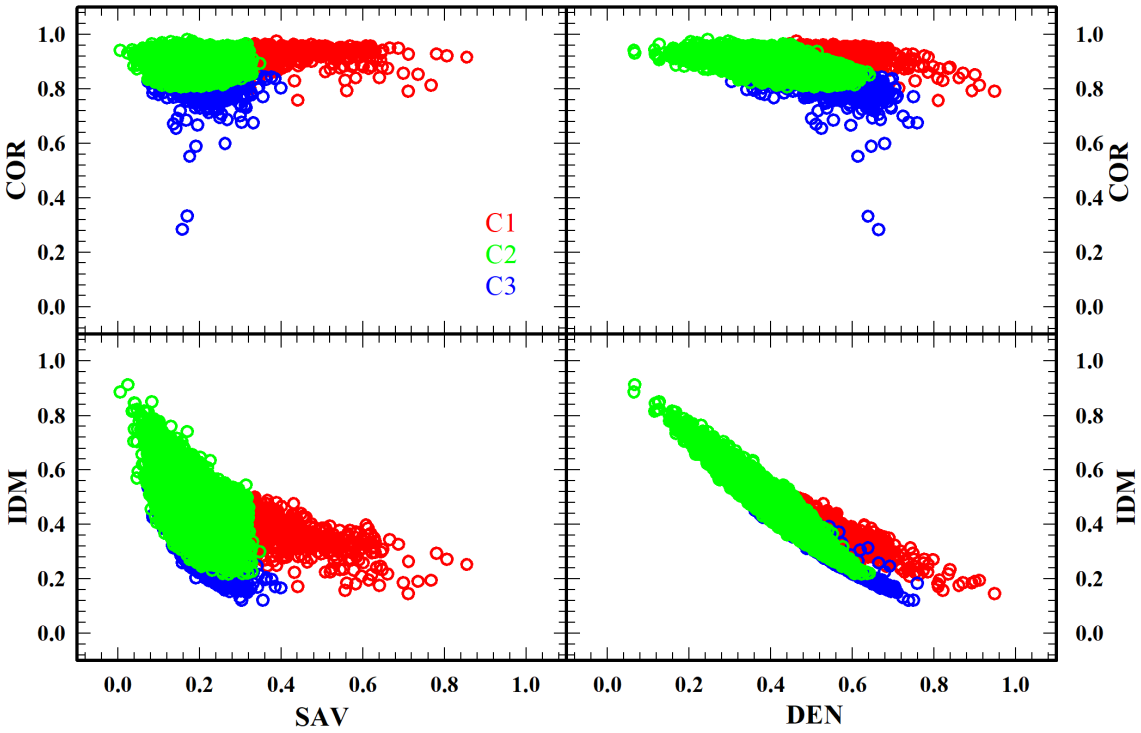

Figure 5.5 Scattering plots of three GLCM parameters extracted from the diffraction images of 11414 PPE cells and normalized between 0 and 1. COR = correlation; IDM = inverse difference moment. SAV =sum of variance.

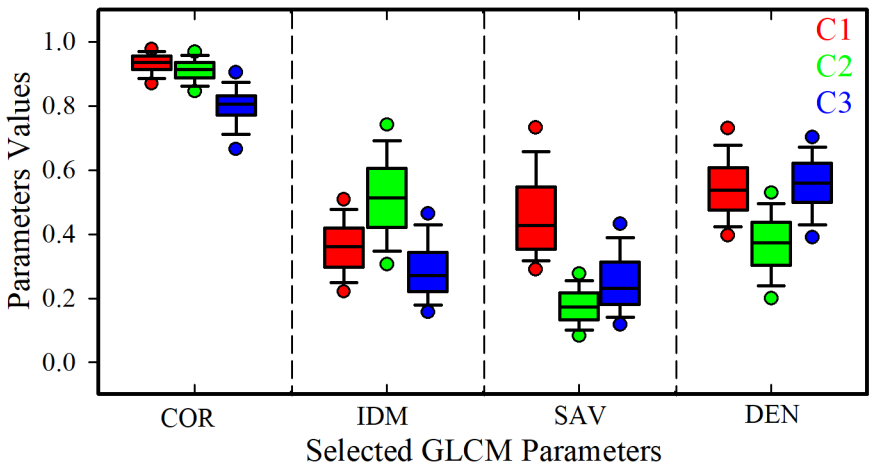

Figure 5.6 Box plot of the selective GLCM parameters.

the difference among these DIs between C1, C2, and C3 cells as well as among P1, P2, and P3 samples. From these results, it is obvious that the method of HC and GMM clustering based on selected GLCM parameters is efficient method for cell classification to investigate the variations in diffraction patterns and the correlations between these patterns and morphological parameters of the cells.

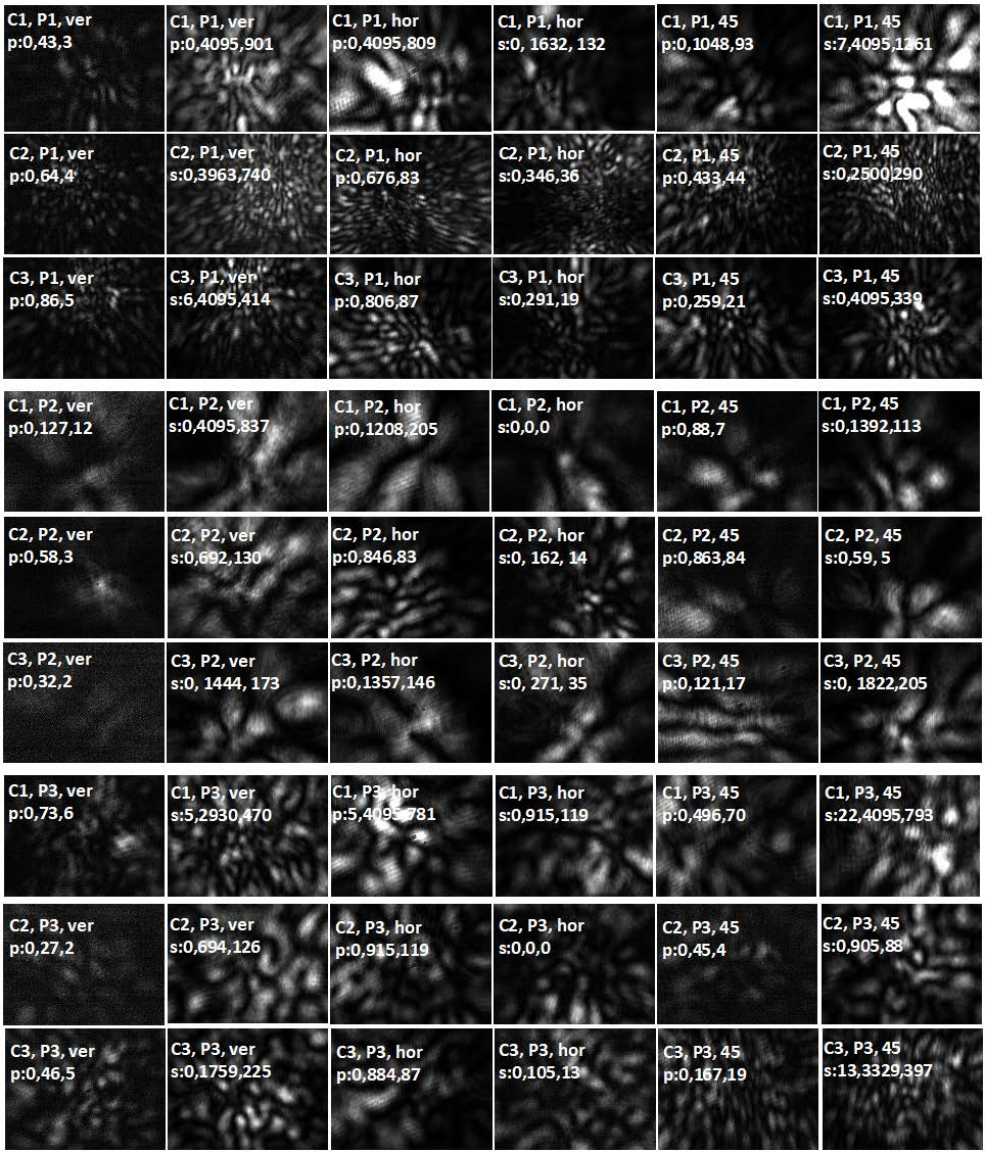

Figure 5.7 Examples of normalized DIs measured for PPE cells extracted from three different patients and clustered into three groups. Each image is marked on the top by cluster and patient ID, incident and scatter polarization, and minimum, maximum, and average pixel intensity.

