Thesis submitted to the Ecole Polytechnique de Louvain in partial fulfillment for the degree of Electrical Engineering

Visual feature extraction to discriminate nucleolus phenotypes in fluorescence microscopy

Pia Muñoz Trallero

Supervisors:
Prof. Christophe De Vleeschouwer
Dr. Pascaline Parisot
Referees:
Prof. Veronica Vilaplana
Jury members:
Prof. Laurent Jacques

Louvain-la-Neuve, 2014

Contents

Abbreviations **1**

Biological Concepts **3**

1 Introduction **7**

2 Images: From culture of cells to cells **11**
2.1 Phenotypes . . . 11
2.2 Database . . . 11
2.3 Segmentation of cells . . . 13
2.3.1 Segmentation of nuclei . . . 13
2.3.2 Obtaining nucleoli individual images . . . 15

3 Mean shift algorithm **17**
3.1 General algorithm . . . 17
3.2 Adaptation to our case . . . 19
3.2.1 Adaptation from points to image intensities . . . 19
3.2.2 Kernel function . . . 19
3.2.3 Clustering . . . 21
3.2.4 Peak seeking . . . 22
3.3 Pseudo-code . . . 22
3.3.1 Mean shift program . . . 22
3.3.2 Subprograms . . . 24
3.4 Post-processing . . . 25

4 Nucleolus phenotype discrimination **31**
4.1 Image pre-processing . . . 31
4.2 Which features? . . . 32
4.2.1 Intensity . . . 34
4.2.2 Mean shift modes . . . 34
4.2.3 Lines . . . 36
4.2.4 Edge detection regions . . . 38
4.2.5 Area ratios . . . 41
4.2.6 Grey Level Aura Matrices . . . 41

5 Experiments and validations **43**
5.1 Validation methodologies . . . 43
5.1.1 Segmentation parameters . . . 43
5.1.2 Characterizing the nucleolus . . . 44
5.1.3 Phenotypes classification . . . 51

5.2 Results 56
5.2.1 Intensity 56
5.2.2 Mean-shift 57
5.2.3 Edge detection 85
5.2.4 Area ratios 100

6 Future work **109**

7 Conclusions **111**

Bibliography **113**

Annex **115**

Abbreviations

DAPI: 4´,6-diamidino-2-phenylindole

DFC: Dense fibrilar components

DNA: Deoxyribonucleic acid

FBL: Fibrillarin

FC: Fibrillar centers

GC: Granular components

mRNA: Messenger RNA

MS: Mean-Shift

NCL: Nucleolin

NPM1: Nucleophosmin

RNA: Ribonucleic acid

RNAi: RNA interference

RNP: Ribonucleoprotein

RpL27: Ribosomal protein L27

rRNA: Ribosomal RNA

siRNA: Small interfering RNA or Silencer RNA

snRNP: Small nuclear ribonucleoprotein

Tif1A: Transcriptional intermediary factor 1 Alpha

GFP: Green Fluorescent Protein

Biological Concepts

In this section we introduce some concepts that are mainly employed in sections 1 and 2 and are useful to the untrained reader to understand the biological background and bases of this thesis:

DAPI (4´,6-diamidino-2-phenylindole) [[1], Introduction, p.1], [[2], Introduction, p.2] is a fluorescent stain that allows easy visualization of the nucleus in interphase cells. It can pass through an intact cell membrane therefore it can be used to stain both live and fixed cells.

Deoxyribonucleic acid (DNA) [3] is a nucleic acid that contains the genetic information for cell growth, division, and function. In eukaryotes, it is chiefly founded in the nucleus of cells.

Depletion [3] is the act or process of emptying or removal of a fluid, as the blood.

Fibrillarin-GFP [[4], Introduction, p.2] is a truncated fibrillarin mutant expressed as fusion proteins with GFP. It is used as a marker for fibrillarin protein localization.

Green Fluorescent Protein (GFP) [[5], Introduction, p.1] traditionally refers to the protein first isolated from the jellyfish Aequorea victoria. Gene fusions using GFP are an alternative to immunofluorescence microscopy as a report for protein localization and their use, in general, improves the sensitivity of molecular detection. This protein has the attribute of exhibiting fluorescence by excitation with specific wavelength light which allows protein localization in living cells.

HeLa cell [6] is the most widely used immortal cell line in biomedical research. HeLa cells were extracted in 1951 from cervical cancer cells taken from a 31 year old African American woman named Henrietta Lacks.

Messenger Ribonucleic acid (mRNA) [3] is a type of RNA that codes the chemical blueprint for a protein during protein synthesis. In eukaryotes, the mRNA is produced in the nucleus during transcription, which is the first step of gene expression in which a particular segment of DNA is copied into RNA by the enzyme RNA polymerase. When the mRNA has been completely processed, it is called a mature mRNA, which will then be transported for translation into the cytoplasm through the nuclear pore.

Nucleolus [[7], p.1-3] is a dynamic structure that disassembles and reforms during each cell cycle around the rRNA gene clusters. Within the nucleolus, three distinct subcompartments are described: fibrillar centers (FC), dense fibrillar components (DFC) and granular components (GC).

Phenotype [3] is the physical appearance or biochemical characteristic of an organism

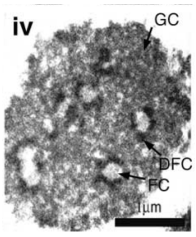


Figure 0.0.1: Subcompartments of the nucleolus (Figure from [[7], Fig. 1.(B).iv, p. 3])

as a result of the interaction of its genotype and the environment, which can be regarded as a natural environment or a built environment.

Ribonucleic acid (RNA) [3] is a nucleic acid that plays a role in transferring information from DNA to protein-forming system of the cell. RNA is a molecule consisting of a long linear chain of nucleotides. Each nucleotide unit is comprised of a sugar, phosphate group and a nitrogenous base. In eukaryotes, it is found in the nucleus and in the cytoplasm. RNAs are involved in protein synthesis (mRNA, rRNA), post-transcriptional modification or DNA replication and gene regulation (siRNA). Together with DNA, RNA comprises the nucleic acids, which, along with proteins, constitute the three major macromolecules essential for all known forms of life.

Ribonucleic acid interference (RNAi) [[8], p.1-5] is a mechanism for gene regulation that can either reduce or abolish gene activity or induce or stimulate gene activity, typically by causing the destruction of specific mRNA molecules.

Ribosomes [9] are cytoplasmic granules composed of RNA and protein, at which protein synthesis takes place. The RNA present in ribosomes, called ribosomal RNA (rRNA), is produced in the nucleolus.

Small interfering RNA (siRNA) [[8], p.1-5] is a short double-stranded RNA (dsRNA) synthesized within the cell and has the property of being able to reduce or abolish gene activity by RNAi-like mechanisms. In biomedical research, siRNA is used as a powerful tool to experimentally elucidate the function of essentially gene in a cell. The injection of a siRNA within the cell leads to an efficient loss of the target mRNA, which is cleaved and subsequently degraded. This process is known as silencing.

Transfection [[10], Introduction, p.1] is the process of deliberately introducing nucleic acids into eukaryotic cells by non-viral methods. This gene transfer technology is a powerful tool to study gene function and protein expression in the context of a cell.

Genes summaries [**[11]**,Summaries]

Fibrillarin (FBL): This gene product is a component of a nucleolar small nuclear ribonucleoprotein (snRNP) particle thought to participate in the first step in processing preribosomal RNA. It is associated with the U3, U8, and U13 small nuclear RNAs and is located in the DFC of the nucleolus. The encoded protein contains

a N-terminal repetitive domain that is rich in glycine and arginine residues, like fibrillarins in other species. Its central region resembles an RNA-binding domain and contains an RNP consensus sequence. Antisera from approximately 8% of humans with the autoimmune disease scleroderma recognize fibrillarin.

Nucleolin (NCL): It is an eukaryotic nucleolar phosphoprotein involved in the synthesis and maturation of ribosomes. It is located mainly in DFC of the nucleolus. Human NCL gene consists of 14 exons with 13 introns and spans approximately 11kb. The intron 11 of the NCL gene encodes a small nucleolar RNA, termed U20.

Nucleophosmin (NPM1): This gene encodes a phosphoprotein which moves between the nucleus and the cytoplasm. The gene product is thought to be involved in several processes including regulation of the ARF/p53 pathway. A number of genes have been characterized as fusion partners, in particular the anaplastic lymphoma kinase gene on chromosome 2. Mutations in this gene are associated with acute myeloid leukemia. More than a dozen pseudogenes of this gene have been identified.

Ribosomal protein L27 (RpL27): This gene encodes a ribosomal protein that belongs to the L27E family of ribosomal proteins and is located in the cytoplasm. As is typical for genes encoding ribosomal proteins, there are multiple processed pseudogenes of this gene dispersed through the genome.

Transcriptional intermediary factor 1 Alpha (Tif1A): Is the protein encoded by the gene TRIM24. Tif1A localizes to nuclear bodies and is thought to be associated with chromatin and heterochromatin-associated factors. The protein is a member of the tripartite motif (TRIM) family.

1 Introduction

The nucleolus is a specialized sub-cellular functional domain found in the nucleus of eukaryotic cells. It is not bound by any membrane, what makes it extremely dynamic. From a functional point of view, it is mainly involved in the assembly of ribosomes. After being produced in the nucleolus, ribosomes are exported to the cytoplasm where they translate messenger RNA (mRNA) into proteins. Because it directly impacts the synthesis of proteins and thus the functions of the cell, the nucleolus is an important organelle of the cell. Interestingly, earlier studies have shown that there is a strong correlation between the 3-D structure of the nucleolus and the potential diseases affecting the cell. Healthy or normal cells are characterized by spherical nucleoli, whilst diseased or abnormal cells present conformation artifacts that can be visually observed through (epi-)fluorescence microscopy. We use the term *phenotype* to denote a specific kind of nucleolus conformation artifact.

The research unit of *RNA Metabolism* of the Faculty of Science in the Université Libre de Bruxelles (ULB), which is linked to the Center for Microscopy and Molecular Imaging (CMMI) in Gosselies, investigates how gene inhibition affects the nucleolar structure. It aims at identifying which proteins - among the 700 proteins present in the nucleolus - are required to maintain a normal conformation of the nucleolus. It also aims at characterizing the different kinds of phenotypes resulting from the inhibition of genes encoding nucleolar components. Based on the assumption that proteins that induce similar deformations of the nucleolus are involved in similar (dis)functions of the cell, the study will help to better understand the role of each protein.

RNA interference (RNAi) is a molecular biology technique, which allows to down-regulate the expression of specific genes of interest. Typically, small RNA molecules named silencers (siRNAs), are transfected into cells where they bind to specific segments of a targeted mRNA molecule triggering its degradation. RNAi has become an extremely powerful research tool, especially in cell cultures, because synthetic silencers introduced into cells can selectively and robustly suppress the expression of specific genes of interest. RNAi may be used to implement large-scale screens that systematically shut down each gene in the cell, which can help to elucidate the workings of normal and diseased cells, for example, by identifying the proteins that are required for a particular cellular process. The research unit of *RNA Metabolism* of the ULB has recently conducted a large scale RNAi screen in HeLa cells by RNAi where each of a selected set of 700 nucleolar proteins were depleted by 3 individual silencers. The HeLa cell line used in this work expressed a fluorescently-tagged nucleolar protein such that the organization of the nucleolus could be followed by microscopy. A large library of images was generated (16 images for each silencer; 3 silencers per gene; 700 genes). In a preliminary analysis of this dataset, a limited number of phenotypes (about a dozen)

have been identified through visual inspection of these images.

The objective of this master thesis project is to go one step further and to develop image analysis tools for automated analysis and quantitative characterization of the visual appearance of the nucleolus conformation such as to classify unambiguously these phenotypes. The main steps of the master thesis are:

- The automatic segmentation of the nucleolus.
- The representation of the segmented nucleolus in terms of a limited number of fundamental structures. This will be carried out with mean shift algorithm, explained in detail in Section 3, and with the edge detection tool introduced in section 4.2.3.
- The definition of a number of features characterizing the spatial organization of those fundamental structures that approximate the nucleolus.
- Phenotypes classification and clustering based on those features.

To infer which features best help discriminate phenotypes, we might rely on the fact that each experiment (for example, corresponding to one silencer) generates many samples with similar phenotypes. Hence, those phenotypes should be part of the same class or cluster.

Briefly, our personal solution of the problem is illustrated in Fig. 1.0.1.

Finally, this memory is structured in 7 main chapters. First the introduction, where we present the biological problem and our particular solution. Second the database is presented and the segmentation of the nucleolus is done. In Section 3, we present mean shift algorithm in detail and our particular adaptation of the algorithm to images. Section 4 contains the whole set of features proposed features to test in Section 5. Section 5 is divided in two blocks. In the first block we discuss the different parameters applied to our solution and introduce the metrics used to consider whether a feature is discriminant. The second block contains the set of results of our experiments. Finally, Sections 6 and 7 are the sections dedicated to the perspective and the conclusions of the thesis.

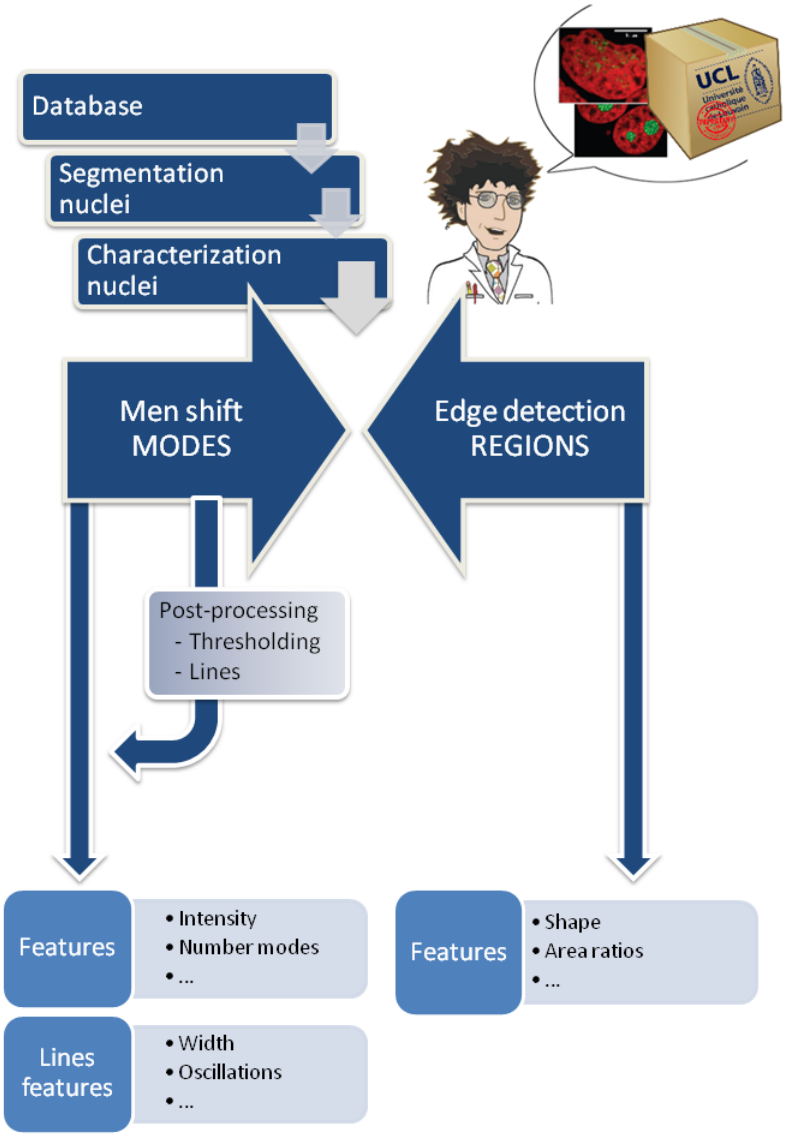


Figure 1.0.1: Solution of the problem

2 Images: From culture of cells to cells

2.1 Phenotypes

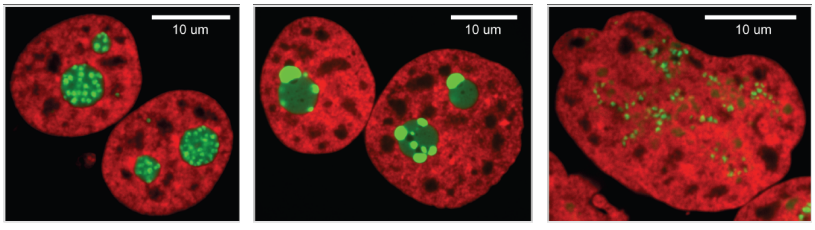
When the nucleolus is disintegrated by the action of a drug or the removal of a specific protein within the cell, we can basically describe two main phenotypes of nucleolar destructuration:

1. Unfolding phenotype: Unraveling of the nucleolus which looks like the unfolding of a necklace. By contrast, a normal phenotype has the appearance of a compacted necklace.
2. Capping phenotype: Delocalization and reassembling of all the nucleolar material at the periphery of the nucleolus. Thereby, the phenotype is more a compaction than an unfolding.