Chapter 6 Conclusion

Through this dissertation research, we have investigated the profiling of primary PPE cells of patients suspected with cancers by the p-DIFC method to lay a foundation for future study of label-free detection of cancer cells in the PPE samples. To achieve this aim, the key issue is to study and establish correlations between the diffraction patterns embedded in p-DI data acquired from PPE cells and their morphological features revealed by the confocal and cytology fluorescence microscopy. Quantitative results of 3D and 2D morphology have been obtained in terms of volume and surface parameters of cell and intracellular organelles that play important roles in light scattering. We have also developed realistic optical cell models for simulation and analysis of light scattering patterns of the calculated p-DI data with various techniques. It has been demonstrated with the simulation of diffraction imaging process that strong correlations exist between the diffraction patterns embedded in the p-DI data and the 3D morphological characteristics. In addition, a suite of tools of imaging processing and statistical analysis have been developed based on existing algorithms through this study.

The investigations of 3D cell morphology have been carried out with confocal microscopy of 499 single PPE cells extracted 12 patients suspected with cancer. We acquired confocal image stack data and perform 3D reconstruction and parameter calculation after double staining of nucleus and mitochondria with fluorescent dyes Syto-61 and Mitotracker Orange. We used an in-house developed Matlab-based software for image segmentation, interpolation, and 3D reconstruction. A total of 27 morphology parameters of volume and surface related to the cell, nucleus, and mitochondria were calculated for quantitative analysis and comparison among different types of cells. Table 3.2 lists the means and standard deviations of all calculated parameters to provide an

overview of the PPE cells morphology. Six scatter plots of the imaged PPE cells with different pairs of morphology parameters are presented in Figure 3.4 as examples to compare their distributions. Despite their dispersed distributions, the symbols representing the cells show significant overlap in all of the scatter plots. But, these scatter plots are difficult to use for finding clear separation lines in the parameter space to discriminate different types of cells of the effusion samples and achieve the goal of cells classification.

For that reason it was imperative to adopt a machine learning algorithm which can be used to objectively and effectively identify the cell distribution in the morphology parameter space. We employed an unsupervised algorithms of the hierarchical clustering (HC) and Gaussian Mixture Model (GMM) for automated cell classification in the 3D parameter space of cell morphology. These algorithms, however, requires a prior knowledge on number of clusters available within the data. For determination of an optimized cluster number k for a given cell data set, we utilized the Akaike information criterion (AIC) and the Bayes information criterion (BIC) as statistical estimators to evaluate the optimal value of k in our classification study. Both of these criteria represent a balance between maximizing the degree of fit of the data point to a cluster model and minimizing number of clusters. The results showed that 3 is the best choice of k and the corresponding clusters were named as C1, C2 and C3. Tables 3.4 and 3.5 present the means and standard deviations of the 27 morphology parameters together with the p-values to test the statistical significance of the differences among the three cell clusters which are mainly characterized by their differences in cell, nuclear, and mitochondrial volumes with C1 indicating the cells of smallest volumes and C3 the cells of largest volumes while C2 the medium values. To examine the differences in morphology closely, we shown in Figure 3.8 of the 3D parameters selected from the Tables 3.4 and 3.5 with p-values < 0.05 for imaged cells to visualize and compare their distributions. These differences are quit obvious in statistical analysis of the mean value and the standard deviation. Even though, the scatter plots of those parameters with p-value of less than 0.05 evince that morphological parameters alone can hardly be used for effective classification of the different cell types.

We have also carried out 2D morphology quantitative analysis of PPE cells based on cytopathological image slides and labels provided by collaborating cytologists for comparison to the previous results 3D analysis. The cells in these slides were examined and evaluated by the cytologists for two major group labels of normal and cancer cells. Quantitative 2D morphological measurements of 560 PPE cells extracted from 6 patients were performed to obtain cell and nuclear areas. The results correlate the nature of cells to the morphological parameters of the cell and nuclear area value. The same clustering technique of HC and GMM with k=3 is used to classify the cells in 2D parameter space. The results of the clustering showed that among the three clusters there are about 98% of the cells in C1 are normal cells and 100% of the cells clustered in C3 are cancer cells. In comparison, cells in C2 have mixed portions of 19% as normal to 81% as cancer. Further analysis through scatter plot (Figure 3.13) for these results showed that most of normal cells are found in the range of small cell area and while the cancer cells are spread in bigger range of relatively large cell area while certain amount of normal cells and cancer cells are overlapped in the middle range.

Simulations of diffraction imaging were performed to investigate correlations between cell morphology and diffraction patterns. The p-DI calculations were based on virtual PPE cell structures obtained from the 3D morphological structures in terms of optical cell models (OCMs) by converting fluorescent light intensity values into refractive indices. We employed a previously in-house developed MATLAB code to build OCMs from confocal imaging based cell structures with three organelles of cytoplasm, nucleus and mitochondria. Equations 4.1 to 4.10 were used to obtain RI values for each voxel in a cellular organelle from the fluorescent intensity values of the fluorescent dyes tagged to the biomolecules inside nucleus or mitochondria. Four types of realistic OCMs have been developed with 9 PPE cell structures (3 small, 3 medium, and 3 large), selected based on the 3D morphology classification results, and adjustable refractive index (RI) parameters of nucleus and mitochondria to study the effects of cell morphology on RI distribution and diffraction patterns in calculated p-DI pairs. In addition We have also investigated OCM improvement by adding artificially an organelle of lysosome that is also important to light scattering. This has been achieved by adding a distribution of small spheres of volume rang from 0.11 to 305μm^3 and

average refractive index of 1.45 and standard deviation of 0.02 inside the cell to study the effect of the lysosomes. The texture features of the calculated p-DIs were extracted by the computation method GLCM to obtain a total of 30 image texture parameters for each cell to quantify diffraction pattern by the cell. The simulation results of 171 cell structure selected as 72 small, 60 medium, and 39 large were used to study the correlation between cell morphological features and diffraction patterns revealed by GLCM parameters. The classification of calculated p-DI data by clustering algorithms of HC and GMM by 5 selected GLCM parameters was used to understand the capability of cell profiling by p-DI to classify the measured p-DI data. The clustering process of calculated p-DI achieved clustering results with average accuracy of 57.8% and standard deviation of 10.0% in term of clustering the three of types of cells (small, medium, and large) into three clusters (C1, C2, and C3) as shown in Table 4.3. These findings provide a basis to understand the the ability of p-DIs for morphology based classification.

The p-DI measurements have been performed with the method of p-DIFC. The acquired p-DI data were first preprocessed to remove those by cellular debris and aggregated noncellular particles. A total of 11,414 p-DI pairs were obtained from PPE samples of 3 patients were imported to GLCM algorithm to extract 30 parameters per image. We used Pearson's, Spearman's and R^2 correlation coefficients to evaluated the correlations among 15 GLCM parameters to identify the set of most independent parameters. This parameter set was further assessed on their ability to classify the PPE cells with the HC and GMM based clustering algorithm. We obtained an optimized set of GLCM parameters that has sufficient information capacity for texture characterization of the p-DI data and investigated their relation to cell size and other parameters based on the clustering analysis. The light scattering patterns obtained through numerical simulation from the cell models and experimentally showed similar characteristics in term of GLCM parameters, indicating strong similarities between the optical models and real cells as shown in Figures 4.17 and 5.5. These results lead to a conclusion that the p-DIFC method has the potential to be developed into a rapid and label-free method for cell essay and morphology based classification to discriminate PPE cells of types based on size and morphology.

Bibliography

[1] Ie Ming Shih, Ritu Salani, Michael Fiegl, Tian Li Wang, Antoninus Soosaipillai, Christian Marth, Elisabeth Müller-Holzner, Gunther Gastl, Zhen Zhang, and Eleftherios P. Diamandis. Ovarian cancer specific kallikrein profile in effusions. *Gynecologic Oncology*, 105(2):501–507, 2007.

[2] Jack A. Kastelik. Management of malignant pleural effusion. *Lung*, 191(2):165–175, 2013.

[3] Justin M. Thomas and Ali I. Musani. Malignant Pleural Effusions A Review. *Clinics in Chest Medicine*, 34(3):459–471, 2013.

[4] V. B. Antony, R. Loddenkemper, P. Astoul, C. Boutin, P. Goldstraw, J. Hott, F. Rodriguez Panadero, and S. A. Sahn. Management of malignant pleural effusions. *European Respiratory Journal*, 18:402–419, 2001.

[5] Helma Motherby, Bahram Nadjari, Patricia Friegel, Johannes Kohaus, Uwe Ramp, and Alfred Böcking. Diagnostic accuracy of effusion cytology. *Diagnostic Cytopathology*, 20(6):350–357, 1999.

[6] Yong-Moon Lee, Ji-Yong Hwang, Seung-Myoung Son, Song-Yi Choi, Ho-Chang Lee, Eun-Joong Kim, Hye-Suk Han, Jin-young An, Joung-Ho Han, and Ok-Jun Lee. Comparison of diagnostic accuracy between CellprepPlus® and ThinPrep® liquid-based preparations in effusion cytology. *Diagnostic Cytopathology*, 42(5):384–390, 5 2014.

[7] Craig F. Bohren and Donald R. Huffman. *Absorption and scattering of light by small particles*. John Wiley & Sons, 1983.

[8] Andrew T. Young. Rayleigh scattering. *Applied Optics*, 20(4):533, 2 1981.

[9] H. C. van de Hulst. *Light scattering by small particles*. Courier Corporation, 1981.

[10] R. Scott Brock, Xin-Hua Hu, Douglas A. Weidner, Judith R. Mourant, and Jun Q. Lu. Effect of detailed cell structure on light scattering distribution: FDTD study of a B-cell with 3D structure constructed from confocal images. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 102(1):25–36, 11 2006.

[11] Konstantin V. Gilev, Maxim A. Yurkin, Ekaterina S. Chernyshova, Dmitry I. Strokotov, Andrei V. Chernyshev, and Valeri P. Maltsev. Mature red blood cells: from optical model to inverse light-scattering problem. *Biomedical Optics Express*, 7(4):1305, 4 2016.

[12] Roger G. Newton. *Scattering Theory of Waves and Particles*. Springer Science & Business Media, 1982.

[13] Akhlesh Lakhtakia and G.W. Mulholland. On two numerical techniques for light scattering by dielectric agglomerated structures. *Journal of Research of the National Institute of Standards and Technology*, 98(6):699, 11 1993.

[14] W. S. Bickel, J. F. Davidson, D. R. Huffman, and R. Kilkson. Application of polarization effects in light scattering: a new biophysical tool. *Proceedings of the National Academy of Sciences*, 73(2):486–490, 2 1976.

[15] Huafeng Ding, Jun Q. Lu, R. Scott Brock, Thomas J. McConnell, Jenifer F. Ojeda, Kenneth M. Jacobs, and Xin-Hua Hu. Angle-resolved Mueller matrix study of light scattering by B-cells at three wavelengths of 442, 633, and 850 nm. *Journal of Biomedical Optics*, 12(3):034032, 2007.

[16] Marina Moran. *Correlating the Morphological and Light Scattering Properties of Biological Cells*. PhD thesis, East Carolina University, 2013.

[17] Renliang Xu. Light scattering: A review of particle characterization applications. *Particuology*, 18:11–21, 2 2015.

[18] M. A. Yurkin, V. P. Maltsev, and A. G. Hoekstra. The discrete dipole approximation for simulation of light scattering by particles much larger than the wavelength. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 2007.

[19] Edward M. Purcell and Carlton R. Pennypacker. Scattering and Absorption of Light by Nonspherical Dielectric Grains. *The Astrophysical Journal*, 186:705, 12 1973.

[20] Bruce T. Draine and Piotr J. Flatau. Discrete-Dipole Approximation For Scattering Calculations. *Journal of the Optical Society of America A*, 11(4):1491, 4 1994.

[21] Maxim A. Yurkin and Alfons G. Hoekstra. User Manual for the Discrete Dipole Approximation Code ADDA 1.3b4, 2014.

[22] Maxim A. Yurkin and Alfons G. Hoekstra. The discrete-dipole-approximation code ADDA: Capabilities and known limitations. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 112(13):2234–2247, 9 2011.

[23] Wenhuan Jiang. *Study of Morphology Based Cell Assay by Diffraction Imaging Flow Cytometry*. PhD thesis, East Carolina University, 2015.

[24] P. F. Mullaney and P. N. Dean. The Small-Angle Light Scattering of Biological Cells. *Biophysical Journal*, 10(8):764–772, 8 1970.

[25] A. Brunsting and P. F. Mullaney. Light Scattering from Coated Spheres: Model for Biological Cells. *Applied Optics*, 11(3):675, 3 1972.

[26] Andrew Dunn and Rebecca Richards-Kortum. Three-dimensional computation of light scattering from cells. *IEEE Journal of Selected Topics in Quantum Electronics*, 2(4):898–905, 1996.

[27] J. R. Mourant, T. M. Johnson, S. Carpenter, A. Guerra, T. Aida, and J. P. Freyer. Polarized angular dependent spectroscopy of epithelial cells and epithelial cell nuclei to determine the size scale of scattering structures. *Journal of Biomedical Optics*, 7(3):378, 2002.

[28] Jun Zhang, Yuanming Feng, Wehnhuan Jiang, Jun Q. Lu, Yu Sa, Junhua Ding, and Xin-Hua Hu. Realistic optical cell modeling and diffraction imaging simulation for study of optical and morphological parameters of nucleus. *Optics Express*, 24(1):366–377, 2016.

[29] Marina S. Moran, Xin-hua Hu, and Jun Q. Lu. Detecting cellular morphological changes through light scattering patterns : comparison of methods. *JOURNAL OF ADVANCED OPTICS AND PHOTONICS*, 1(1):23–34, 2018.

[30] Maxim A. Yurkin, Konstantin A. Semyanov, Peter A. Tarasov, Andrei V. Chernyshev, Alfons G. Hoekstra, and Valeri P. Maltsev. Experimental and theoretical study of light scattering by individual mature red blood cells by use of scanning flow cytometry and a discrete dipole approximation. *Applied Optics*, 2005.

[31] Shuting Wang, Jing Liu, Jun Q. Lu, Wenjin Wang, Safaa A. Al-Qaysi, Yaohui Xu, Wenhuan Jiang, and Xin-Hua Hu. Development and evaluation of realistic optical cell models for rapid and label-free cell assay by diffraction imaging. *Journal of Biophotonics*, page e201800287, 11 2018.

[32] J. R. Mourant, M. Canpolat, C. Brocker, O. Esponda-Ramos, T. M. Johnson, A. Matanock, K. Stetter, and J. P. Freyer. Light scattering from cells: the contribution of the nucleus and the effects of proliferative status. *Journal of Biomedical Optics*, 2000.

[33] Rebekah Drezek, Martial Guillaud, Thomas Collier, Iouri Boiko, Anais Malpica, Calum Macaulay, Michele Follen, and Rebecca R. Richards-Kortum. Light scattering from cervical cells throughout neoplastic progression: influence of nuclear morphology, DNA content, and chromatin texture. *Journal of Biomedical Optics*, 2003.

[34] Jun Q. Lu, Ping Yang, and Xin-Hua Hu. Simulations of light scattering from a biconcave red blood cell using the finite-difference time-domain method. *Journal of Biomedical Optics*, 10(2):024022, 2005.