These phenotypes actually occur when a specific function of the nucleolus is affected. As we have already discussed, the function of the nucleolus is to produce ribosomes, which are made of rRNAs.Those RNAs are first synthesized as one big RNA in the middle part of the nucleolus and then this big RNA is cleaved and modified in the peripheral parts of the nucleolus. On one side, if you touch the big RNA synthesis function, a capping phenotype will manifest. On the other side, if you affect the RNA cleavage and modification function, you will get an unfolding phenotype. Fig. 2.1.1 shows an example contextualized within a cell.

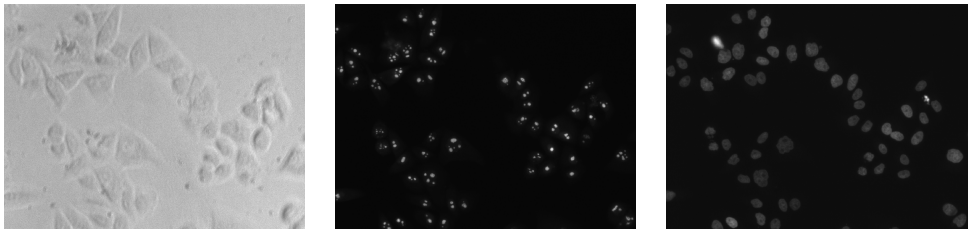
2.2 Database

Our database contains images from 13 independent wells. The method used for the analysis is a depletion of specific genes by siRNAs using a cell line expressing Fibrillarin-



(a) Normal nucleoli **(b)** Capping phenotype **(c)** Unfolding phenotype

Figure 2.1.1: Phenotypes at high resolution



(a) w1 illumination **(b)** w2 illumination **(c)** w3 illumination

Figure 2.2.1: Three different illuminations of a sample site from a control without addition of any siRNA

GFP and with nuclei stained in blue with DAPI. These specific genes are GFP, Fibrillarin, Nucleophosmin, Nucleolin, Tif1A and RpL27. For each well, 16 sites have been imaged with three different illuminations, which are illustrated in Fig. 2.2.1:

- **w1:** corresponds to a transmitted light image and is not used for the analysis
- **w2:** corresponds to the image in the GFP channel, which correspond to the nucleoli
- **w3:** corresponds to the image in the blue channel highlighting the DAPI marker staining the nuclei of the cells

Accordingly, w3 is used as a seeding image to segment the nuclei and then w2 is used to characterize the nucleoli.

Characteristics of the 13 wells:

- 8 wells correspond to the controls present within each plate of the screening:
  - A12: control without addition of any siRNA. Corresponds to normal cells.
  - B12: control with a siRNA scramble with no specific target in the genome. Corresponds also to normal cells.
  - C12: cells + siRNA against GFP. Corresponds to cells exhibiting a weaker intensity of the signal as it targets the fluorescent gene.
  - D12: cells + siRNA against Fibrillarin. Corresponds also to cells exhibiting a weaker intensity of the signal as it targets the Fibrillarin fused to the GFP fluorescent protein.
  - E12: cells + siRNA against Nucleophosmin. Corresponds to cells exhibiting the strongest unfolding phenotype of the controls.
  - F12: cells + siRNA against Nucleolin. These cells exhibit a weaker unfolding phenotype of the nucleoli closer to normal cells.
  - G12: cells + siRNA against Tif1A. These cells exhibit a capping phenotype where the Fibrillarin-GFP signal is accumulating in highly fluorescent spots within cells.

  - H12: cells + siRNA against RpL27. These cells exhibit various phenotypes different from the normal cells. Some cells present an unfolding of the nucleoli, others present a phenotype related to cap formation.

- The 5 remaining wells (A08, C03, C04, C07 and E02) correspond to examples directly extracted from the screening and were chosen because they exhibited strong and variable modifications of the nucleolus structure.

To ease the process, we have decided to start with the first 8 wells, which are well-defined as part of a class (normal/healthy or abnormal/diseased) by themselves thus they will facilitate the clustering in the end. Besides, in our experiments, wells A12 and B12 set up the healthy or normal class and wells C12, D12, E12, F12, G12 and H12 set up the diseased or abnormal class. In the Annex, you can find some sample of wells A12, B12, E12, F12, G12 and H12.

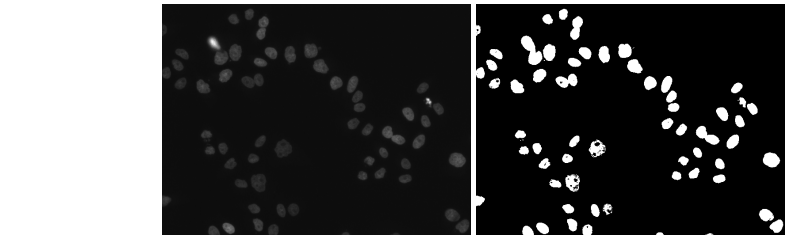
2.3 Segmentation of cells

All our experiments are carried out over nucleoli individual images. Therefore, in this section we explain how we obtain the nucleoli individual images from the images of illuminations w2 and w3. We use the third illumination w3 to segment the nuclei and obtain a labeled image of markers or masks allowing us to find the location of the nucleoli of each cell in w2. We use the sample site in Fig. 2.2.1 to explain the segmentation step by step.

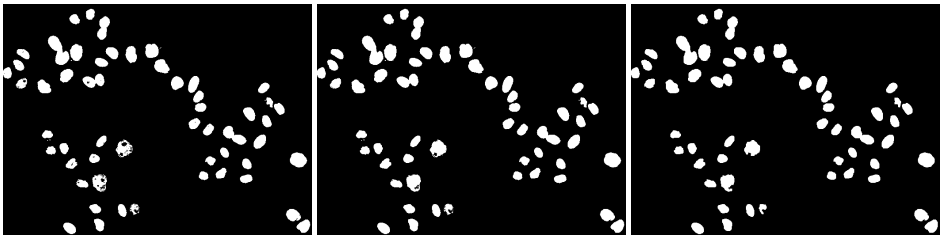
2.3.1 Segmentation of nuclei

Firstly, we binarize the image in Fig. 2.2.1(c) by a thresholding value $binary_{th}$ and obtain an early approximation of our mask image shown in Fig. 2.3.1(a). Then, in Fig. 2.3.1(b) we erase connected components touching image boundaries, which are incomplete nuclei, and in Fig. 2.3.1(c) we apply a hole filling to fill the small holes within the connected components in order to obtain a better labeling. After that, an image opening with a structuring element SE_{open} is applied in Fig. 2.3.1(d) to remove a number of small connected components that can be considered as noise. Next steps are presented in Fig. 2.3.1 from (e) to (g) and consist on the removal of non standard connected components, which are usually due to poor segmentation on account of an abrupt change of intensity in the original image or a clustering between two cells during binarization and hole filling processes. We consider non standard connected components, those connected components which are outside the limits of shape and size, which are established by thresholding. To detect and remove them we set the following thresholds:

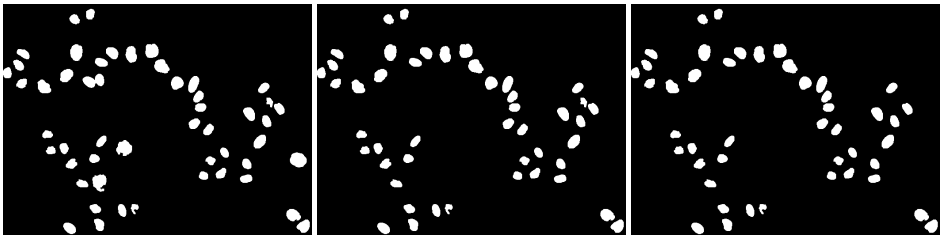
- A maximum ratio $ratioEllipse_{th}$ between the major and minor axis of a hypothetical ellipse approximating a connected component to detect elongated connected components belonging to clustered nuclei.
- A maximum area $areaBox_{th}$ for the bounding boxes surrounding the connected



(a) Nuclei image **(b)** Binarization by thresholding (109 labels)

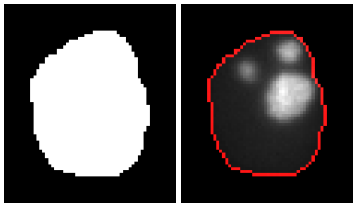


(c) Incomplete cells removal (102 labels) **(d)** Hole filling (94 labels) **(e)** Removal of small connected components (53 labels)



(f) Removal of elongated markers (50 labels) **(g)** Removal of big markers detected with bounding boxes (46 labels) **(h)** Removal of small markers (45 labels)

Figure 2.3.1: Segmentation of nuclei step by step. In each sub-figure it is indicated the number of labels or detected cells in that specific step of the segmentation.



(a) Mask image **(b)** Nucleolus image

Figure 2.3.2: Nucleolus location

components also to detect clustered nuclei. A bounding box is defined as the smallest rectangle containing a connected component.

- A minimum area $areaCell_{th}$ to detect too small connected components.

2.3.2 Obtaining nucleoli individual images

Finally, we label the remaining masks in Fig. 2.3.1(g) to disjoint each segmented nuclei as a seeding image and locate its own nucleolus. In the end, we create a rectangular background image in absolute black for each cell, to which the mask image and the nucleolus image are joined, and we save them apart. In Fig. 2.3.2(a) we have the first mask looking from left to right in Fig. 2.3.1(g) isolated from the rest and, at last, in Fig. 2.3.2(b) it is presented the nucleolus individual image obtained, where we have added a red line delimiting the nucleus perimeter. Every pixel outside the perimeter has zero value.

3 Mean shift algorithm

3.1 General algorithm

Mean shift [12] is a non-parametric feature-space technique, commonly known as mode detection or mode seeking technique, proposed for the feature space multimodal analysis. As mean shift is a density estimation-based technique, the feature space can be regarded as the empirical probability density function (PDF) of the represented parameter. Thereby, dense regions in the feature space correspond to local maxima and minima of the PDF, that is, to the modes of the unknown density $f(r)$. These modes are located among the zeros of the density gradient:$\nabla f(r) = 0$. Meanwhile, mean shift procedure is an elegant way to locate these zeros without estimating the density. Once the location of a mode is determined, clustering depends on the local structure of the feature space.

Given a set $S = \left\{ r_i \in \mathbb{R}^d \right\}_{i=1..n}$ of n data points in a d-dimensional space $\mathbb{R}^d$, mean shift procedure consists on an iterative method that starts with an initial point or estimate r, which is a d-vector. As a Kernel Density Estimation (KDE) is a fundamental data smoothing tool based on finite data samples, mean shift uses a kernel function k to determine the weights of nearby points to r. The kernel density estimation technique works with a symmetric positive definite $d \times d$ bandwidth matrix H, simplified in practice by $H = h^2 I_d$, where h is a positive scalar and I_d is the $d \times d$ identity matrix. Matrix H controls the orientation of the smoothing induced. Accordingly, let's define mean shift equation:

$$m_{h,g}(r) = \frac{\sum_{i=1}^{n} r_i g\left(\left\| \frac{r - r_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{r - r_i}{h} \right\|^2 \right)} - r \tag{3.1.1}$$

where $g(t)$ is the derivative of the kernel k:

$$g(t) = -k'(t) \tag{3.1.2}$$

The iterative process consists on the translation of the kernel window through the mean shift vector $m(r)$ by setting $r_{new} = r_{old} + m(r_{old})$ and repeating the mean shift procedure again with the new estimate, until $m(r)$ converges at a nearby point where the PDF has zero gradient, that is to say, where $r_{new} = r_{old}$. The described iterative process can be repeated for the whole data set of points or just for a representative sample in the feature space. Fig. 3.1.1 shows an example of mean shift application.

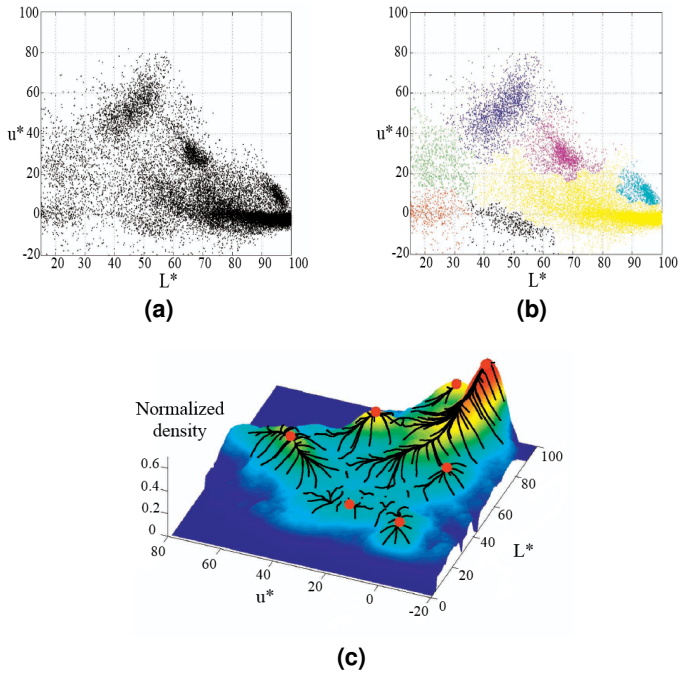


Figure 3.1.1: Example of a 2D feature space analysis. (a)Two dimensional data set of 110.400 points representing the first two components of the L*u*v space from an example image. (b)Decomposition obtained by running 159 mean shift procedures with different initializations. (c)Trajectories of the mean shift procedure drawn over the Epanechnikov density. The peaks retained for the final classification are marked with red dots. (Figure from [[12], Fig. 2, p. 7])

3.2 Adaptation to our case

As we apply mean shift algorithm directly to gray scale images, we have adapted the general algorithm to our particular case. Hence, we work with a 2-dimensional space $\mathbb{R}^2$ where we define P as the set of Cartesian coordinates of an image I of size $H \times W$, where H is the height of the image (number of rows) and W is the width of the image (number of columns) :

$$P = \{r_i = (x_i, y_i) \mid x_i \in [1, W] \,\&\, y_i \in [1, H]\}_{i=1..H\times W} \tag{3.2.1}$$

3.2.1 Adaptation from points to image intensities

Thus, assuming each pixel in the image I with an intensity level greater than zero as an initial estimate or seed $r = (x, y)$, we apply mean shift iterative procedure on the following set S of data:

$$S = \{r_i \in P \mid I(r_i) > 0\}_{i=1..n} \tag{3.2.2}$$

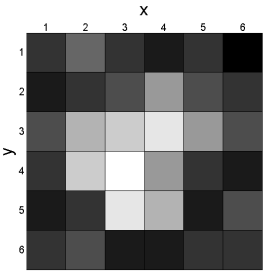
Fig. 3.2.1(a) and Fig. 3.2.1(b) show an example of a simple gray level image of 6x6 pixels with eleven integer intensity levels [0-10]. As pixels in the image are equidistant among them(see Fig. 3.2.1(c)), they must be distinguished for their intensity level. In particular, Fig. 3.2.1(d) represents the conversion from the intensity level of each pixel to a proportional number of density points, which represent the density level of the image intensity. These density points are drawn with crosses in the figures. Fig. 3.2.1(e) shows one particular section of the axis *y*, where we can relate the intensity level of the six pixels from $y = 5$, represented in the axis of ordinates by green circles, to the appropriated number of density points, represented by red crosses. Finally, applying these ideas to the Equation 3.1.1, we obtain this expression:

$$m_{h,g}(r) = \frac{\sum_{p=1}^{n} r_p I(r_p) g\left(\left\|\frac{r-r_p}{h}\right\|^2\right)}{\sum_{p=1}^{n} I(r_p) g\left(\left\|\frac{r-r_p}{h}\right\|^2\right)} - r \tag{3.2.3}$$

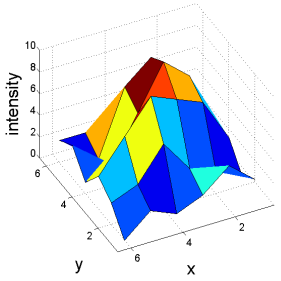
3.2.2 Kernel function

Following, we define a kernel function to convert our discrete sampled image into a continuous one. To simplify calculations, we have decided to use a triangular distribution as kernel function, also known as hat function (see Fig. 3.2.2):

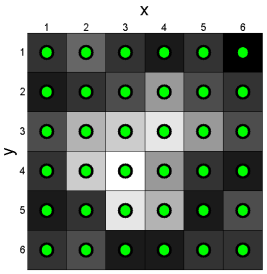
$$k(t) = triangle(t) = \bigwedge(\mathrm{t}) = \begin{cases} 1 - |t| & |t| < 1 \\ 0 & otherwise \end{cases} \tag{3.2.4}$$



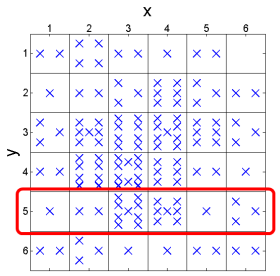
(a) Sample image



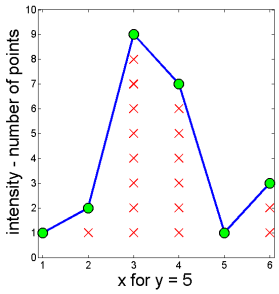
(b) Sample image in 3D



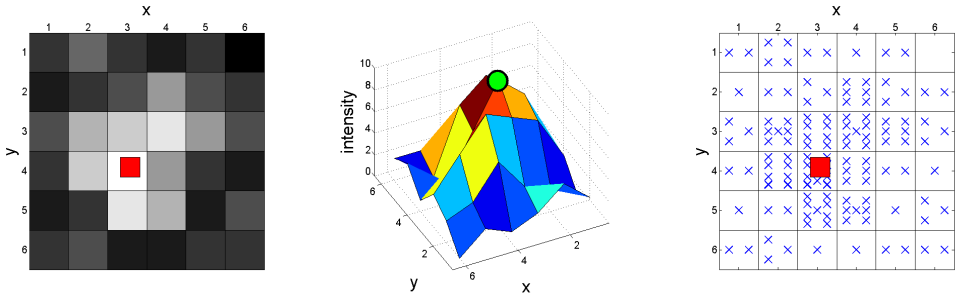
(c) Green circles represent equidistant pixels in the image



(d) Proportional conversion from intensity levels to density points



(e) Density points represented for a specific section of y axis



(f) Mean Shift results

Figure 3.2.1: Concept of mean shift density from image intensity levels

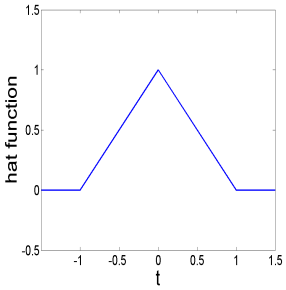


Figure 3.2.2: Triangular function or hat function

That way Equation 3.2.3 is simplified due to:

$$g(t) = -k'(t) = \begin{cases} \frac{t}{|t|} & |t| < 1 \\ 0 & otherwise \end{cases}$$

Due to this simplification, our kernel function only determines which neighboring pixels are involved on the mean shift calculation of a specific point. Accordingly, kernel function fixes the mean shift window size, which is defined by a square window of size $bw_{MS} \times bw_{MS}$ centered on the point r. In conclusion, letting $N(r)$ be a set of pixels falling within the mean shift window and defined by a infinity norm, the final equation for our particular application looks like this:

$$m(r) = r^{MS} - r \tag{3.2.5}$$

$$where\, r^{MS} = \frac{\sum_{r_p \in N(r)} r_p I(r_p)}{\sum_{r_p \in N(r)} I(r_p)} \tag{3.2.6}$$

$$and\, N(r) = \{u \in S \mid ||r - u||_\infty \leq bw_{MS}/2\}\ where\, bw_{MS}/2 = h \tag{3.2.7}$$

As we work in a continuous feature space, mean shift vector $m(r)$ can be any irrational number with infinite decimals. Therefore, to ensure that it converges at some point, a tolerance value $mTol$ is set. When mean shift vector reaches $mTol$ or a lower value, that is, $||m(r)||_2 \leq mTol$, it is considered that $m(r)$ has converged. Then, mean shift iterative process ends and the new estimate or mean shift point $\tilde{r}^{MS}$ is set.

3.2.3 Clustering

Once we have applied mean shift procedure to each starting point or pixel r, we obtain an individual result from each one, namely, a mean shift point $\tilde{r}^{MS}$ located in the continuous feature space where the mean shift vector has converged.

Next step consists on clustering as it is shown in Fig. 3.1.1(b) with different colors. Clustering is based on the idea of gathering together all the points which are close enough within the appropriated space and following the subsequent measuring model of distance. In our case, we look at the 2-dimensional space of a gray scale image and establish a threshold $cTol$ to determine if two points are close enough to belong to the same cluster. Thus we define the clustering condition for two sample points $\tilde{r}_1^{MS}$and $\tilde{r}_2^{MS}$ as:

$$\|\tilde{r}_1^{MS} - \tilde{r}_2^{MS}\|_2 \leq cTol \tag{3.2.8}$$

3.2.4 Peak seeking

Finally, a mode seeking or peak seeking technique is necessary to find the main peak of each cluster. Our mode seeking process consists on going through all the mean shift points r_c^{MS} assigned to a particular cluster and searching for the higher mean intensity calculated around each r_c^{MS} within the clustering window, which is a square window of size $bw_{cluster} \times bw_{cluster}$ centered on r_c^{MS}. The mean intensity around a particular point r_c^{MS} is calculated from the intensities of the pixels in $M(r_c^{MS})$, which is the set of pixels falling within the clustering window and defined by the infinity norm.

$$mean_{calc}(r) = \frac{\sum_{r_p \in M(r)} I(r_p)}{|M(r)|} \tag{3.2.9}$$

$$where\, M(r) = \{v \in S \mid ||r - v||_\infty \leq bw_{cluster}/2\} \tag{3.2.10}$$

and $|A|$ is the cardinality of A

This way, we can guarantee that we find modes located at a summit of a spot in the original image instead of being an isolated pixel with high intensity.

Results from applying the mean shift procedure executed with Equation 3.2.5, the clustering procedure and the mode seeking just exposed to the sample image in Fig. 3.2.1(a) are shown in Fig. 3.2.1(f). In this particular case, there is only one cluster with its own suitable mode.

3.3 Pseudo-code

3.3.1 Mean shift program

In Algorithm 3.1., we perform Mean shift for the whole data set of initial points $r = (x, y)$ going over the S array defined by Equation 3.2.2 position by position. In each position we call iterationMS_procedure program iteratively, updating the mean shift vector m in each iteration until it converges, that is to say, when the euclidean distance of the mean shift vector $||m||_2$ is smaller or equal to the mean shift tolerance $mTol$. When mean shift converges, we save the position of the new point $\tilde{r}^{MS}$ in the $points_{MS}$ array. Once we have compiled a new estimation for the whole set of pixels, we proceed to carry out the clustering to gather the nearby points in a single cluster. Finally, we obtain a few number of peaks by a mode seeking technique and keep them on $modes$ array.

Algorithm 3.1 Mean shift program

Input
- → I: Image matrix of size HxW
- → bw_{MS}: Mean shift window width
- → $bw_{cluster}$: Clustering window width
- → mTol: Mean shift tolerance for convergence
- → cTol: Maximum distance between two points for clustering

Output
- → nModes: Number of clustering modes
- → modes: Array of size [nModes,2] gathering the coordinates of the clustering modes
- → $index_{cluster}$: Array of size nSeeds containing the indexes of the clustering modes for which each mean shift point in $points_{MS}$ belongs to

```
1  Let nSeeds:= number of initial points
2  Let S:= array of size [nSeeds,2] containing the coordinates of the
       initial points r
3  Let points_MS:= empty array of size [nSeeds,2] that will keep the
       coordinates of the points after applying mean shift procedure r_c^MS
4
5  for n=1..nSeeds do
6      m:=[mTol mTol]^T
7      r:=S(n,:)
8      while ‖ m ‖_2 > mTol do
9          r^MS:=iterationMS_procedure(r,I,bw_MS)
10         m:=r^MS − r
11         r:=r^MS
12     end while
13     points_MS(n,:):=r^MS
14 end for
15 [nModes,modes,index_cluster]:=clustering_procedure(I,S,points_MS,cTol,bw
       _cluster)
```

Algorithm 3.2 iterationMS_procedure program

Input
- → r: 2 dimensional array containing the point to update with the new point
- → I: Image matrix of size HxW
- → bw_{MS}: Mean shift window width

Output
- → r^{MS}: 2 dimensional array wich will gather the new point

1 **Let** xMin:= max(1, ceil(r(1)−bw_{MS}/2))
2 **Let** xMax:= min(W, floor(r(1)+bw_{MS}/2))
3 **Let** yMin:= max(1, ceil(r(2)−bw_{MS}/2))
4 **Let** yMax:= min(H, floor(r(2)+bw_{MS}/2))
5
6 $r^{MS} := \frac{\sum_{x=xMin}^{xMax}\sum_{y=yMin}^{yMax}[x,y]^T I(x,y)}{\sum_{x=xMin}^{xMax}\sum_{y=xMin}^{xMax} I(x,y)}$

3.3.2 Subprograms

In this section, we present the pseudo-code of the subprograms called from the main program: iterationMS_procedure, clustering_procedure and mean_procedure.

3.3.2.1 IterationMS_procedure program

The mean shift iterative process is given in Algorithm 3.2., where iterationsMS_procedure receives an arbitrary point $r = (x, y)$ and proceeds to apply the Equation 3.2.6 returning r^{MS}, where $N(r)$ is the set of pixels defined by Equation 3.2.7 falling within the mean shift window, which is a square window of size $bw_{MS} \times bw_{MS}$ centered on the given point r. In our implementation we define $N(r)$ belonging to P set instead of S (see Equations 3.2.1 and 3.2.2), because adding some zero intensity values within the summation of r^{MS} operation does not change the result but, in exchange, it simplifies the code.

3.3.2.2 Mean_procedure program

The intensity mean operation around a given point is described in Algorithm 3.3., where mean_procedure receives an arbitrary point $r = (x, y)$ and proceeds to apply Equation 3.2.9 returning $mean_{calc}$, where $M(r)$ is the set of pixels defined by Equation 3.2.10 falling within the clustering window, which is a square window of size $bw_{cluster} \times bw_{cluster}$ centered on the given point r.

Algorithm 3.3 mean_procedure program

Input
- → r: 2 dimensional array containing the point to calculate the mean intensity
- → I: Image matrix of size HxW
- → S: Array of size [nSeeds,2] containing the coordinates of the initial points r
- → $bw_{cluster}$: Clustering window width

Output
- → $mean_{calc}$: Mean intensity around the given point r

1 **Let** xMin:= max(1, ceil(r(1)−bw_{MS}/2))
2 **Let** xMax:= min(W, floor(r(1)+bw_{MS}/2))
3 **Let** yMin:= max(1, ceil(r(2)−bw_{MS}/2))
4 **Let** yMax:= min(H, floor(r(2)+bw_{MS}/2))
5 **Let** M(r):= { v ∈ S | xMin ≤ v(1) ≤ xMax and yMin ≤ v(2) ≤ yMax }
6
7 $mean_{calc} := \frac{\sum_{r_p \in M(r)t} I(r_p)}{|M(r)|}$

3.3.2.3 Clustering_procedure program

Clustering_procedure is given in Algorithm 3.4. and it consists on gathering together into a cluster all the mean shift points in $points_{MS}$ array that satisfy the clustering condition defined in Equation 3.2.8, merging neighbor points step by step. Finally, the mode seeking technique allows us to find a representative peak for each cluster applying $mean\_procedure$ to all the mean shift points belonging to the same cluster and keeping the point with the higher result in $modes$ array.

```
/* A \ B is the relative complement of array A in array B, also termed the
   set-theoretic difference of B and A */
/* [A
   B] is the concatenation of the arrays A and B */
```

3.4 Post-processing

As it has already been told, mean shift procedure locates the gradient zeros of the density function. Consequently, although we are only interested in local maxima to locate the white spots on the images of the cell nucleolus, mean shift also finds local minima. Accordingly, we need to apply some post-processing after mean shift to remove these minima and definitely find the useful modes.

To better understand what we mean, we have illustrated all previous concepts in an example: we have applied a mean shift procedure on a normalized nucleolus image shown in Fig. 3.4.1(a) with a mean shift window width bw_{MS} of 10, a mean shift tolerance $mTol$ of 0.10, a maximum distance for clustering $cTol$ of 2 and a clustering window width $bw_{cluster}$ of 3. Fig. 3.4.1(b) and Fig. 3.4.1(d) show the modes after clustering and

Algorithm 3.4 clustering_procedure program

Input

- → I: Image matrix of size HxW
- → S: Array of size [nSeeds,2] containing the coordinates of the initial points r
- → $points_{MS}$: Array of size [nSeeds,2] containing the coordinates of the mean shift points
- → cTol: Maximum distance between two points for clustering
- → $bw_{cluster}$: Clustering window width

Output

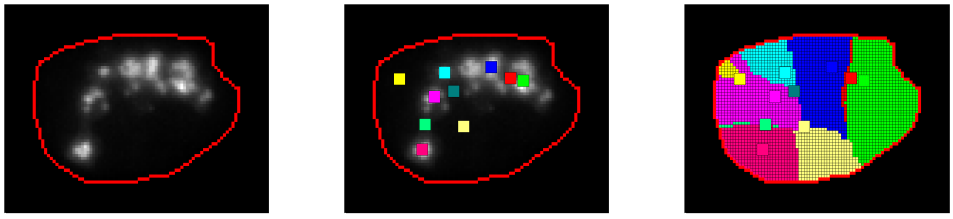
- → nModes: Number of clustering modes
- → modes: Array of size [nModes,2] gathering the coordinates of the clustering modes
- → $index_{cluster}$: Array of size nSeeds containing the indexes of the clustering modes for which each mean shift point in $points_{MS}$ belongs to

```
1  Let modes:= []
2  Let nModes:= 0
3  Let set_notVisited:= [1:1:nSeeds], array containing a set of indexes of
       the mean shift points that have not been assigned to a cluster yet
4  Let means_calc:= array of size nSeeds that will keep the mean
       intensities around the mean shift points in points_MS
5  Let set_neighbors:= array containing a set of indexes of the points from
       which we have to search neighbors, it is redefined at every mean
       shift point
6  Let merge_going:= array containing a set of indexes of the mean shift
       points belonging to the same cluster, it is redefined at every
       cluster
7
8  means_calc(i):= mean_procedure(points_MS(i,:),I,S,bw_cluster), ∀ i=1..nSeeds
9  while length(set_notVisited)>0 do
10     set:=set_notVisited(1)
11     set_neighbors:=set
12     /* Merging neighbor points step by step */
13     merge_going:=[]
14     merge_prev:=[]
15     while length(set_neighbors)>0 do
16         for all set ∈ set_neighbors do
17             merge_going:=merge_going ∪ {i∈set_notVisited |
18                         ‖points_MS(i,:)−points_MS(set,:)‖_2 ≤cTol}
19         end for
20         set_neighbors:=merge_going \ merge_prev
21         merge_prev:=merge_going
22     end while
23     /* Find the clustering mode for nModes*/
24     index_mode_n:=argmax_{i∈merge_going} means_calc(i)
25     nModes:=nModes+1
26     modes:=[modes
27             points_MS(index_mode_n,:)]
28     index_cluster(i):=nModes, ∀ i ∈ merge_going
29     /* Index of mean shift points that have not been assigned to a
           cluster */
30     set_notVisited:=set_notVisited \ merge_going
31 end while
```

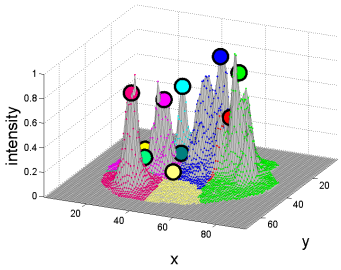
Fig. 3.4.1(c) shows the shapes and sizes of every cluster. In this particular nucleolus image, there are five well-defined peaks: from left to right they are the dark pink, the light pink, the sky-blue, the navy blue and the green clusters. Thus, there are five useless clusters which must be removed. To remove them, we need to apply different tools:

1. Thresholding: to erase lowest useless modes we just need to determine a threshold and remove the modes below it. Results from applying this tool to the sample image are presented in Fig. 3.4.1(e), where a cutting section dividing removed modes and remaining modes is displayed, and finally in Fig. 3.4.1(f) and Fig. 3.4.1(g), where only remaining modes are exposed. If we apply a too aggressive thresholding, a lot of small but very significant maxima can be removed. Hence, it is better to apply a cautious thresholding although useless modes are not entirely removed.