[35] Benjamin Rappaz, Pierre Marquet, Etienne Cuche, Yves Emery, Christian Depeursinge, and Pierre J. Magistretti. Measurement of the integral refractive index and dynamic cell morphometry of living cells with digital holographic microscopy. *Optics Express*, 13(23):9361, 11 2005.

[36] Wonshik Choi, Christopher Fang-Yen, Kamran Badizadegan, Seungeun Oh, Niyom Lue, Ramachandra R. Dasari, and Michael S. Feld. Tomographic phase microscopy. *Nature Methods*, 4(9):717–719, 9 2007.

[37] Marshall Don Graham. The Coulter Principle: Foundation of an industry, 2003.

[38] Sean C. Bendall, Garry P. Nolan, Mario Roederer, and Pratip K. Chattopadhyay. A deep profiler's guide to cytometry. *Trends in Immunology*, 33(7):323–332, 7 2012.

[39] Howard M. Shapiro. Overture. In *Practical Flow Cytometry*, pages 1–60. John Wiley & Sons, Inc., Hoboken, NJ, USA, 1 2005.

[40] Xuantao Su, Shanshan Liu, Xu Qiao, Yan Yang, Kun Song, and Beihua Kong. Pattern recognition cytometry for label-free cell classification by 2D light scattering measurements. *Cell analysis*, 23(21), 2015.

[41] David A. Basiji, William E. Ortyn, Luchuan Liang, Vidya Venkatachalam, and Philip Morrissey. Cellular Image Analysis and Imaging by Flow Cytometry. *Clinics in Laboratory Medicine*, 27(3):653–670, 9 2007.

[42] Kenneth M. Jacobs, Jun Q. Lu, and Xin-Hua Hu. Development of a diffraction imaging flow cytometer. *Optics Letters*, 34(19):2985–2987, 2009.

[43] Kenneth M. Jacobs, Li V. Yang, Junhua Ding, Andrew E. Ekpenyong, Reid Castellone, Jun Q. Lu, and Xin-Hua Hu. Diffraction imaging of spheres and melanoma cells with a microscope objective. *Journal of Biophotonics*, 2(8-9):521–527, 9 2009.

[44] Ke Dong, Yuanming Feng, Kenneth M. Jacobs, Jun Q. Lu, R. Scott Brock, Li V. Yang, Fred E. Bertrand, Mary A. Farwell, and Xin-Hua Hu. Label-free classification of cultured cells through diffraction imaging. *Biomedical Optics Express*, 2(6):1717, 6 2011.

[45] Edmund S. Cibas and Barbara S. Ducatman. *Cytology: Diagnostic Principle and Clinical Correlates*. Saunders, an imprint of Elsevier Inc., Philadelphia, PA, third edition, 2009.

[46] Richard W. Light. *Pleural disease*. LIPPINCOTT WILLIAMS & WILKINS, a Wolters Kluwer business, Philadelphia, PA 19103 USA, sixth edition, 2013.

[47] Mateen H. Uzbeck, Francisco A. Almeida, Mona G. Sarkiss, Rodolfo C. Morice, Carlos A. Jimenez, Georgie A. Eapen, and Marcus P. Kennedy. Management of malignant pleural effusions. *Advances in Therapy*, 27(6):334–347, 2010.

[48] A. M. Egan, D. Mcphillips, S. Sarkar, and D. P. Breen. Malignant pleural effusion. *Qjm*, 107(3):179–184, 2014.

[49] William W. Johnston. The Malignant Pleural Effusion. *Cancer*, 56(4):905–909, 1985.

[50] Steven E. MUTSAERS. Mesothelial cells: Their structure, function and role in serosal repair. *Respirology*, 7(3):171–191, 9 2002.

[51] Ebru Cakir, Funda Demirag, Mehtap Aydin, and Yurdanur Erdogan. A review of uncommon cytopathologic diagnoses of pleural effusions from a chest diseases center in Turkey. *CytoJournal*, 8(1):13, 2011.

[52] Morgan R. Davidson, Adi F. Gazdar, and Belinda E. Clarke. The pivotal role of pathology in the management of lung cancer. *Journal of thoracic disease*, 5 Suppl 5(SUPPL.5):463–78, 10 2013.

[53] John D. Minna, Jack A. Roth, and Adi F. Gazdar. Focus on lung cancer. *Cancer Cell*, 1(1):49–52, 2 2002.

[54] James G. Ravenel. Lung cancer staging. *Seminars in Roentgenology*, 39(3):373–385, 7 2004.

[55] Alan D. L. Sihoe and Anthony P. C. Yim. Lung cancer staging. *Journal of Surgical Research*, 117(1):92–106, 3 2004.

[56] Wei-Bing Yang, Qiu-Li Liang, Zhi-Jian Ye, Chun-Mi Niu, Wan-Li Ma, Xian-Zhi Xiong, Rong-Hui Du, Qiong Zhou, Jian-Chu Zhang, and Huan-Zhong Shi. Cell Origins and Diagnostic Accuracy of Interleukin 27 in Pleural Effusions. *PLoS ONE*, 7(7), 2012.

[57] Junhua Ding, Dongmei Zhang, and Xin-Hua Hu. A Framework for Ensuring the Quality of a Big Data Service. In *2016 IEEE International Conference on Services Computing (SCC)*, pages 82–89. IEEE, 6 2016.

[58] Ying Zhang, Yuanming Feng, Calvin R. Justus, Wenhuan Jiang, Zhigang Li, Jun Q. Lu, R. Scott Brock, Matthew K. Mcpeek, Douglas A. Weidner, Li V. Yang, and Xin-Hua Hu. Comparative study of 3D morphology and functions on genetically engineered mouse melanoma cellsw. *Integr. Biol. Integr. Biol*, 4(4):1428–1436, 2012.

[59] Yuhua Wen, Zhan Chen, Jianfen Lu, Elizabeth Ables, Jean-Luc Scemama, Li V. Yang, Jun Q. Lu, and Xin-Hua Hu. Quantitative analysis and comparison of 3D morphology between viable and apoptotic MCF-7 breast cancer cells and characterization of nuclear fragmentation. *PLoS ONE*, 12(9)(e0184726):1–12, 2017.

[60] C. Bouveyron, S. Girard, and C. Schmid. High-dimensional data clustering. *Computational Statistics & Data Analysis*, 52(1):502–519, 9 2007.

[61] Rama Chellappa, Ashok Veeraraghavan, Narayanan Ramanathan, Chew-Yean Yam, Mark S. Nixon, Ahmed Elgammal, Jeffrey E. Boyd, James J. Little, Niels Lynnerup, Peter K. Larsen,

and Douglas Reynolds. Gaussian Mixture Models. In *Encyclopedia of Biometrics*, pages 659–663. Springer US, Boston, MA, 2009.

[62] N. M. Laird A. P. Dempster and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[63] Noam Shental, Aharon Bar-Hillel, Tomer Hertz, and Daphna Weinshall. Computing Gaussian Mixture Models with EM using Equivalence Constraints. In *Advances in Neural Information Processing Systems*, pages 465–472, 2004.

[64] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575, 1 2003.

[65] Johannes Blömer and Kathrin Bujna. Adaptive Seeding for Gaussian Mixture Models. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9652 LNAI, pages 296–308. Springer, Cham, 4 2016.

[66] Volodymyr Melnykov and Igor Melnykov. Initializing the EM algorithm in Gaussian mixture models with an unknown number of components. *Computational Statistics & Data Analysis*, 56(6):1381–1395, 6 2012.

[67] Marina Meila and David Heckerman. An Experimental Comparison of Several Clustering and Initialization Methods. *Archives of Internal Medicine*, 171(1):386–395, 1 2013.

[68] Peter Willett. Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management*, 24(5):577–597, 1 1988.

[69] MathWorks. Fit Gaussian mixture model to data - MATLAB fitgmdist.

[70] Moises Noe Sanchez Garcia. *Fractal Dimension for Clustering and Unsupervised and Supervised Feature selection*. PhD thesis, Cardiff University, 2011.

[71] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 12 1974.