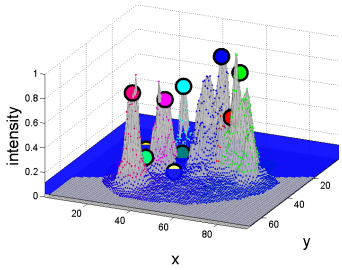
2. Lines: there are some cases where useless modes are above the previous thresholding. For those cases, it is required to search for another common characteristic feature. Since we have observed that useless higher modes use to belong to lengthened and narrow clusters, as shown in Fig. 3.4.1(h), we can consider them as drawn lines oriented towards a particular direction. Thus, to detect them, we utilize one pixel width lines drawn within the cluster boundaries as you can see on the schema in Fig. 3.4.3, passing through the mode assigned to that particular cluster and oriented towards different directions. In case that one of these lines cuts across a specific cluster sharing only the central pixels with that cluster, we can say we have found a useless cluster and we proceed to remove it. If the cluster has one pixel width, it is possible to find empty lines, that is, lines with no pixels shared. Lines are also helpful in some cases where useless clusters intersect with others, in which we find dashed lines. In our implementation, we use four different lines: vertical, horizontal and the two diagonals of 45 degrees relative to the axes of the nucleolus image. Fig. 3.4.2 shows some examples of lines. For instance, you can see a one pixels line cutting across the cluster in (1), a dashed line over the green cluster in (2) and an empty line on the cluster in the middle at (3). Hence, the aforementioned clusters are considered as useless and erased in post-processing.



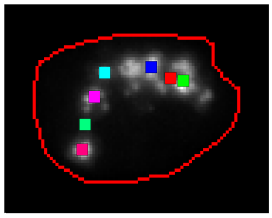
(a) Nucleolus of a sample cell **(b)** Mean shift modes after clustering **(c)** Mean shift clusters before thresholding



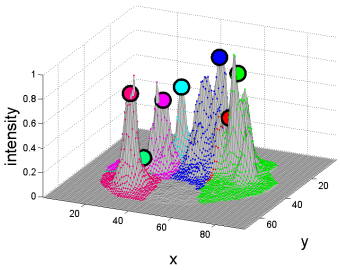
(d) Mean shift modes after clustering in 3D



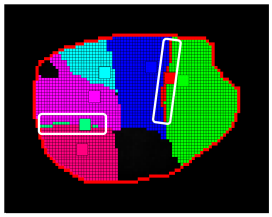
(e) Graphic illustration of thresholding



(f) Mean shift modes after thresholding

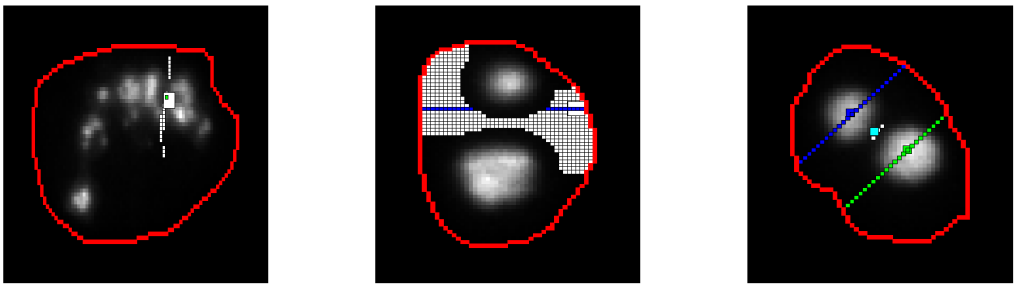


(g) Mean shift modes after thresholding in 3D



(h) Mean shift clusters after thresholding

Figure 3.4.1: Useful modes selection example



(1) (2) (3)

Figure 3.4.2: Lines examples to erase useless modes

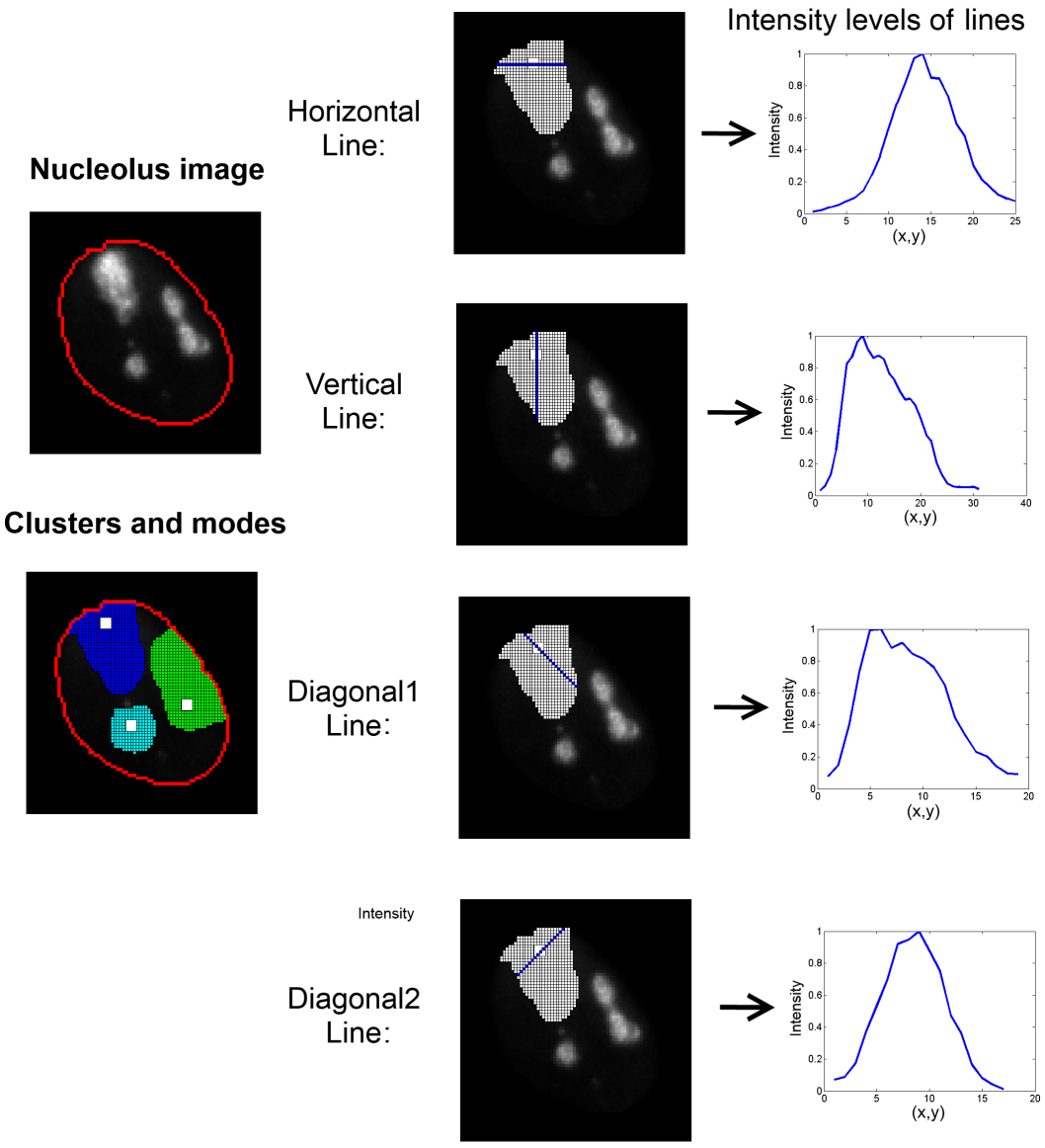


Figure 3.4.3: Lines obtaiting and display

4 Nucleolus phenotype discrimination

In this section, we work with the isolated nucleoli images. A nucleolus image is defined as a rectangular image with an absolute black background and a centered nucleolus, as seen in Fig. 4.1.2, where we have added a red line delimiting the nucleus perimeter. Every pixel outside the perimeter has zero value.

4.1 Image pre-processing

Throughout the development of the various experiments carried out in this thesis, we work with two types of nucleolus images in terms of intensity: globally normalized images and individually mapped images:

- **Global normalization (norm):** This pre-processing is appropriate for those features directly extracted from the original images without being processed prior, and related to the intensity levels. As the name suggests, this is a normalization by a global adjust. We use the absolute maximum intensity of the whole set of nucleoli images to normalize the intensity levels thus we guarantee that intensity values are in a range between zero and one.
- **Individual mapping (gamma):** This pre-processing is suitable for those images that will be processed in some way before feature extraction. Individual mapping is particularly effective in samples with low intensity levels or with a narrow intensity range. Our individual mapping consists in widening the image intensity range in order to take advantage of the whole range allowed by the character of the image so that we can get a better visualization of details. Additionally in our particular case, brighter pixels are our center of attention. Hence, we apply a gamma correction to emphasize higher intensity levels (brighter pixels) to the detriment of lower ones (darker pixels).

  Mapping is a range transform operation where input and output range can also be established. This means that you can determine to use the whole range of the image or you can establish a minimum and maximum intensity levels instead. For the input image, boundary intensities will be “low” and “high” values in Fig. 4.1.1. Intensities out of this range will disappear and their pixels will be reassigned to the boundary intensities. “Bottom” and “top” values will be the boundary intensities assigned to the output image. In our case, for every nucleolus individual image, we have applied a mapping with an input range from the minimum to the maximum of the particular image, and an output range from zero to one. That way, we can assure intensity levels are used as efficiently as possible.

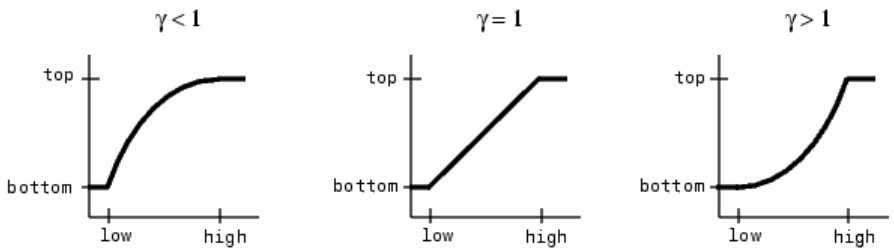


Figure 4.1.1: Gamma correction

Gamma correction is a mapping defined by the following power-law expression:

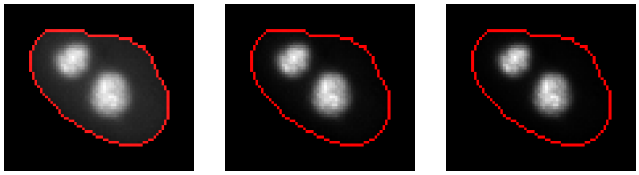
$$Level_{output} = (Level_{input})^{gamma} \tag{4.1.1}$$

As can be seen in Fig. 4.1.1, gamma value specifies the shape of the curve describing the relationship between the intensity values of the input and output images. For gamma values greater than 1, the mapping is weighted towards lower values, that is, brighter intensity levels will be highlighted while some dark levels will be lost. On the contrary, for gamma values under 1, the mapping is weighted towards higher output values and, in consequence, darker intensity levels will be highlighted to the detriment of brighter ones. At last, if gamma is equal to 1, a linear mapping will be applied. As we want to highlight brighter values to distinguish white spots from background, we use gamma values greater than 1.

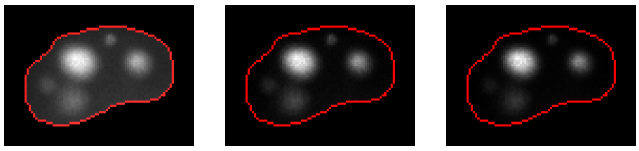
In particular, we have chosen a sample of 6 representative nucleolus images to test mean shift configurations. Fig. 4.1.2 shows three different images per cell. Fig. 4.1.2 (1) are the nucleolus images ordinarily normalized by their particular maximums. In turn, Fig. 4.1.2 (2) and Fig. 4.1.2 (3) are the nucleolus images after applying gamma correction with a gamma value of 1.25 and 1.50 respectively.

4.2 Which features?

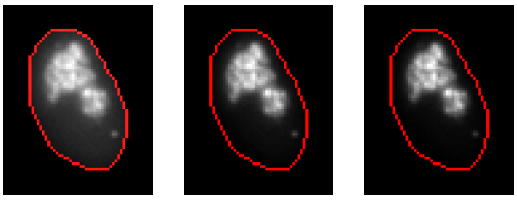
In this section we introduce a wide variety of features that have been tested in the nucleoli images. In Section 5, results are presented. We work with two different types of features: features directly extracted from the nucleolus images and features extracted from a previous characterization, in our case from mean shift or edge detection. For the characterization of the nucleolus and the feature extraction, we use the two types of images previously introduced: norm and gamma.



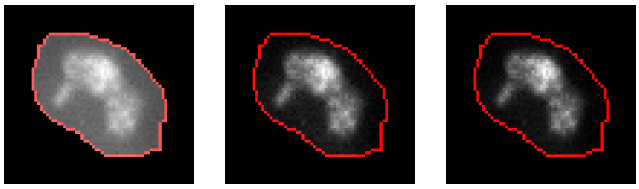
(a) Cell 1



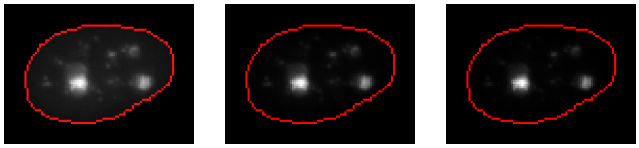
(b) Cell 2



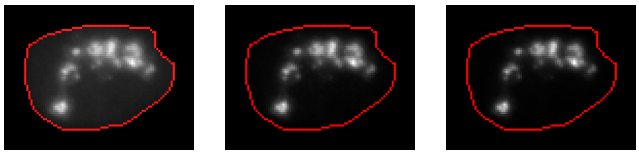
(c) Cell 3



(d) Cell 4



(e) Cell 5



(1) (2) (3)

(f) Cell 6

Figure 4.1.2: Sample of cells. (1) Raw nucleolus images. (2) Nucleolus images after gamma correction with a gamma value of 1.25 (3) Nucleolus images after gamma correction with a gamma value of 1.50

4.2.1 Intensity

In addition to the phenotypes described in Section 2.1, a third visual discrimination is also contemplated: image intensity levels can be affected differently at each experiment, for instance, at each set of samples corresponding to one silencer. Particularly, in section 2.2 we present the database used in our experiments, which is composed of 8 wells corresponding to 2 different classes: healthy or normal class (wells A12 and B12) and diseased or abnormal class (wells C12, D12, E12, F12, G12 and H12). In wells C12 and D12, the silencer targets directly the GFP fluorescent protein and the Fibrillarin fused with the GFP respectively, causing a loss of intensity. Hence, maximum intensity $intensity_{max}$ extracted from norm image is tested.

4.2.2 Mean shift modes

As it has been told in Section 1, earlier studies have shown that there is a strong correlation between the 3-D structure of the nucleolus and the potential diseases affecting the cell. Accordingly, we have decided to apply the mean shift algorithm to gamma image to detect the most representative maxima or modes and characterize the nucleolus based on these modes and their clusters. An overview of the most significant features extracted from mean shift is expounded below:

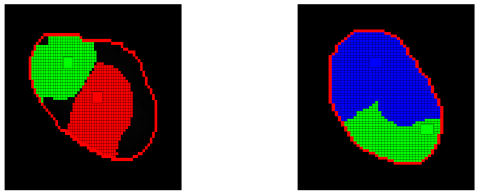
1. In a specific mean shift configuration, which is defined by the parameters bw_{MS}, mTol, cTol and $\text{bw}_{\text{cluster}}$, we extract the following features for each nucleolus:

   a) $numModes$: Number of modes

   b) When a feature is extracted directly from a mode, we have a set of values that corresponds to the set of resulting modes and we need to choose a value among the set. These features are:

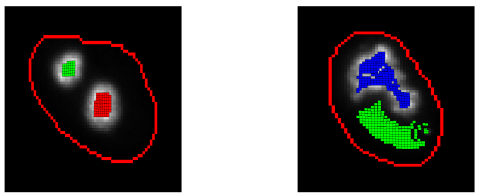
      i. $intensity_{modes}$: Intensity of modes extracted from norm image. We save the maximum and the mean of the set of intensities.

      ii. Shape features: As mean shift assigns all the pixels in the nucleus to a cluster, we are not able to extract shape information of the white spots within the image from clusters. In section 3.2.3, we have already explained that we obtain an individual result $\tilde{r}^{MS}$ from applying mean shift procedure to each starting pixels r and then we apply a clustering towards $\tilde{r}^{MS}$ points because they are compacted in some areas among the zeros of the density gradient.
      On the contrary, if we observe the shape designed by the set of points $\tilde{r}^{MS}$ belonging to a cluster, we notice some differences between a typical normal and abnormal behavior. Thereby, in Fig. 4.2.1 you can see that normal cells are characterized by compact, small and circular shapes while abnormal cells presenting unfolding phenotypes are characterized by extended shapes. We set up a closed region of pixels from the points $\tilde{r}^{MS}$ belonging to a cluster and test some shape features on these regions:



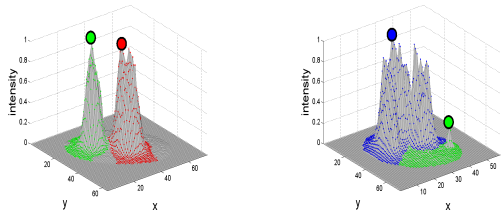
Typical normal cell Typical abnormal cell

(a) Pixels belonging the cluster



Typical normal cell Typical abnormal cell

(b) Points $\tilde{r}^{MS}$ belonging the cluster



Typical normal cell Typical abnormal cell

(c) Pixels in 3D

Figure 4.2.1: Shape in mean shift clusters

A. $area_{modes}$: number of pixels within a region

B. ecc_{modes}: Eccentricity: ratio of the distance between the foci of a hypothetical ellipse composed by the region and its major axis length. The value is between 0 and 1, being a circle an ellipse whose eccentricity is 0 and a line segment an ellipse whose eccentricity is 1.