[72] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

[73] Xiaoyi Chen, Milena Hasan, Valentina Libri, Alejandra Urrutia, Benoît Beitz, Vincent Rouilly, Darragh Duffy, Étienne Patin, Bernard Chalmond, Lars Rogge, Lluis Quintana-Murci, Matthew L. Albert, and Benno Schwikowski. Automated flow cytometric analysis across large numbers of samples and cell types. *Clinical Immunology*, 157(2):249–260, 4 2015.

[74] Ken Aho, Dewayne Derryberry, and Teri Peterson. Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, 95(3):631–636, 3 2014.

[75] James Coates, Luis Souhami, and Issam El Naqa. Big Data Analytics for Prostate Radiotherapy. *Frontiers in Oncology*, 6(June):149, 2016.

[76] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, Jean Yves Tinevez, Daniel James White, Volker Hartenstein, Kevin Eliceiri, Pavel Tomancak, and Albert Cardona. Fiji: An open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682, 2012.

[77] Caroline A. Schneider, Wayne S. Rasband, and Kevin W. Eliceiri. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, 9(7):671–675, 7 2012.

[78] Curtis T. Rueden, Johannes Schindelin, Mark C. Hiner, Barry E. DeZonia, Alison E. Walter,

Ellen T. Arena, and Kevin W. Eliceiri. ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics*, 18(1):1–26, 2017.

[79] Jun Zhang, Gang Wang, Yuanming Feng, and Yu Sa. Comparison of contourlet transform and gray level co-occurrence matrix for analyzing cell-scattered patterns. *Journal of Biomedical Optics*, 21(8):086013, 8 2016.

[80] Robert Barer. Refractometry and Interferometry of Living Cells. *Journal of the Optical Society of America*, 47(6):545, 6 1957.

[81] Moritz Friebel and Martina Meinke. Model function to calculate the refractive index of native hemoglobin in the wavelength range of 250-1100 nm dependent on concentration. *Applied Optics*, 45(12):2838, 4 2006.

[82] Oana C. Marina, Claire K. Sanders, and Judith R. Mourant. Correlating light scattering with internal cellular structures. *Biomedical Optics Express*, 3(2):296, 2 2012.

[83] M. A. Yurkin and A. G. Hoekstra. The discrete dipole approximation: An overview and recent developments. *Journal of Quantitative Spectroscopy & Radiative Transfer*, 106:558–589, 2007.

[84] Ran Pan, Yuanming Feng, Yu Sa, Jun Q Lu, Kenneth M Jacobs, and Xin-Hua Hu. Analysis of diffraction imaging in non-conjugate configurations. *Optics Express*, 22(25):31568, 12 2014.

[85] Wenjin Wang, Yuhua Wen, Jun Q. Lu, Lin Zhao, Safaa A. Al-Qaysi, and Xin-Hua Hu. Rapid classification of micron-sized particles of sphere, cylinders and ellipsoids by diffraction image parameters combined with scattered light intensity. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 224:453–459, 2 2019.

[86] Scott Prahl. Mie Scattering Calculator.

[87] M. I. Mishchenko and J. W. Hovenier. Depolarization of light backscattered by randomly oriented nonspherical particles. *Optics Letters*, 20(12):1356, 6 1995.

[88] Punal M. Arabi, Gayatri Joshi, and N. Vamsha Deepa. Performance evaluation of GLCM and pixel intensity matrix for skin texture analysis. *Perspectives in Science*, 8:203–206, 2016.

[89] Dhanashree Gadkari. *Image Quality Analysis Using GLCM*. PhD thesis, University of Central Florida, 2004.

[90] Andrea Baraldi and Flavio Parmiggiani. An investigation of the textural characteristics associated with gray level cooccurrencematrix statistical parameters, 1995.

[91] Robert M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[92] Taewoo Kim, Renjie Zhou, Mustafa Mir, S. D. Babacan, P. S. Carney, Lynford L. Goddard, and Gabriel Popescu. White-light diffraction tomography of unlabelled live cells. *Nature Photonics*, 8(3):256–263, 2014.

[93] Francesco Merola, Pasquale Memmolo, Lisa Miccio, Roberto Savoia, Martina Mugnano, Angelo Fontana, Giuliana D'Ippolito, Angela Sardo, Achille Iolascon, Antonella Gambale, and Pietro Ferraro. Tomographic flow cytometry by digital holography. *Light: Science and Applications*, 6(4):1–7, 2017.

[94] Yu Sa, Yuanming Feng, Kenneth M. Jacobs, Jun Yang, Ran Pan, Ioannis Gkigkitzis, Jun Q. Lu, and Xin-Hua Hu. Study of low speed flow cytometry for diffraction imaging with different chamber and nozzle designs. *Cytometry Part A*, 83(11):1027–1033, 11 2013.

[95] Yuanming Feng, Ning Zhang, Kenneth M. Jacobs, Wenhuan Jiang, Li V. Yang, Zhigang Li, Jun Zhang, Jun Q. Lu, and Xin-Hua Hu. Polarization imaging and classification of Jurkat T and Ramos B cells using a flow cytometer. *Cytometry Part A*, 85(9):817–826, 9 2014.

[96] He Wang, Yuanming Feng, Yu Sa, Yuxiang Ma, Jun Q. Lu, and Xin-Hua Hu. Acquisition of cross-polarized diffraction images and study of blurring effect by one time-delay-integration camera. *Applied Optics*, 54(16):5223, 6 2015.

[97] He Wang, Yuanming Feng, Yu Sa, Jun Q. Lu, Junhua Ding, Jun Zhang, and Xin-hua Hu. Pattern recognition and classification of two cancer cell lines by diffraction imaging at multiple pixel distances. *Pattern Recognition*, 61:234–244, 1 2017.

[98] Safaa Al-Qaysi, Heng Hong, Yuhua Wen, Jun Q. Lu, Yuanming Feng, and Xin-Hua Hu. Profiling pleural effusion cells by a diffraction imaging method. In Daniel L. Farkas, Dan V. Nicolau, and Robert C. Leif, editors, *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XVI*, page 57. SPIE, 2 2018.

[99] M. M. Mukaka. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal : the journal of Medical Association of Malawi*, 24(3):69–71, 9 2012.

[100] Douglas G. Altman. *Practical statistics for medical research*. CRC press, 1990.

[101] Richard Taylor. Interpretation of the Correlation Coefficient: A Basic Review. *Journal of Diagnostic Medical Sonography*, 6(1):35–39, 1 1990.

[102] Kristin Yeager. LibGuides: SPSS Tutorials: Pearson Correlation.

[103] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 11 1973.

[104] Richard W. Conners, Mohan M. Trivedi, and Charles A. Harlow. Segmentation of a high-resolution urban scene using texture operators. *Computer Vision, Graphics, and Image Processing*, 25(3):273–310, 3 1984.

[105] Manish H. Bharati, J. Jay Liu, and John F. MacGregor. Image texture analysis: Methods and comparisons. *Chemometrics and Intelligent Laboratory Systems*, 72(1):57–71, 2004.

[106] J. M. Carstensen. *Description and Simulation of Visual Texture*. PhD thesis, Technical University of Denmark, Lyngby, Denmark, 1992.

Appendix A CELLS ISOLATION PROTOCOL

This protocol is for extracting Pleural and Peritoneal Effusion (PPE) cells from the PPE fluid. The key is to keep the fluid sample on ice during the transportation and use caution to avoid spilling during operation.

1. Keep the fluid sample on ice for transportation.
2. Centrifuge the fluid sample in 1500 RPM for 5 minutes and discard the supernatant.
3. Break up the cells sediment pellet by tapping at the bottom of the tube.
4. Use 10 mL red blood cell lysis buffer and shake at room temperature for 10 minutes to remove the red blood cells. Repeat this step if needed.
5. Filter cells using 70 μm cell strainer to get rid particles such as fat or chunk of tissue in the suspension (if needed).
6. Re-suspend the cells in PBS/BSA buffer and get them ready for counting process.

Appendix B CELL COUNTING PROTOCOL

Cell counting should be applied after cells extraction and before staining. Also, it is necessary to count the cell before confocal and p-DIFC imaging processes to estimate the viability and concentration of cells.

1. Clean the hemocytometer using 70% ethanol before use.

2. Mix the cell suspension to be counted well by either gentle agitation of the tube containing the cells or another appropriate method (e.g., using a serological pipette if required).

3. Use a 30 μL pipette to take out 15 μL the cells suspension and drop them gently in a small tube. Take out 15 μL Trypan blue and mix.

4. Use a micropipette to draw up 15 μL cell suspension containing Trypan blue. Carefully fill the hemocytometer by gently resting the end of the micropipette tip at the grooved edge of the chambers. Let, the sample to be sucked out of the micropipette by capillary action; the cell suspension should run to the edges of the grooves only. Try not to overfill the chamber. Re-load the micropipette with cells suspension and fill the second chamber if needed.

5. Focus the 10x objective of the light microscope on the hemocytometer grid lines. Focus on one set of 16 corner squares.

6. Use an eight key manual differential cell counter to count the number of cells in this area of 16 squares. When counting, always count only live cells that look healthy. Count cells that are within the chosen square and any located on the right hand or bottom boundary line. Dead cells stained blue with Trypan blue can be counted separately for a viability count.

7. Focus the microscope to another set of 16 corner squares and carry on counting until all four sets of 16 corner squares are counted.

8. According to hemocytometer design, the total number of cells in one set of 16 corner squares is equivalent to the number of cells $\times 10^4$/mL, and the total number of cells is equivalent to the sum of the total number of cells in four sets of one hemocytometer grid.

9. So divide the total count by 4 to get the average number of cell in one square. Then adjust the number cells to 1:2 dilution in Trypan blue by multiplying by 2 to adjust. These two steps are equivalent to dividing the cell count by 2. For example: if the cell count is 130, the cell density is: $\frac{130}{2} = 65 \times 10^4$/mL.

Appendix C CELL STAINING PROTOCOL

The double staining protocol is for acquiring confocal image stack.

1. Collect cells in 15 mL conical tube.
2. Centrifuge cells suspension at 1500 RPM for 5 minutes.
3. Aspirate the supernatant media on top of cells to get cell sediment pellet.
4. Tap the bottom of the tube to break up cell pellet. Then, re-suspend cells in 5 mL of media.
5. Pipette the cell suspension several times to obtain a single cell suspension.
6. Add 1 μL of nucleus stain (Syto-61) and mitochondria stain (MitoTracker Orange) to the 5 mL cell suspension. The final concentrations of stains are 1 μM and 0.2 μM for Syto-61 and MitoTracker Orange respectively.
7. Invert tube several times to mix media well.
8. Incubate stained cells suspension at 37°C and 5% CO_2 for 40 minutes.
9. Centrifuge cells suspension at 1500 RPM for 3 minutes.
10. Break up cell pellet and re-suspend cells in 2 mL of media as previous procedure is the start of first wash.
11. Incubate washed cells at 37°C and 5% CO_2 for at least 5 minutes.
12. Centrifuge washed cells at 1500 RPM for 5 minutes.

13. Aspirate the supernatant media on top of cells to get cell sediment pellet, and this is the end of the first wash.

14. If too much fluorescence background appears in the confocal images, a second wash may be taken by repeating above steps.

15. Re-suspend cell sediment pellet in 5 mL media to have concentrated cell suspension for confocal observation –cell suspension for imaging.

16. Add 150 μL to depression slide, put a glass cover slide on top and invert the assembly for the inverted microscope viewing. If cells are too dense, dilute cell suspension with more media.

Appendix D 3D PARAMETERS DEFINITION

Parameter	Symbol	Units	Definition
Cell grid perimeter	GP_c	μm	$GP_c = N_{s,cyto} a_{vxl}$ (a)
Cell surface area	S_c	μm^2	$S_c = N_{s,cyto} s_{vxl}$ (b)
Cell volume	V_c	μm^3	$V_c = (N_{v,cyto} + N_{v,nucl} + N_{v,mito}) v_{vxl}$ (c)
Surface to volume ratio of cell	SVr_c	μm^{-1}	$SVr_c = S_c / V_c$
Index of surface irregularity of cell	SIi_c	$\mu m^{-1/2}$	$SIi_c = GP_c / \sqrt{V_c}$
Cell equivalent spherical radius	ER_c	μm	$ER_c = (3V_c/4\pi)^{1/3}$
Cell volume sphericity index	VSi_c		$VSi_c = 4\pi ER_c^2 / S_c = (36\pi V_c^2)^{1/3} / S_c$
Average distance of cell membrane voxels to centroid	$< R_c >$	μm	$< R_c >= \sum_{i=1}^{N_{s,cyto}} R_c(i) / N_{s,cyto}$ (d)
Standard deviation of $< R_c >$	ΔR_c	μm	$\Delta R_c = \{\frac{1}{N_{s,cyto}} \sum_{i=1}^{N_{s,cyto}} (R_c(i) - \langle R_c \rangle)^2\}^{1/2}$
Nuclear grid perimeter	GP_n	μm	$GP_n = N_{s,nucl} a_{vxl}$
Nuclear surface area	S_n	μm^2	$S_n = N_{s,nucl} s_{vxl}$
Nuclear volume	V_n	μm^3	$V_n = N_{v,nucl} v_{vxl}$
Nuclear surface to volume ratio	SVr_n	μm^{-1}	$SVr_n = S_n / V_n$
Index of surface irregularity of nucleus	SIi_n	$\mu m^{-1/2}$	$SIi_c = GP_c / \sqrt{V_c}$
Nuclear equivalent spherical radius	ER_n	μm	$ER_n = (3V_n/4\pi)^{1/3}$
Nuclear volume sphericity index	VSi_n		$VSi_n = 4\pi ER_n^2 / S_n = (36\pi V_n^2)^{1/3} / S_n$
Average distance of nuclear membrane voxels to centroid	$< R_n >$	μm	$< R_n >= \sum_{i=1}^{N_{s,nucl}} R_n(i) / N_{s,nucl}$
Standard deviation of $< R_n >$	ΔR_n	μm	$\Delta R_n = \{\frac{1}{N_{s,nucl}} \sum_{i=1}^{N_{s,nucl}} (R_n(i) - \langle R_n \rangle)^2\}^{1/2}$
Mitochondrial grid perimeter	GP_m	μm	$GP_m = N_{s,mito} a_{vxl}$
Mitochondrial surface area	S_m	μm^2	$S_m = N_{s,mito} s_{vxl}$
Mitochondrial volume	V_m	μm^3	$V_m = N_{v,mito} v_{vxl}$
Surface to volume ratio of mitochondria	SVr_m	μm^{-1}	$SVr_m = S_m / V_m$
Index of surface irregularity of mitochondria	SIi_m	$\mu m^{-1/2}$	$SIi_m = GP_m / \sqrt{V_m}$
Mitochondrial equivalent spherical radius	ER_m	μm	$ER_m = (3/SVr_m \pi)^{1/3}$
Distance between the centroids of nucleus and cell	CD_m	μm	$\lvert \mathbf{r}_{nc} - \mathbf{r}_{cc} \rvert$ (e)
Volume ratio of nucleus to cell	Vr_{nc}		$Vr_{nc} = V_n / V_c$
Volume ratio of mitochondrion to cell	Vr_{mc}		$Vr_{mc} = V_m / V_c$

(a) a_{vxl} = voxel side length = $dx = dy \approx dz$, where dx and dy is the pixel size along $x-$ and $y - axis$, respectively, while dz is the distance between two neighboring interpolated slices.

The side length dz is obtained by requiring $dz \cdot (N_{int} + 1) \approx \Delta z \cdot f$ with N_{int} as the integer number of interpolated slices between two raw slices and $\Delta z = 0.5$ μm as the translation step size along $z - axis$, $f = 0.87$ is the correction factor for light refraction.