For each shape feature we obtain a set of values corresponding to the set of clusters. Next step consists on defining the choice criteria to convert a set of features into a value. In $area_{modes}$ we look for the ratio between the minimum area and the maximum area found in the nucleolus while for ecc_{modes} we save the maximum, minimum and mean of the set of values.

2. Features to compare between different mean shift configurations:

   $numModes$: In cells presenting unfolding phenotype, we find different peaks of intensity in the same white spot. Depending on the mean shift configuration the algorithm can detect these different peaks dividing the white spot in different clusters or just detect one peak considering the white spot as a single cluster.

4.2.3 Lines

We use Lines to characterize the different spots shaping the nucleolus in mean shift technique. We define a Line as a one line pixel width line drawn within the cluster boundaries, passing through the mode assigned to that particular cluster and oriented towards different directions. In our implementation, we use four different Lines: horizontal line, vertical line and two diagonal lines of 45 degree relative to the axes of the nucleolus image. Lines allow us to look for some features related to the 3-D patterns of the different clusters. Thus we have four lines per mode and a number of modes which depends on the mean shift results. To decide which line to use for the feature extraction, we use different criteria:

1. Length: we choose the largest line of the cell (line1)
2. Number of oscillations: we choose the line with more oscillations of the cell (line2)
3. Intensity value: we choose the mode with higher intensity value. Then, to select one of the four directions of that mode, we still need to use the criteria:
   a) Length (line31)
   b) Number of oscillations (line32)

Following, an overview of the features to look at on the chosen line:

- For each line, we define 3 interesting locations and for each location we find the features $position$, $normVal$ and $gammaVal$. The locations are:
  - $height$:

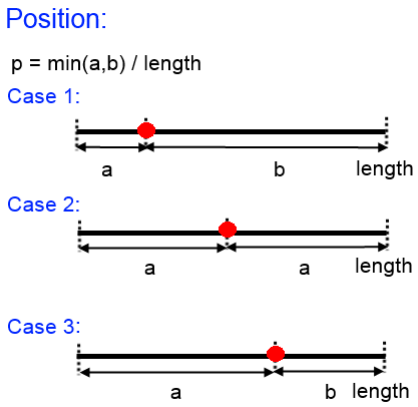


Figure 4.2.2: $position$ calculation

  * $maxH$: Absolute maximum of the line
  * $modeH$: Mode found with mean shift
- $absMin$: Absolute minimum of the line. Though, to do the choice of the $absMin$ in a line, a minimum located between two maxima is a priority compared to the minima in the ends of the lines (see Fig. 5.1.17). Thus, we will only choose a minima in the end when there is only one oscillation and, consequently, just the two minima at the ends.

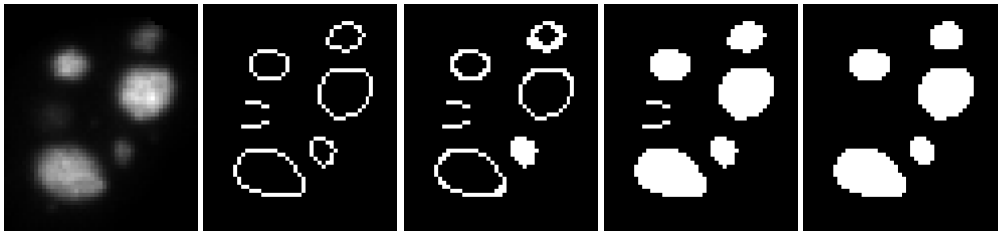
- $position$: Normalized index which indicates the position of a location in the length. We present feature $position$ in Fig. 4.2.2 where red points are the placement of the location:

$$position = min(a, b)/length$$

  As we want a rotation invariant system, we need a normalization of the position where we get the same result on cases 1 and 3 in Fig. 4.2.2 when *a* length from case 1 is equal to *b* length from case 3. In this normalization, we obtain values in a range [0-0.66] where values close to 0 belong to locations placed near to one end of line and values within the range [0.5-0.66] belong to centered locations (case 2 in Fig. 4.2.2).

  This measure should help us to detect capping phenotype due to this phenotype consists on the compaction of material at the periphery of the nucleolus and this compaction should lead to lines with heights placed close to the ends of line. Also this feature should help to detect unfolding phenotype, because in images affected by this phenotype, mean shift finds centered locations $absMin$ while in a typical normal cell, we use to find minima at the ends of the line.

- $normVal$: Intensity value of that particular location on norm image.
- $gammaVal$: Intensity value of that particular location on gamma image.
- $numOsc$: Number of oscillations, which is equivalent to the number of maxima of



(a) Sample nucleolus image **(b)** Laplacian of Gaussian filter **(c)** Closings **(d)** Hole filling **(e)** Openings

Figure 4.2.3: Edge detection step by step

the line.

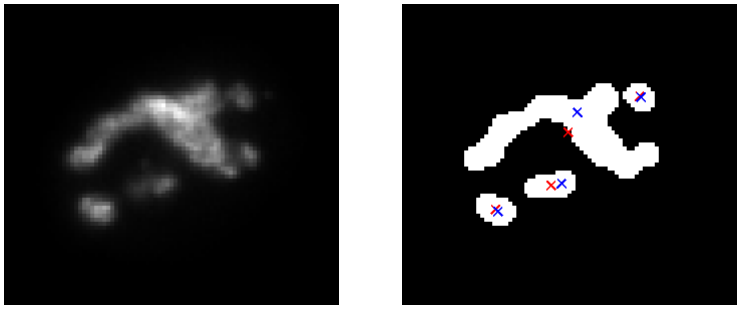
- $lengthLine$: Length of the line.
- $width$: Width calculated over the height locations and for the 50% and the 80% of the intensity value of that height: $width50max$, $width50mode$, $width80max$ and $width80mode$.
- $ratioLine$: this ratio is a measure for the oscillations of the lines and is only calculated for those lines that have $numOsc > 1$. Fig. 5.1.18 illustrates the necessary metrics to calculate this ratio. We define it as:

$$ratioLine = \frac{min(d1, d2)}{\left(\frac{h1+h2}{2}\right)} \tag{4.2.1}$$

4.2.4 Edge detection regions

As mean shift clusters are not following the shapes of the white spots in the nucleolus, there are some measures of shape and size that we are not able to do. We propose a tool that complements mean shift providing shape and size measures from the white spots in the nucleolus. To find the spots in the nucleolus, our edge detection technique follows next steps, exposed in Fig. 4.2.3:

- Laplacian of Gaussian filter to detect contours on gamma image.
- Several image tools:
  - Image closings with a structuring element SE_{close} in different directions to close some contours that still opened
  - Hole filling to fill the small holes within connected components in order to obtain a better labeling
  - Image openings with a structuring element SE_{open} in different directions to remove some small remaining lines.



(a) Nucleolus image (b) Modes and physical centers

Figure 4.2.4: Mode seeking technique applied in edge detection regions

- Finally, a labeling with a connection $connect$ is done to obtain an individual region for each white spot.
- Last, we erase some regions smaller than a given threshold $area_{th}$ that have not been removed nor assigned to any big connected component during the process .

This tool does not take into account the variations of intensity as mean shift does and for that reason it finds small regions with low intensity and big regions with high intensity equally, and detects some modes impossible to obtain with mean shift. Besides, with this tool we can have a more accurate idea of the white spots shapes and sizes.

Furthermore, to introduce the intensity factor to the proposal, we apply the mode seeking technique explained in section 3.2.4 to our regions, exchanging the points belonging to a cluster for the pixels belonging to a region in this case. Results are shown in Fig. 4.2.4, where physical centers of the regions are marked with red crosses while resulting modes from mode seeking are marked with blue crosses.

An overview of the most significant features extracted from edge detection is expounded below:

1. $numRegions$: Number of regions
2. For features extracted directly from the mode or region, we have a set of features that corresponds to the set of resulting modes or regions from a single nucleolus and we need to choose a value among the set of features. These features are:
   a) $intensity_{regions}$: Intensity of modes extracted from norm image. We save the maximum, minimum and the mean.
   b) $distance_{regions}$: Distance between modes. We save the smallest and the largest.
   c) $distanceCentroid_{regions}$: Distance between the mode and the physical center of a region. Since normal nucleoli have compacted and symmetric spots while abnormal nucleoli can be affected by unfolding phenotype causing the unraveling of the nucleolus, the distance between a mode and a physical

(a) Typical normal cell (b) Bounding boxes in typical normal cells

(c) Typical abnormal cell (d) Bounding boxes in ypical abnormal cells

Figure 4.2.5: Unfolding phenotypes with edge detection tool

center in a normal cell should be shorter than in an abnormal cell with unfolding phenotype where the modes can be located close to the border of the region, as is is shown inFig. 4.2.4, where we also see that the physical center of the big region is located out of the region. We keep the largest, the second largest and the smallest distance.

d) Shape features for regions:

   i. $MA_{regions}$: Major axis of the ellipse approximating the region.

   ii. $ecc_{regions}$: Eccentricity of the ellipse

   iii. $R_{regions}$: Ratio between the area of the region and the area of the bounding box around the region. As you can see at the set of samples of well E12, in the Annex (see Fig. 7.0.3), unfolding phenotype causes wide and formless spots, which with edge detection implementation are converted in regions with some spreadings like in Fig. 4.2.5, which lead to a bigger bounding box than in the compacted regions from normal cells, where the bounding box and the region share a higher number of pixels.

   iv. $areasR$: Ratio between the minimum area and the maximum area inside the nucleolus.

   For features i,ii and iii we save the maximum and minimum of the set of values obtained for each feature. We also save these values for the biggest region in the nucleolus with the aim of finding a pattern in wells H12, on which we have observed a large white spot characterizing the nucleoli repeatedly, as you can see in the set samples in the Annex (see Fig. 7.0.6). In particular, we also keep the mean of the set of values in $MA_{regions}$.

3. Comparison of number of modes obtained with edge detection $numRegions$ and with mean shift $numModes$. We expect to find a lowest number of modes in edge

detection comparing to a mean shift configuration that finds the different peaks in a connected component, at least for the wells presenting unfolding phenotypes.

4.2.5 Area ratios

Depending on the phenotype found in the nucleolus, connected components will be more compacted or extended inside the nucleus perimeter. For instance, in case of a capping phenotype, the proportion of whites should be smaller than in an unfolding phenotype. To measure this assessments, we propose several size features:

1. Occupation ratio with thresholding: we binarize the images by different intensity thresholds and, for each threshold, we obtain a value $OccupationTH$ that determines the ratio of white pixels within the nucleus.
2. Nucleolus occupation ratio with edge detection: we use the labeled image obtained with the edge detection tool explained in section 4.2.3 and obtain a value $OccupationED$ that determines a ratio between the number of pixels in the nucleus.
3. Region occupation ratio with edge detection: We apply the same operation as in (2), but this time we separate each region obtained with edge detection. This way we can find the ratio $OccupationRegion$ between the number of the pixels in a region and the number of pixels in the nucleus or in the nucleolus, which is the sum of regions. Once we get the ration $OccupationRegion$ for each region, we proceed to choose a criterion to select one. These measures should allow us to look for the large spot that characterizes the class H by looking for the proportion of the largest region within the whole nucleolus. We also look for the ratio between the smallest region and the largest region.

4.2.6 Grey Level Aura Matrices

Grey Level Aura Matrices (GLAM) [13] is a similarity measure originally proposed for modeling textures. Using a conventional notation, an image X is modeled as a finite rectangular lattice of mxn grids:

$$S = \{s = (i, j) \mid 0 \leq i \leq m - 1, 0 \leq j \leq n - 1\} \tag{4.2.2}$$

with a neighborhood system $N = \{N_s, s \in S\}$, where N_s is the neighborhood at site s. Neighborhood system N has two properties to satisfy:

1. Site s is excluded from its neighborhood
2. The neighborhood is symmetric

Each neighborhood system has an order number associated to determine its size at each site s. In an order d neighborhood system, the neighborhood at site s is given by:

$$N_{s=(i,j)} = \left\{r = (k, l) \mid 0 < (k - i)^2 + (l - j)^2 \leq d\right\} \tag{4.2.3}$$

To simplify, an order two neighborhood is used.

Given two subsets $A, B \subseteq S$, the Aura of A with respect to B for neighborhood system $N = \{N_s, s \in S\}$ is given by:

$$\vartheta_B(A, N) = \cup_{s \in A}(B \cap N_s) \tag{4.2.4}$$

And Aura measure of A with respect to B is given by:

$$m(A, B) = \sum_{s \in A} \mid B \cap N_s \mid \tag{4.2.5}$$

where for a given subset $A \subseteq S$, $\mid A \mid$ is the total number of elements in A.

Defining a partition of the lattice S as $\Im = \{S_i \mid 0 \leq i \leq G-1\}$, where G is the total number of gray levels in the image. Then, the Aura Matrix of $\Im$ over S is given by:

$$A(\Im) = [m(S_i, S_j)]$$

Intuitively, Aura measure $m(A, B)$ evaluates the amount of mixing sites between subsets A and B. A large value of $m(A, B)$ implies that the subsets A and B are mixed together while a small value implies that A and B are separated from each other.

Aura matrix is rotation invariant, which implies that the aura matrix of a given image will be the same no matter how the image is rotated. This property make GLAM suitable for analyze nucleolus images textures.

5 Experiments and validations

The environment and programming language used for all the code implementation and also for the image displays has been MATLAB, which has been developed by Mathworks Inc. corporation.

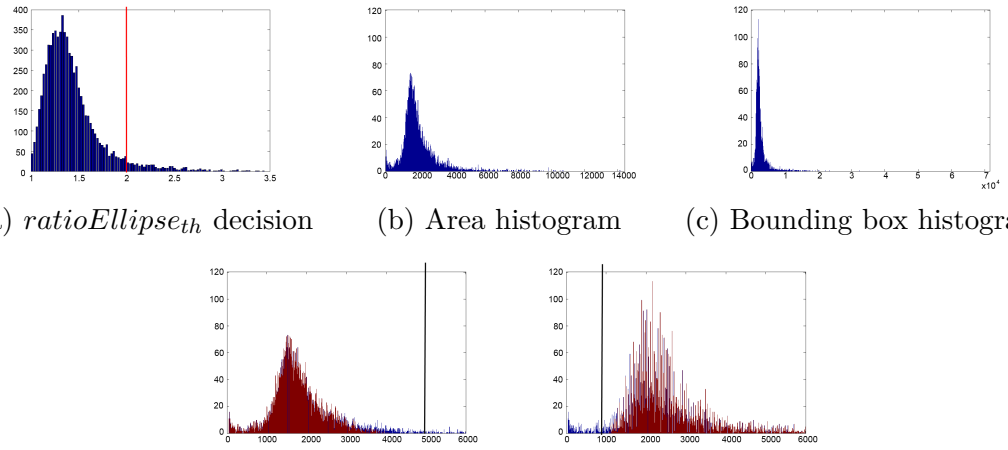
5.1 Validation methodologies

In this section, we discuss the values of the parameters that provide the best results for the image segmentation and the characterization by means of mean shift and edge detection. Following, we propose some metrics to determine whether two wells are similar or not based on each feature.