(b) $N_{s,(c,nucl,\,or\,mito)}$ = number of surface or membrane voxels for the organelle $(c, nucl, ormito)$, $s_{vxl} = 1.414 \cdot a_{vxl} \cdot dz$ = area of the diagonal plane of one voxel as the average surface area of the membrane voxels. A surface voxel of a specific organelle is defined as the one which has at least one of the six neighboring voxels belong to another organelle or the host medium outside of the cell.

(c) $N_{v,(c,nucl,\,or\,mito)}$ = number of volume voxels for the organelle $(c, nucl,\ or\ mito)$ which includes the surface voxels and interior voxels, $V_{vxl} =$ volume of voxel $= dx \cdot dy \cdot dz$.

(d) $R_c = |\mathbf{r}_{cs}(i) - \mathbf{r}_{cc}|$ with $\mathbf{r}_{cs}(i) = (x_{cs}(i), y_{cs}(i), z_{cs}(i))d_{av}$ as the position vector of the i^{th} voxel on the cell surface or membrane and $\mathbf{r}_{cc} = (x_{cc}, y_{cc}, z_{cc})d_{av}$ as the position vector of the cell centroid, $d_{av} = (dx + dy + dz)/3$. The component coordinates of $\mathbf{r}_{cc}$ are defined as $x_{cc} = \sum_{i=1}^{N_{v,cell}} with\, N_{v,cell} = N_{v,cyto} + N_{v,nucl} + N_{v,mito}$, etc..

(e) $\mathbf{r}_{nc} = (x_{nc}, y_{nc}, z_{nc})d_{av}$ is the position vector of the nuclear centroid with its components defined as $x_{nc} = \sum_{i=1}^{N_{v,nucl}} x(i)/N_{v,nucl}$, etc..

Appendix E GMM ALGORITHM

GMM clustering algorithm starts by fitting a Gaussian mixture distribution of clustering components k to the data matrix X of $(N \times D)$ dimensions. N represents the rows of data matrix match the PPE cells in our measurement (449 cells); D represents for the columns of the data matrix match to the extracted 3D morphological parameters (27 morphological parameters).

$$X(N \times D) = \begin{pmatrix} x_{1,1} & \cdots & x_{1,D} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,D} \end{pmatrix} \tag{E.1}$$

Assuming that the data is coming from k clusters, the algorithm will randomly split the data matrix into k sub-matrices and use maximum likelihood parameters estimation (MLPE) method to estimate the mean (μ) and covariance matrices (α)s' for each column of the data matrix. The estimated μ is $\frac{\sum x}{n}$, and the mean vectors for the k clusters are:

$$\mu(k, D) = \begin{pmatrix} \mu_{1,1} & \cdots & \mu_{1,D} \\ \vdots & \ddots & \vdots \\ \mu_{1,k} & \cdots & \mu_{k,D} \end{pmatrix}, \tag{E.2}$$

and the covariance matrices of each cluster with the corresponding column variances along the diagonal

$$\alpha_{k=1} = \begin{pmatrix} \sigma^2_{1,1} & \cdots & \sigma^2_{1,D} \\ \vdots & \ddots & \vdots \\ \sigma^2_{n,1} & \cdots & \sigma^2_{n,D} \end{pmatrix}, \alpha_{k=2} = \begin{pmatrix} \sigma^2_{n+1,1} & \cdots & \sigma^2_{n+1,D} \\ \vdots & \ddots & \vdots \\ \sigma^2_{N,1} & \cdots & \sigma^2_{N,D} \end{pmatrix} \tag{E.3}$$

both matrices will be the initial values for the expectation maximization (EMax) algorithm with assigned equal prior probabilities to each cluster. In the "Expectation" step, the algorithm calculates the probability that each data point belongs to each cluster using current mean vectors and covariance matrices. Since the algorithm deals with multiple inputs variables, the equation for the probability density function of a multivariate Gaussian is:

$$g_k(x|\mu_i\alpha_i) = \frac{1}{(2\pi)^{\frac{D}{2}}|\alpha_i|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu_i)'\alpha_i^{-1}(x-\mu_i)\} \tag{E.4}$$

At this point, we will end up with a probability density function (PDF) of (N × k) matrix,

$$PDF = \begin{pmatrix} g_{1,1} & \cdots & g_{1,k} \\ \vdots & \ddots & \vdots \\ g_{N,1} & \cdots & g_{N,k} \end{pmatrix} \tag{E.5}$$

The probability $w_i^{(k)}$ that the data point belongs to cluster k can be calculated by multiplying each PDF value by the prior probability for cluster then divide by the sum of weighted probability for each cluster. By that, the algorithm ends up with a matrix of one row per data point and one column per cluster. i.e.,

$$W(N,k) = \begin{pmatrix} w_{1,1} & \cdots & w_{1,k} \\ \vdots & \ddots & \vdots \\ w_{N,1} & \cdots & w_{N,k} \end{pmatrix} \tag{E.6}$$

In the "Maximization" step, algorithm re-calculates the cluster means and covariance based on the probabilities calculated in the expectation step. The new mean calculated for each cluster by finding the weighted average of all data points.

$$new\mu_k = \frac{\sum_{i=1}^{m} w_i^{(k)} x_i}{\sum_{i=1}^{m} w_i^{(k)}} \Rightarrow new\mu(k, D) = \begin{pmatrix} \mu_{1,1} & \cdots & \mu_{1,D} \\ \vdots & \ddots & \vdots \\ \mu_{1,k} & \cdots & \mu_{k,D} \end{pmatrix} \tag{E.7}$$

Appendix F GLCM PARAMETERS

The gray tone of a rectangular input image I with N_x horizontal resolution pixels, and N_y vertical resolution pixels is quantized by N_g levels. $L_x = 1, 2, \ldots, N_x$ and $L_y = 1, 2, \ldots, N_y$, are the horizontal and the vertical spatial domains respectively, and $G = 0, 1, 2, \ldots, N_g$ is the quantized gray-level tone. For the 8-bit gray-level image, G is equal 255, and N_g is equal 254. The set $L_x \times L_y$ represent the set of pixels of the image sorted by their row-column pixel gray-level labels. It is assumed that the texture information in an image I is contained in the overall or average spatial relationship. Let's denote $p_{i,j}$ as the "co-occurrence" frequency of two neighboring pixels with gray-level values i and j that are separated by certain distance d. Where $d = 1$ can be defined nearest neighbored pixels at specific angular direction θ. Which is equal to $0°, 45°, 90°, and 135°$ for horizontal, vertical, and diagonal direction. [103].

Figure F.1 shows an example of GLCM calculation of an input image with four gray-level values. The pixel value representation for the input image is ranging from 0 the black color to 3 the white color. The gray-level co-occurrence frequencies can be calculated and represented by matrices for horizontal ($\theta = 0°$), vertical ($\theta = 90°$), and diagonal direction ($\theta = 45° and 135°$) respectively.

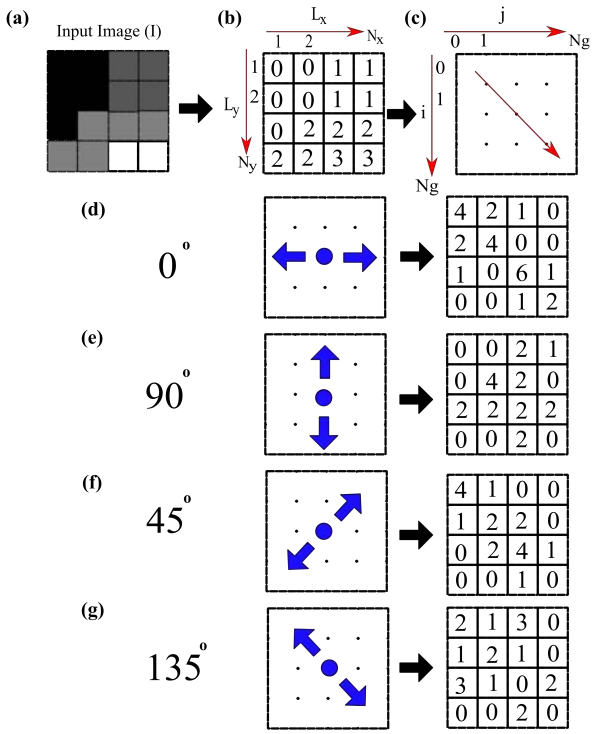

Figure F.1 Calculation of GLCM (a) The input image sample with gray-levels. (b) Pixel value representation of the input image with four gray-level. (c) The co-occurrence matrix representation. (d), (e), (f), and (g) The co-occurrence matrix of the input image with $d = 1$ and $\theta = 0°, 90°, 45°$, and $135°$ respectively.