5.1.1 Segmentation parameters

The parameters finally set in the segmentation process are introduced below:

- $binary_{th}$ is set with a Matlab function based on Otsu's method, which chooses the threshold to maximize the separability of the resultant classes in gray levels.
- SE_{open}: we use a 7x7 matrix of ones as the structuring element to remove noise with the opening.
- $ratioEllipse_{th}$ is set to 2 due to the histogram at Fig. 5.1.1(a).
- To set the threshold for the area of the region $areaCell_{th}$ and the area of the bounding box $areaBox_{th}$ we first test some values to remove all the connected components outside an area and a bounding box area limits and next we merge them as you can see in Fig. 5.1.1(d) and (e). Fig. 5.1.1(d) shows the area histogram in a range [0,6000] overlapped with the same area histogram sliced for bounding box area greater than 5000. The figure proves that setting $areaBox_{th}$ to 5000 guarantees the removal of all the connected components with an area greater that 4500. Thus we set $areaBox_{th}$ to 5000.
- Fig. 5.1.1(e) presents the bounding box histogram in a range [0,6000] overlapped with the same bounding box histogram sliced for area values smaller than 900. The figure shows that setting $areaCell_{th}$ to 900 guarantees the removal of all the connected components with a bounding box area smaller than 1000. Thereby, $areaCell_{th}$ is set to 900.



(a) $ratioEllipse_{th}$ decision (b) Area histogram (c) Bounding box histogram

(d)$areaBox_{th}$ decision (e) $areaCell_{th}$ decision

Figure 5.1.1: Segmentation thresholds decision

5.1.2 Characterizing the nucleolus

In this section we introduce two different proposals to represent the segmented nucleolus image in terms of a limited number of fundamental structures. After these simplifications, we are able to define a number of features characterizing the spatial organization of those fundamental structures that approximate the segmented nucleolus and compare the results between cell classes.

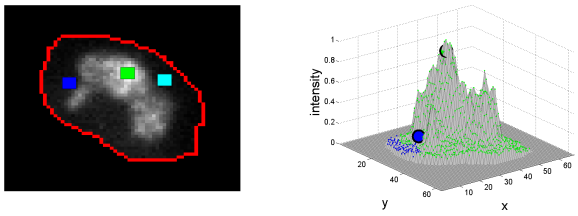
5.1.2.1 Mean shift in the nucleolus

When we apply mean shift algorithm to the nucleolus image, we obtain a set of modes and clusters characterizing the nucleolus. Together with lines, modes and clusters comprise the fundamental structures that characterize the nucleolus in terms of mean shift technique.

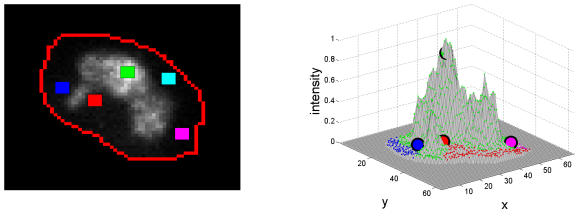
To determine the best combination of variables to obtain the most accurate results, we have tested the sample of nucleolus images shown in Figure Fig. 4.1.2. Variables checked to this purpose have been the next ones:

1. **Image:** Mean shift performance is tested on gamma image with two different gamma values 1.25 and 1.50 (see Fig. 4.1.2 (2) and Fig. 4.1.2 (3)). After some test, we chose the gamma of 1.50 because, despite mean shift finds more irrelevant modes with the higher gamma, as you can see in Fig. 5.1.2, irrelevant modes have lower intensity in these images.

2. **Mean shift values:** As we have explained in Section 3.4, after applying mean shift algorithm to the images, the resulting modes are post-processed to remove useless ones. For this reason, our objective with the choice of mean shift parameters is keeping all necessary modes safe even if it means obtaining some useless



(a) Mean shift applied on a nucleolus image mapped with a gamma of 1.25



(b) Mean shift applied on a nucleolus image mapped with a gamma of 1.50

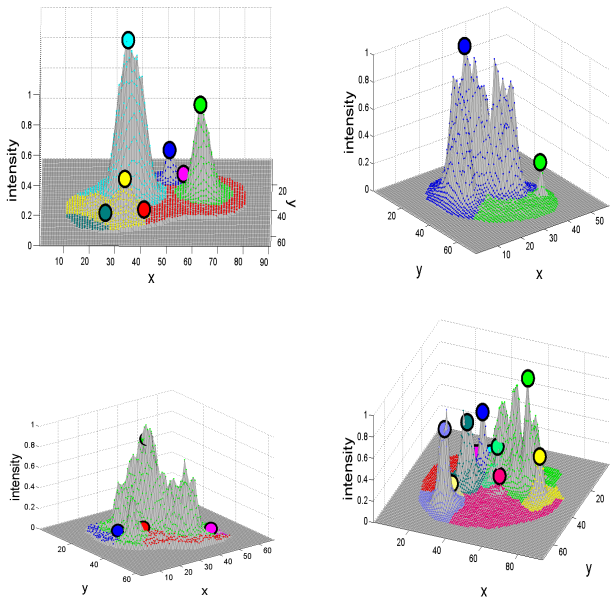
Figure 5.1.2: Image choice for mean shift

modes initially.

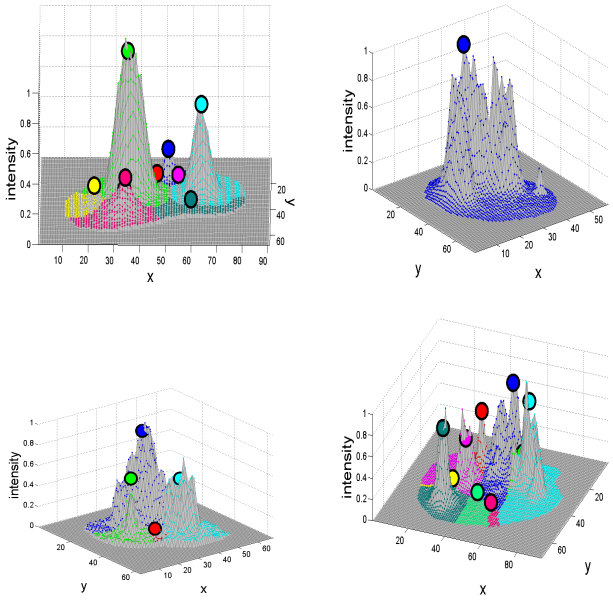
a) **Mean shift window width (bw_{MS}):** Tested in a range from 6 to 15 pixels. Test results show that a width of 6 is too small for nearby peaks but necessary to locate small spots when we have very big and small spots in the same nucleolus. On the other hand, values from 10 to 12 seem the most fitting values with which we find different peaks in a big spot affected by unfolding phenotype. Finally, a window width over 12 appears to be too big for tight peaks. Accordingly, we have decided to carry out our experiments with values 6 and 10, choosing the latter value for its greater efficiency against the window of width 12. Some examples are presented in Fig. 5.1.3 and Fig. 5.1.4where we show our conclusions graphically over the sample cells 2, 3, 4 and 6 of Fig. 4.1.2.

b) **Mean shift tolerance (mTol)**: Tested in a range from 0.01 to 0.50. Mean shift tolerance is a very relevant value in terms of efficiency because it is the key element that solves the number of mean shift iterations required to converge. Value 0.50 is not valid for some results while 0.01 is unnecessarily exhaustive, giving mostly the same results than a tolerance of 0.10 excluding some cases where it finds more modes which, however, are not relevant modes, as can be seen in Fig. 5.1.5. Therefore, we have decided to use a mean shift tolerance of 0.10 for our experiments.

c) **Clustering distance (cTol):** Tested in a range from 2 to 6 pixels. As mean shift tolerance is smaller than the distance among pixels, there is no need to establish the clustering distance as big as the mean shift window. Better results are obtained with values between 2 and 4 as shown in Fig. 5.1.6. We have decided to use a clustering distance of 3 because we have found that,

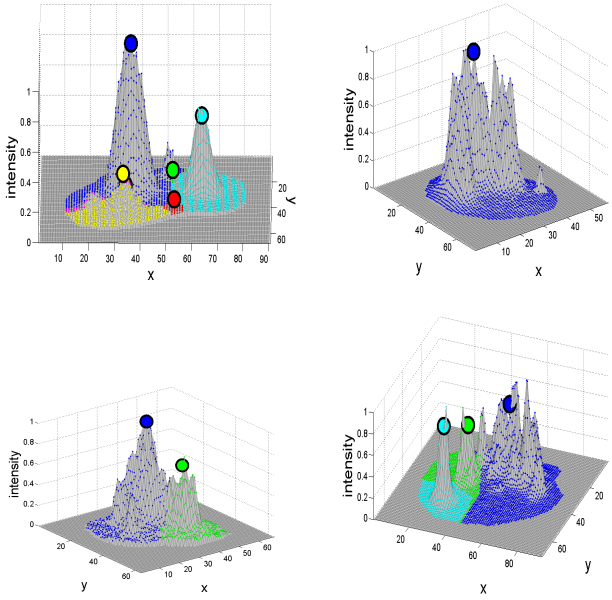


(a) $\mathrm{bw}_{\mathrm{MS}} = 6$



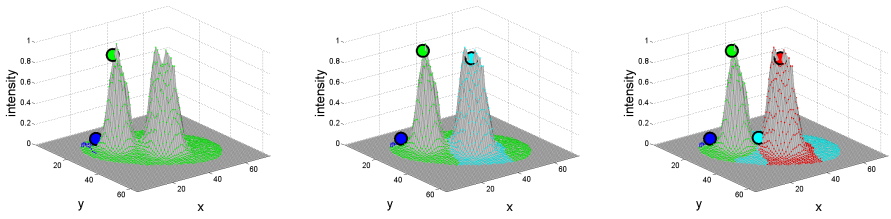
(b) $\mathrm{bw}_{\mathrm{MS}} = 10$

Figure 5.1.3: Choice of mean shift window width (1)



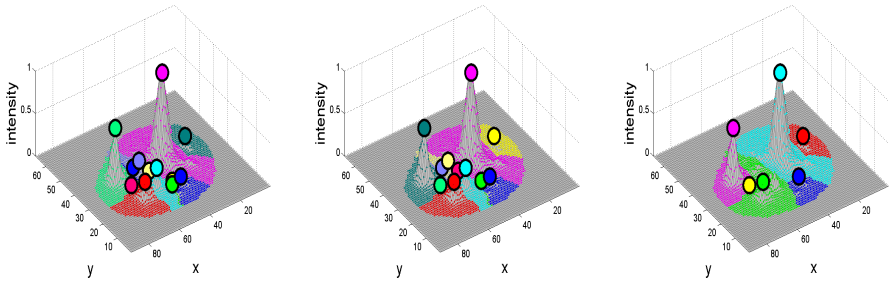
(c) $\mathrm{bw}_{\mathrm{MS}} = 15$

Figure 5.1.4: Choice of mean shift window width (2)



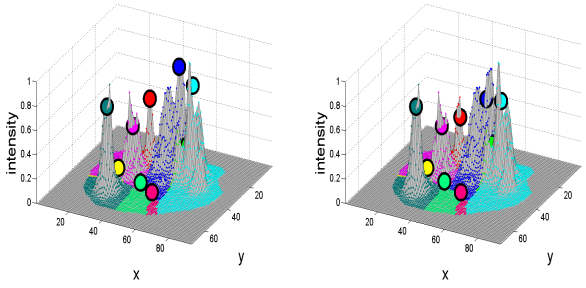
(a) mTol $= 0.50$ **(b)** mTol $= 0.10$ **(c)** mTol $= 0.01$

Figure 5.1.5: Choice of mean shift tolerance



(a) cTol $= 2$ **(b)** cTol $= 3$ **(c)** cTol $= 6$

Figure 5.1.6: Choice of clustering distance



(a) $\mathrm{bw}_{\mathrm{cluster}} = 3$ **(b)** $\mathrm{bw}_{\mathrm{cluster}} = 6$

Figure 5.1.7: Choice of clustering window width

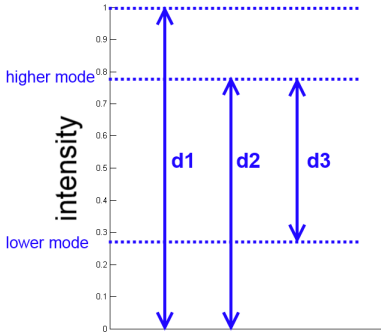
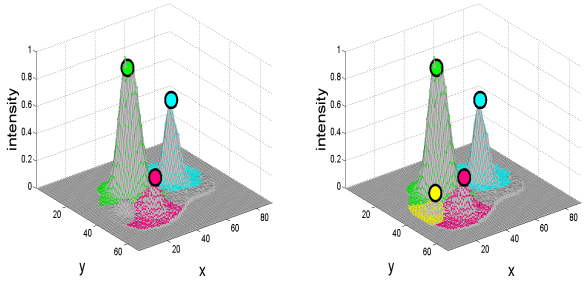


Figure 5.1.8: Distance proposals for thresholding

in some cases, a lower value solves many intermediate modes that must be removed in post-processing afterward while, in some other cases, higher values discard some useful modes.

d) **Clustering window width ($\mathbf{bw_{cluster}}$):** Tested in a range from 2 to 6 pixels, with better results obtained between 3 and 4 because, if we use bigger window sizes in images like cell 6 in Fig. 4.1.2, where we find very close and narrow modes, we get too low peaks because modes remain at the intersections between peaks. This fact is proved in Fig. 5.1.7 and urges us to choose a window width of 3 for our experiments.

3. **Thresholding values:** After applying mean shift, we proceed to remove useless modes with our two post-processing tools: thresholding and lines. Thresholding has two variables: distance and percentage of the distance. In Fig. 5.1.8 we present our three distance proposals for thresholding. Distance d1 fixes a threshold value which only depends on the percentage chosen, without being adjusted to the intensity range of the peaks. Consequently, d1 is too aggressive with those images characterized by low level modes as it is illustrated in Fig. 5.1.9. Distances d2 and d3 adjust their threshold values depending on the range of the modes achieving more accurate results. As we set the mode location with a mode seeking technique, the modes for clusters belonging to local maxima in the image



(a) 10% of d1 **(b)** 10% of d2 or 10% of d3

Figure 5.1.9: Choice of thresholding distance and percentage

are placed in the summit of the white spots and the difference between d2 and d3 does not affect them. Following, we have also done some test by setting threshold values to three different percentages of the proposed distances: 10%, 12% and 15% of the distance. In most tests, results are practically the same but sometimes 12% and 15% remove more modes than 10%. As higher useless modes are generally erasable with lines, our criterion to choose the suitable threshold value is that it is preferable to apply a cautious thresholding, although useless modes are not entirely removed, and then use lines to erase the remaining ones than to apply a greater threshold and remove useful modes. Thus, we decide to use a thresholding of 10% of distance d2.

4. **Lines values:** we have studied the different lines exemplified in Fig. 3.4.2 looking for several mean shift results after thresholding and we have decided to remove those clusters which lead to lines with a size smaller than 4 pixels and with a distance between pixels smaller than 2, for the dashed lines.

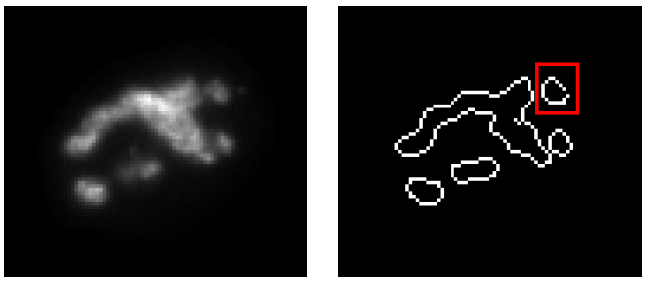
5.1.2.2 Edge detection in the nucleolus

When we apply the Edge detection tool to the nucleolus image, we obtain a set of modes and regions characterizing the nucleolus. Modes and regions comprise the fundamental structures that characterize the nucleolus in terms of edge detection technique.