Let's assume $J(x, y)$ is the 12-bit diffraction image before normalization, and $I(x, y)$ is the 8-bit diffraction image data after normalization. The GLCM of $I(x, y)$ elements usually represented by the normalized frequencies $p(i, j)$. Where $p(i, j) = P(i, j)/R$, and R is the total pair of the neighboring pixels for calculated matrix P [103]. The probability of the reference pixel value i in the grey level image is

$$p_x(i) = \sum_{i=0}^{G-1} p(i, j), \tag{F.1}$$

and the probability of the neighbor pixel value j in grey level image is

$$p_y(j) = \sum_{j=0}^{G-1} p(i, j), \tag{F.2}$$

While, the probabilities of the sum and the difference of the main diagonal line with gray-level image are

$$\begin{aligned} p_{x+y}(k) &= \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} p(i,j), \quad k = i + j, (k = 0, 1, 2, \ldots, 2G-2), \\ p_{x-y}(l) &= \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} p(i,j), \quad l = |i - j|, (l = 0, 1, 2, \ldots, G-1), \end{aligned} \tag{F.3}$$

$p_{x+y}(k)$ and $p_{x-y}(l)$ provide a histogram of the of gray-levels from pixel pairs of input image I [103].

In this study, a total of 17 parameters have been extracted for each of the p-DI pairs which include 15 texture parameters defined through $p(i, j)$ and 2 parameters of maximum and minimum pixel intensities from the 12-bit image $J(z, y)$. Therefore, each p-DI pair yields 34 parameters to represent each imaged cell by the p-DIFC method. The definitions of the GLCM parameters extracted from a diffraction image are discussed below.

1. The average value of gray-level p_x and p_y is [104]

$$\begin{aligned} \mu_x &= \sum_{i=0}^{G-1} i p_x(i) = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} i p(i,j), \\ \mu_y &= \sum_{j=0}^{G-1} j p_y(j) = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} j p(i,j), \end{aligned} \tag{F.4}$$

2. The sum average (SAV) is

$$SAV = \sum_{i=0}^{2G-2} k p(i,j), \tag{F.5}$$

3. The difference average, which also known as dissimilarity (DIS) is [91]

$$DIS = \sum_{i=0}^{2G-2} l p(i,j), \tag{F.6}$$

4. The standard deviation of the gray-level is

$$\sigma_x = \sqrt{\sum_{i=0}^{G-1}(i-\mu_x)^2 p_x(i)},$$
$$\sigma_y = \sqrt{\sum_{j=0}^{G-1}(j-\mu_y)^2 p_y(j)}, \tag{F.7}$$

5. The variance (VAR) is [90]

$$VAR_x = \sum_{i=0}^{G-1}(i-\mu_x)^2 p_x(i),$$
$$VAR_y = \sum_{j=0}^{G-1}(j-\mu_y)^2 p_y(j), \tag{F.8}$$

   VAR is relatively high for the elements that differ from the average value of $p(i, j)$ [90].

6. Angular Second Moment (ASM) measures the uniformity of an image. A uniform image contains only similar gray-level pixels, resulting in a GLCM with a few but relatively high values of $p(i, j)$. Thus, the higher values of ASM indicate more uniform texture images [103].

$$ASM = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1}\{p(i,j)\}^2, \tag{F.9}$$

7. Contrast (CON) in an image presents the amount of local variation measure within the GLCM [105]. An image with a large amount of local variation has a higher contrast value.

$$CON = \sum_{j=0}^{G-1} l^2 p(i,j)^2, \quad l = |i-j|, \tag{F.10}$$

8. Correlation is a measure of gray-level linear dependency in a gray-level image. An image with lower correlation values consists of mostly constant gray-level values [90].

$$COR = \frac{\sum_{i=0}^{G-1}\sum_{j=0}^{G-1}(i-\mu_x)(j-\mu_x)p_x(i,j)}{\sigma_x\sigma_y},$$
$$COR = \frac{\sum_{i=0}^{G-1}\sum_{j=0}^{G-1}(ij)p_x(i,j)\mu_x\mu_y}{\sigma_x\sigma_y}, \tag{F.11}$$

9. Inverse Difference Moment (IDM) is a measure of image local homogeneity. Due to the weighting factor $(1+(i-j)^2)^{-1}$, the IDM will get high contributions from homogeneous areas ($i = j$), and small contribution from inhomogeneous areas ($i \neq j$) [90].

$$IDM = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1}\frac{1}{1+(i-j)^2}p(i,j), \tag{F.12}$$

10. Entropy (ENT) is measure of image texture randomness due to intensity distribution. Homogeneous images have high entropy, while inhomogeneous images have low entropy [90].

$$ENT = -\sum_{i=0}^{G-1}\sum_{j=0}^{G-1}p(i,j)\cdot log(p(i,j)), \tag{F.13}$$

11. The sum entropy (SEN) is

$$SEN = -\sum_{i=0}^{2G-2}p_{x+y}(k)\cdot log(p_{x+y}(k)), \tag{F.14}$$

12. The difference entropy is

$$DEN = -\sum_{i=0}^{G-1}p_{x-y}(l)\cdot log(p_{x-y}(l)). \tag{F.15}$$

13. The sum variance (SVA) is [106]

$$SVA = -\sum_{i=0}^{2G-2}(k-SEN)^2 p_{x+y}(k) \tag{F.16}$$

14. The Difference variance (DVA) is [106]

$$DVA = CON - (\sum_{i=0}^{G-1} kp_{x-y}(k))^2 \tag{F.17}$$

15. Cluster shade (CLS) is a measure of GLCM skewness. It can be used to predicts matrix uniformity. High CLS means more asymmetry of the image [91].

$$CLS = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1}(i + j - \mu_x - \mu_y)^3 p(i, j), \tag{F.18}$$

16. Cluster prominence (CLP) is a fourth power measure of asymmetry. When CLP value is high, the image is not symmetric. [91].

$$CLP = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1}(i + j - \mu_x - \mu_y)^4 p(i, j), \tag{F.19}$$

Appendix G EXAMPLE SCRIPT FOR ADDA

```
# This shell script file created to simulate micro-sphere.
# The absolute refractive index of the sphere is m= 1.588 + 0.00034 at λ =0.532μm.
# The absolute refractive index of the host media is 1.334.
# The relative refractive index of the sphere is 1.588/1.334 = 1.190.
# The relative incident wavelength is 0.399μm.
# Theta range from -180 to 180 degree, and phi=90 degree.

for a in 5
do

INPUT_DIR="/home/Single_Sphere_OCM/$a"
OUTPUT_DIR="/home/Single_Sphere_OCM/$a"
SCAT_DIR="/home/Single_Sphere_OCM"
ADDA_MPI="/home/adda_1.3b4/src/mpi/adda_mpi"
ARGS_MPI="-lambda 0.399 -shape read $INPUT_DIR/Geometry_files.dat -dpl 1
-scat_grid_inp $SCAT_DIR/scat_params_a.dat -store_scat_grid"
ARGS_MPI_INDEX="-m ‘cat $INPUT_DIR/refractive_index.txt‘"
EXEC_MPI_ARG="mpiexec -n 72"
cd $OUTPUT_DIR
$EXEC_MPI_ARG $DIR_NAME $ADDA_MPI $ARGS_MPI $ARGS_MPI_INDEX
done
```

Appendix H INSTITUTIONAL REVIEW BOARD (IRB)

EAST CAROLINA UNIVERSITY

University & Medical Center Institutional Review Board

4N-64 Brody Medical Sciences Building· Mail Stop 682

600 Moye Boulevard · Greenville, NC 27834.

IRB approval reference number: UMCIRB 10-0016