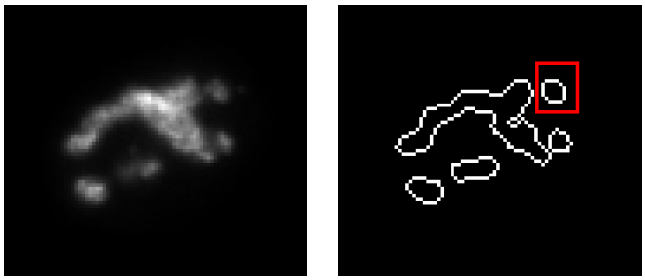
Following, we introduce the values of the parameters that we have set for edge detection configuration:

1. Image: we have tested gamma image with two gamma values of 1.25 and 1.50. As the greater gamma value is more aggressive with low levels, region boundaries are strongly defined and we avoid some opened contours that we obtain with gamma value 1.25 after the Laplacian of Gaussian filter. An example is shown in Fig. 5.1.10

2. SE_{close}: we implement four successive closings to ensure that regions contours have been completely closed before the hole filling process. Our four structuring

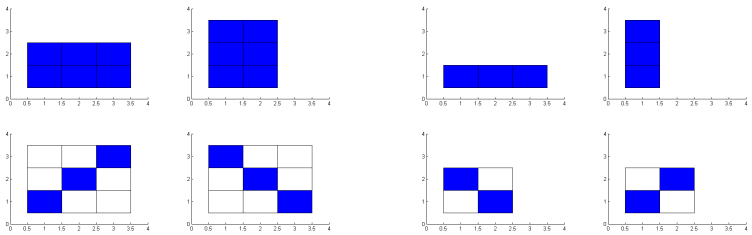


(a) Laplacian of gaussian filter with gamma 1.25



(b) Laplacian of gaussian filter with gamma 1.50

Figure 5.1.10: Gamma image choice for edge detection application



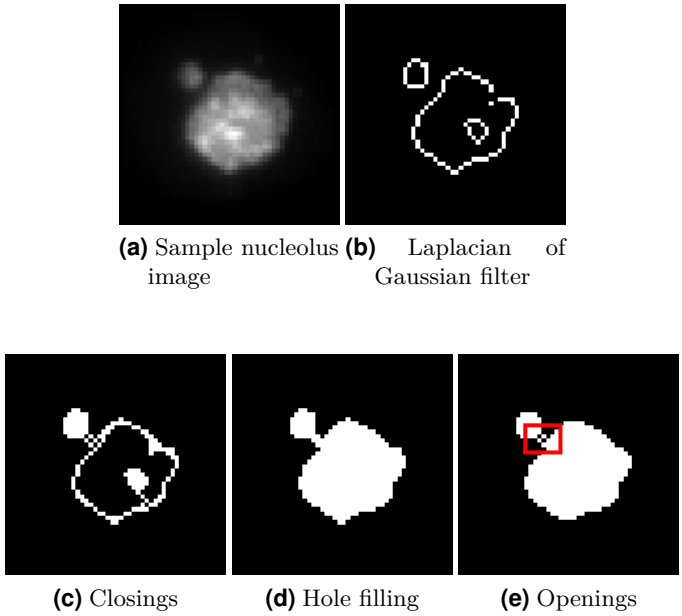
(a) SE for closings **(b)** SE for openings

Figure 5.1.11: Structuring elements in edge detection implementation

elements SE_{close} are defined by the four arrays in Fig. 5.1.11(a).

3. SE_{open}: to erase lines which where not closed before hole filling process mostly because they do not belong to any white spot in the image, we implement four successive openings with the structuring elements SE_{open}defined by the four arrays in Fig. 5.1.11(b).

4. $connect$: we set the connection to 4 in order to avoid clustering between nearby regions. As you can observe in Fig. 5.1.12(e), the openings implemented after the hole filling recover some regions that were merged with the hole filling process. In some cases, some one pixel regions remain between the two separated regions. If we apply a labeling with connection 4 to Fig. 5.1.12(e), we obtain 4 regions: the two real regions and the two pixels between them. Otherwise, if we apply a labeling with connection 8, we get just one region.



(a) Sample nucleolus image **(b)** Laplacian of Gaussian filter

(c) Closings **(d)** Hole filling **(e)** Openings

Figure 5.1.12: Edge detection example

5. $area_{th}$: we set this threshold to 3, just to erase those remaining pixels between regions.

5.1.3 Phenotypes classification

In the Annex, you can find a set of 6 random samples for wells A12, B12, E12, F12, G12 and H12 in our database. If you take a look, you will easily distinguish the unfolding phenotype in well E12 as well as the capping phenotype in well G12 or the large spot found in the nucleolus of well H12. In this section, we try to find the way to discriminate these variations using the features presented in section 4.2.

We measure the discrimination of a feature by computing the histograms for the different wells separately and, right after, compare histograms between them. To be precise, we compare the wells from normal class A12 and B12 among them in order to prove their similarity and, afterward, we compare normal class altogether with the different wells from the abnormal class. Results show that it is not feasible to compare normal class to abnormal class by merging all the abnormal wells together. This is because abnormal wells present different phenotypes and, consequently, they respond differently to the various tested features thus when merging together, the histogram expands its range making impossible to discriminate any feature.

The metric we use to compare two histograms is the Earth Mover's Distance (EMD) [[14], Sections 1 and 2, page1-2]. EMD is based on the cost that must be paid to transform one distribution into the other and is defined for signatures of the form $\{(x_1, p_1)..(x_m, p_m)\}$,where x_i is the center of the histogram bin i and p_i is the num-

EMD evaluation criterion	
Well A12 versus Well B12	Normal Class versus Abnormal Well
0-9	>15-20
10-19	> 30-35
20-29	> 45-50
30-39	> 75-80
40-49	> 100
50-59	> 120
60-69	> 140
> 70	> 160

Table 5.1: Thresholding values to consider whether a feature is discriminant according to the reference well A12 versus well B12

ber of cells that belong to i. Given two signatures $P = \{(x_1, p_1)..(x_m, p_m)\}$ and $Q = \{(x_1, p_1)..(x_n, p_n)\}$, the EMD is defined in terms of an optimal flow $F = f_{ij}$, which minimizes:

$$W(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij} \tag{5.1.1}$$

where $d = d(x_i, y_j)$ is the Euclidean distance between x_i and y_j. In the EMD terminology $W(P, Q, F)$ is the work required to move earth from one signature to another. Once the optimal flow f_{ij}^* is found, the Earth Mover's distance between P and Q is defined as:

$$EMD(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}^* d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}^*} \tag{5.1.2}$$

We have used the EMD Matlab code in [15] for our experiments.

Based on the observation of the EMD values, to determine whether two distributions are similar enough to consider that a feature is discriminant, we fix the thresholding distance independently for each feature taking as a reference the distance obtained for well A12 versus well B12, while the EMD values change depending on the type of distribution observed and the range of the feature. Meanwhile, we present a table in Tab. 5.1 with the thresholding values in reference with the EMD values from well A12 versus well B12. Besides, when looking at the EMD values, we have to take into account that our ideal target is to obtain two separate histograms like in Fig. 5.1.13 which are unalike enough to let us set a thresholding value, which determines whether a cell belongs to one histogram or to the other by a binary comparison. For this reason, we have to use EMD values carefully, because EMD can find distributions with a high cost

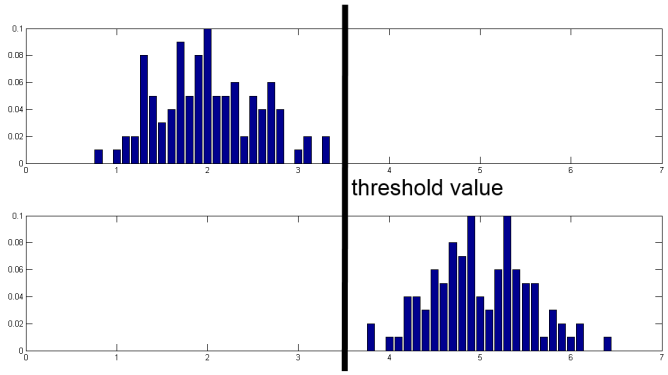
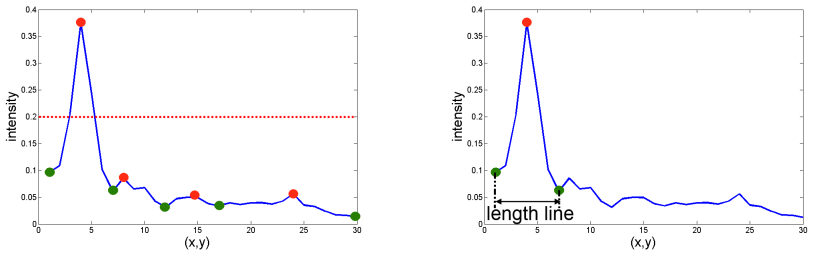


Figure 5.1.13: Ideal target of the thesis



(a) Line extracted from a cluster (b) cut Line

Figure 5.1.14: Cutting lines before feature extraction

required to transform one distribution into the other but the particular differences that make it costly may not fit our objectives. In section 5.2 we present the tables of EMD values for the tested features of interest always referenced with normal wells values, and their respective histograms.

Before the line choice expounded in Section 4.2.3, we proceed to cut the line with a thresholding value of 0.2. This means that modes below this threshold are not taken into account in the choice procedure. Besides, the cutting removes those parts of the line that do not belong to the white spot represented by the cluster and redefines the line length. Cutting is illustrated in Fig. 5.1.14. In case there is a maximum above the cutting line between two minima below it, we maintain the maximum and also the minima between the two maxima.

To validate the lines from mean shift clusters, we have designed a typical pathological cell model that characterizes two types of lines obtained from typical normal modes and two types of lines obtained from typical abnormal modes. We present the model in Fig. 5.1.15. On the one hand, typical normal line 1 is clean and small while typical normal line 2 is big and can have oscillations on the top, like it is shown in Fig. 5.1.18. On the other hand, typical abnormal line 1 is tall and thin while typical abnormal line 2 is big and have oscillations on the bottom. Fig. 5.1.16 shows the two widths proposed in Section 4.2.5 for our model lines and Fig. 5.1.17 shows the locations (heights in red and minima in green). In Fig. 5.1.18 it is illustrated the impact of $ratioLine$ over typical

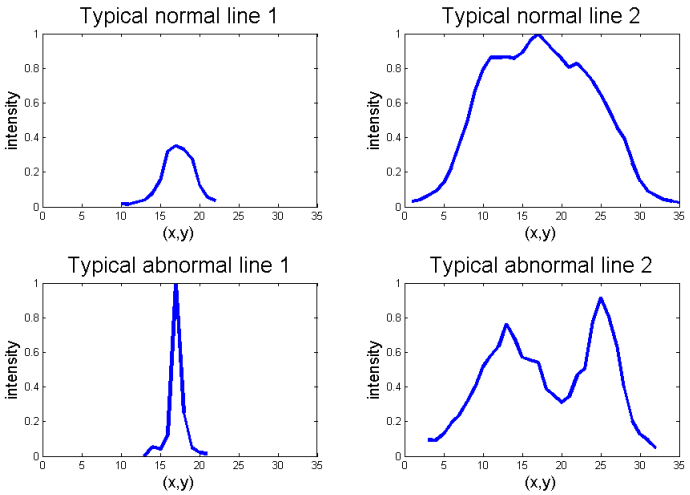


Figure 5.1.15: Model for typical normal and abnormal lines

normal and abnormal lines 2 of our model.

We measure the validity of our model by a graphical comparison of normal class and abnormal wells between two features. This comparison is implemented by drawing together two features from the same class or well, one in each axis of the 2-D line plot, and see how they are related one another. The line features chosen for this measure are presented in Tab. 5.2. The aim of this comparisons is to separate the two typical lines for normal and abnormal behavior, which are merged in the histograms, and set a thresholding line between normal and abnormal samples.

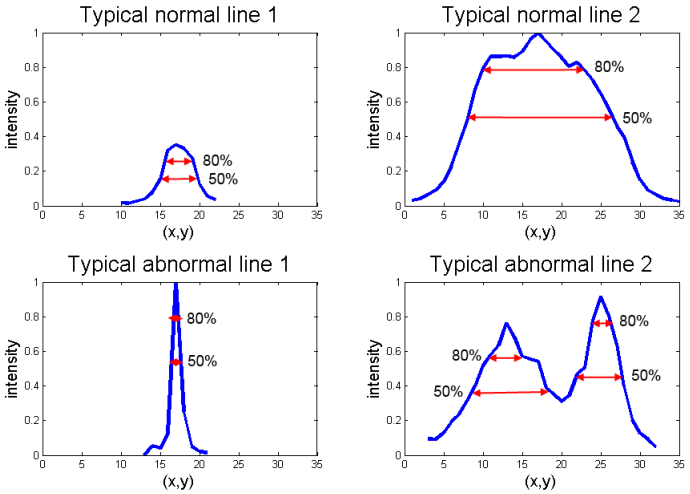


Figure 5.1.16: *width* in typical normal and abnormal lines

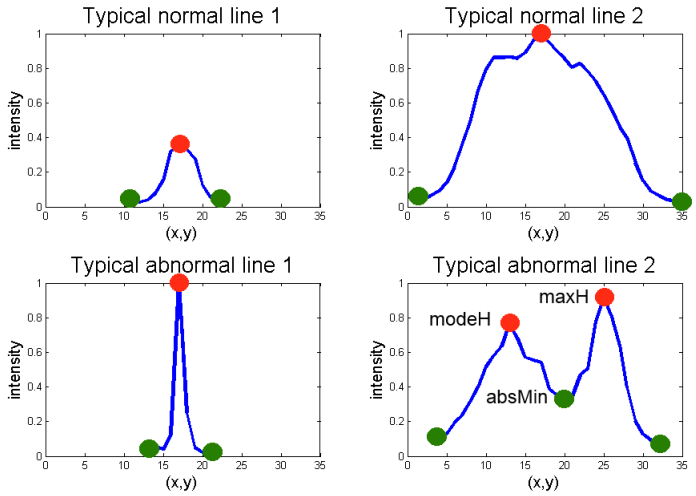


Figure 5.1.17: Locations: height ($modeH$ and $maxH$) in red and $absMin$ in green for typical normal and abnormal lines

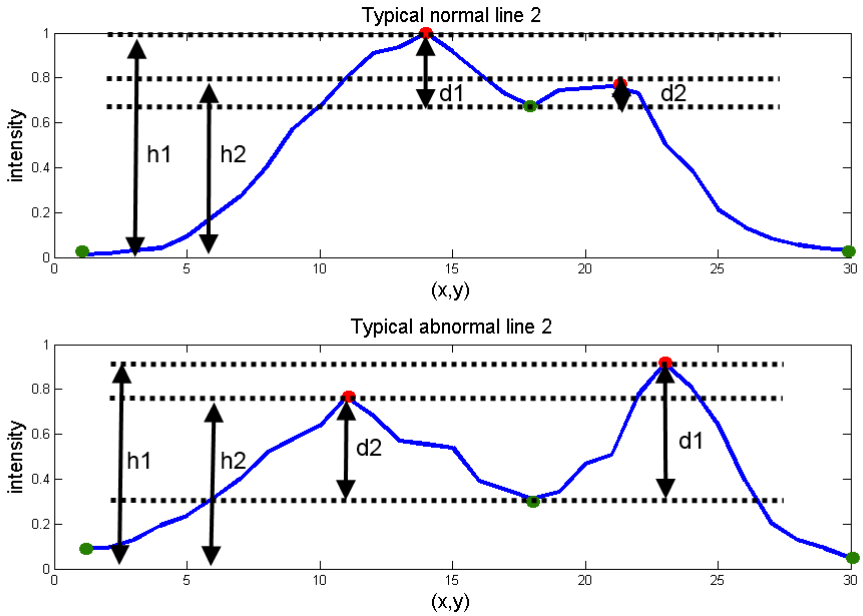


Figure 5.1.18: $ratioLine$ in typical normal and abnormal lines

Feature in X axis	Feature in Y axis
$height$	$width$
$height$	$numOsc$
$ratioLine$	$width$
$position$	$lengthLine$
$width$	$lengthLine$

Table 5.2: Measuring the relation between features in Lines

EMD	
Well A12 versus Well B12	24.8571
Normal class versus Well C12	**60.8143**
Normal class versus Well D12	**59.0238**
Normal class versus Well E12	**49.9238**
Normal class versus Well F12	40.2857
Normal class versus Well G12	44.8048
Normal class versus Well H12	33.7000

Table 5.3: EMD values for feature $intensity_{max}$

5.2 Results

In this section we test the features presented in Section 4.2 and provide the results obtained by applying the metrics proposed in Section 5.1. It is worth mentioning that features with a wide range have been normalized to be in a range within 0 and 1 for the metrics calculation and displays.

5.2.1 Intensity

We present the EMD criterion for the feature $intensity_{max}$ over norm image in Tab. 5.3. These EMD values prove the high discrimination that presents this feature in wells C12 and D12, which can also be checked visually with the histograms in Fig. 5.2.1. Due to these satisfying EMD values, we continue the rest of our experiments discarding wells C12 and D12, since we want to classify the rest of the set of abnormal wells: E12, F12, G12 and H12.

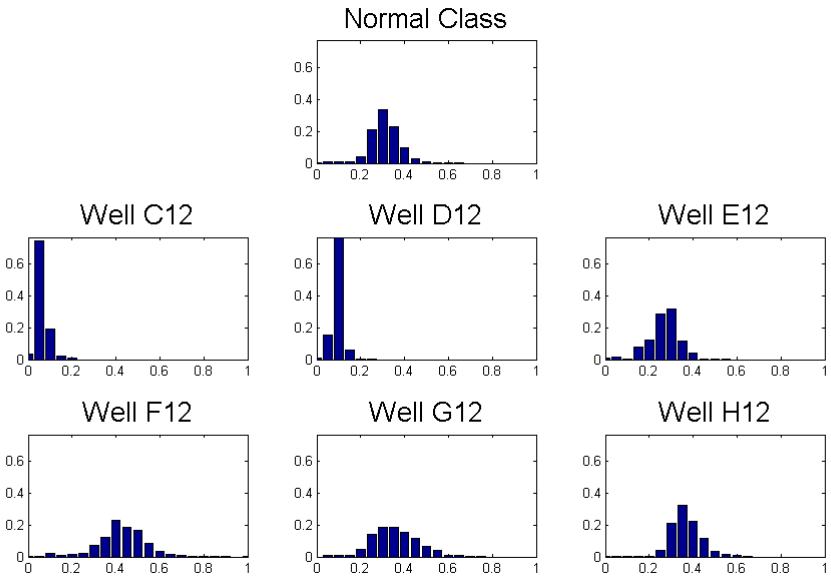


Figure 5.2.1: Histograms for feature $intensity_{max}$

EMD	
Well A12 versus Well B12	40.6222
Normal class versus Well E12	90.9778
Normal class versus Well F12	55.5778
Normal class versus Well G12	68.8444
Normal class versus Well H12	40.6222

Table 5.4: EMD values for feature $numModes$

5.2.2 Mean-shift

Before testing the best parameters for mean shift configuration in Section 5.1.2.1, we have chosen two different variants for our experiments:

- Variant v1: bw_{MS}= 6, mTol = 0.1, cTol = 3 and $bw_{cluster}$= 3
- Variant v2: bw_{MS}= 10, mTol = 0.1, cTol = 3 and $bw_{cluster}$= 3

Feature $numModes$ has been tested for the two variants, getting several histograms in a range from 0 to 9. In Fig. 5.2.2, we present the histograms for v1 which show that this feature is not discriminative at all in exception with the moderately satisfactory result on G12, in which we see that the average number of modes is slightly higher compared to the rest. EMD values from v2 are almost identical to values from v1. We list them in Tab. 5.4.

$intensity_{modes}$	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
maximum v1	22.6143	**47.1190**	37.9952	41.1762	30.3286
mean v1	20.8333	**44.0190**	**38.0190**	36.4714	29.4905
maximum v2	24.9143	**52.0286**	34.9571	**46.5571**	34.3190
mean v2	20.3524	**43.3524**	32.9048	28.8095	20.3524

Table 5.5: EMD values for feature $intensity_{modes}$

We also test feature $intensity_{modes}$ for the two variants. We present EMD values in Tab. 5.5 and their respective histograms in Fig. 5.2.3, Fig. 5.2.4, Fig. 5.2.5 and Fig. 5.2.6. We observe that the maximum of $intensity_{modes}$ from well E12 tend to be lower than the maximum from normal class while the maximum from wells F12, G12 and H12 tend to be greater. The same occurs for the mean of $intensity_{modes}$ mainly with F12.

For shape features, we present a table with the resulting EMD values in Tab. 5.6 and the respective histograms in Fig. 5.2.7, Fig. 5.2.8, Fig. 5.2.9, Fig. 5.2.10, Fig. 5.2.11, Fig. 5.2.12, Fig. 5.2.13 and Fig. 5.2.14. Logically, we notice that shape features are more discriminant over v1, being the variant with bw_{MS}= 6, because $bw_{MS} = 10$ allows us to find the different peaks within a spot affected by unfolding phenotype while $bw_{MS} = 6$ is not able to find nearby peaks merging them in the same cluster leading to large lines in abnormal cases. We observe that shape features in mean shift are particularly discriminative with E12 and G12 in opposite directions, being this wells the wells with a strongest unfolding and capping phenotypes respectively. Feature $area_{modes}$ is normalized by the area of the nucleus.

For the same reason that has been explained above, we do the $numModes$ comparison subtracting the number of modes in bw_{MS}= 6 to the number of modes in bw_{MS}= 10 thus we subtract v1 to v2. Results are expounded in Tab. 5.7 and their histograms in Fig. 5.2.15. Being well E12 the cells exhibiting the strongest unfolding phenotype of the controls, it is not surprising that it obtains the higher EMD values for this particular feature.

$area_{modes}$	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
ratio v1	14.5286	**35.0238**	11.6381	**35.2238**	14.5286
ratio v2	27.1714	**58.7762**	16.1190	28.8905	39.7905

ecc_{modes}	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
maximum v1	49.6857	**117.3286**	66.9857	**106.8810**	89.2571
minimum v1	21.3429	**52.4429**	34.3000	**70.7429**	33.3000
mean v1	37.0571	**87.9238**	54.0286	**94.0667**	64.1476
maximum v2	45.3000	**109.3000**	76.2190	**122.9095**	77.8238
minimum v2	10.6810	22.9381	14.9048	**38.2381**	5.2381
mean v2	26.0095	**60.5333**	42.7857	**78.7571**	35.8857

Table 5.6: EMD values for shape features in mean shift

EMD	
Well A12 versus Well B12	72.0000
Normal class versus Well E12	**181.4167**
Normal class versus Well F12	91.5278
Normal class versus Well G12	150.5278
Normal class versus Well H12	131.5278

Table 5.7: EMD values for comparison of $numModes$

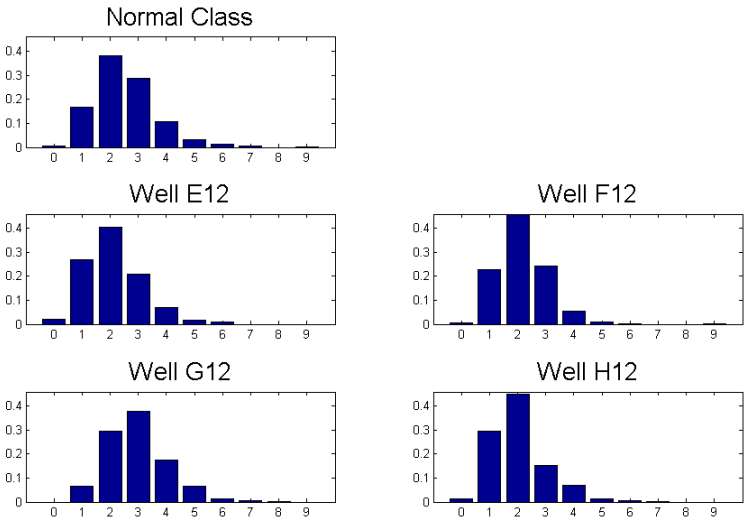


Figure 5.2.2: Histograms for feature $numModes$

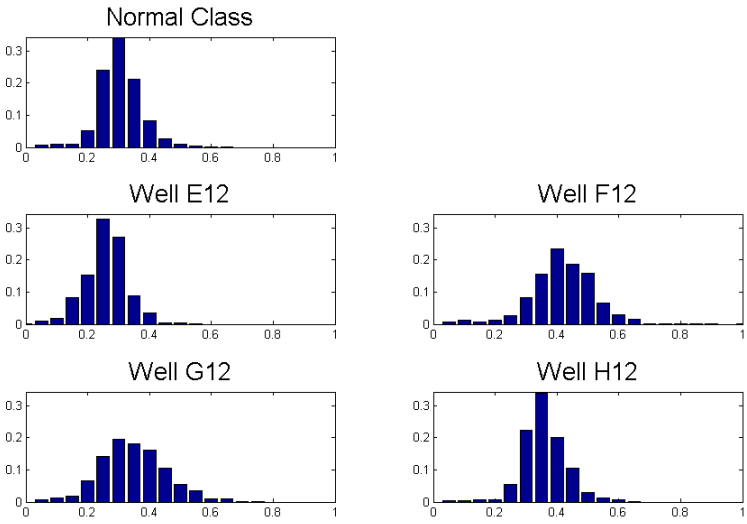


Figure 5.2.3: Histograms for maximum of $intensity_{modes}$ in v1

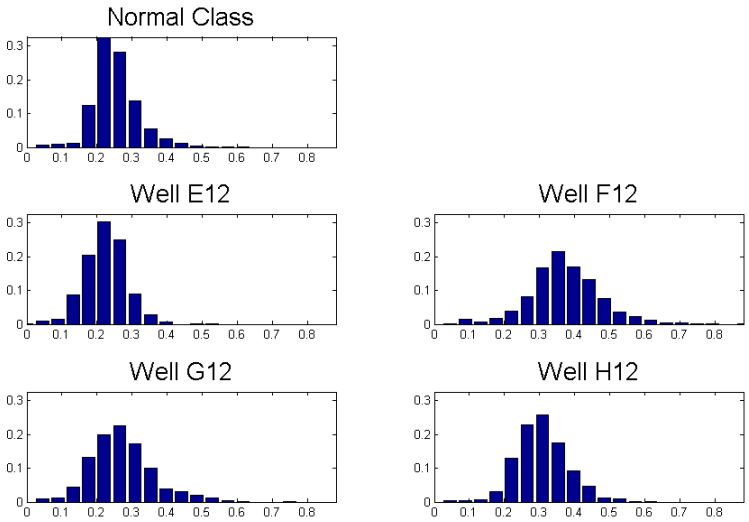


Figure 5.2.4: Histograms for mean of $intensity_{modes}$ in v1

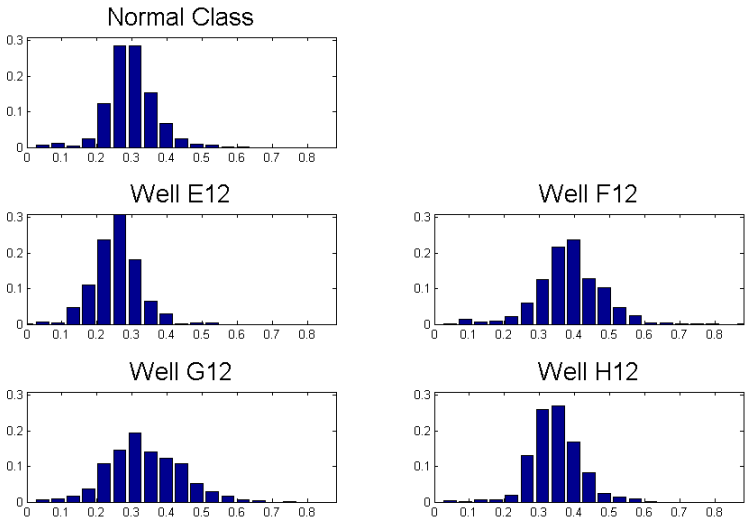


Figure 5.2.5: Histograms for maximum of $intensity_{modes}$ in v3

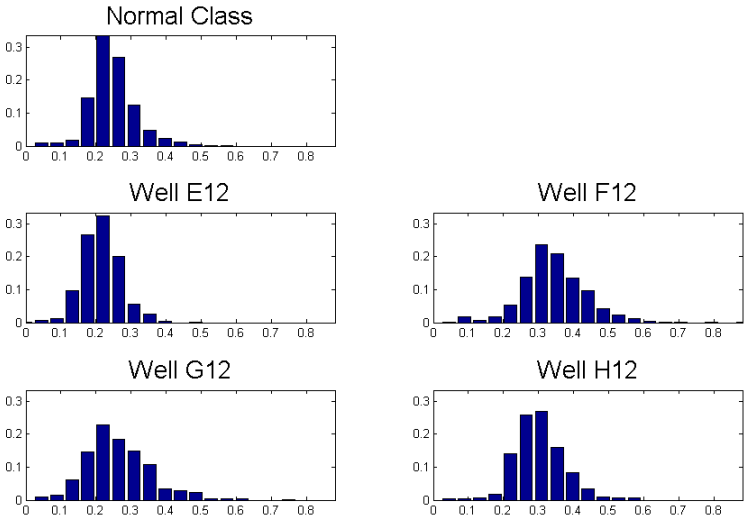


Figure 5.2.6: Histograms for mean of $intensity_{modes}$ in v3

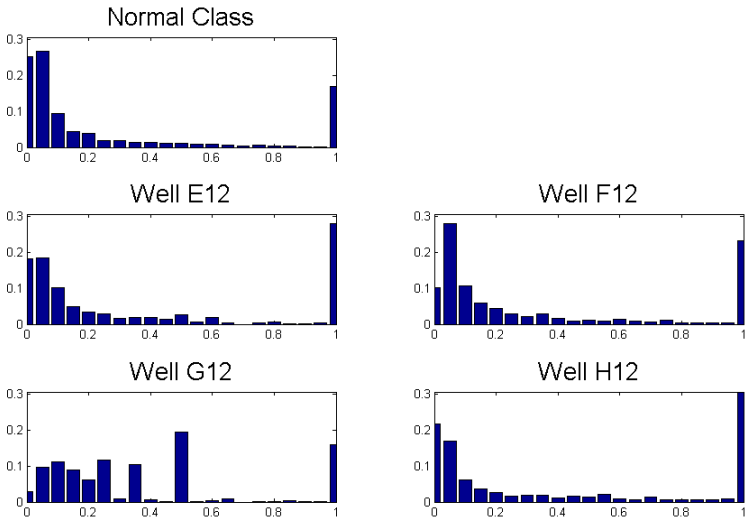


Figure 5.2.7: Histograms for ratio of $area_{modes}$ in v1

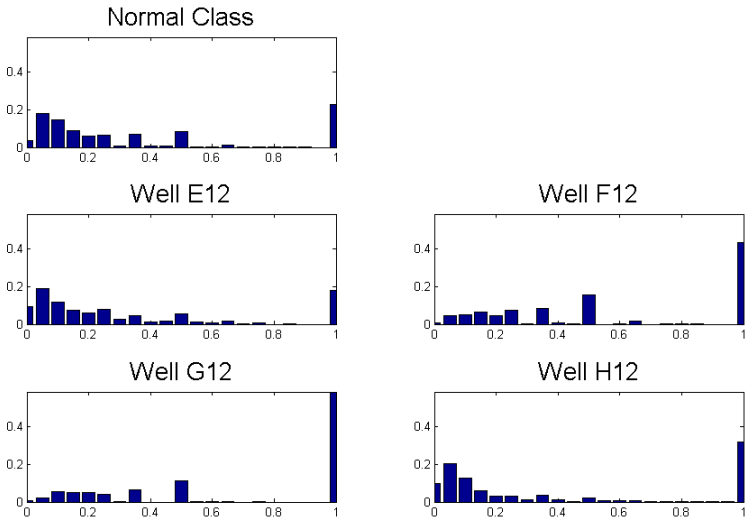


Figure 5.2.8: Histograms for ratio of $area_{modes}$ in v2

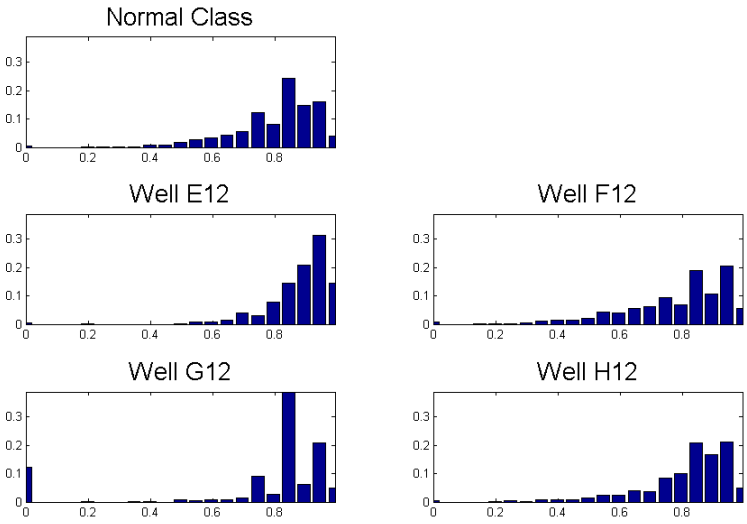


Figure 5.2.9: Histograms for maximum of ecc_{modes} in v1

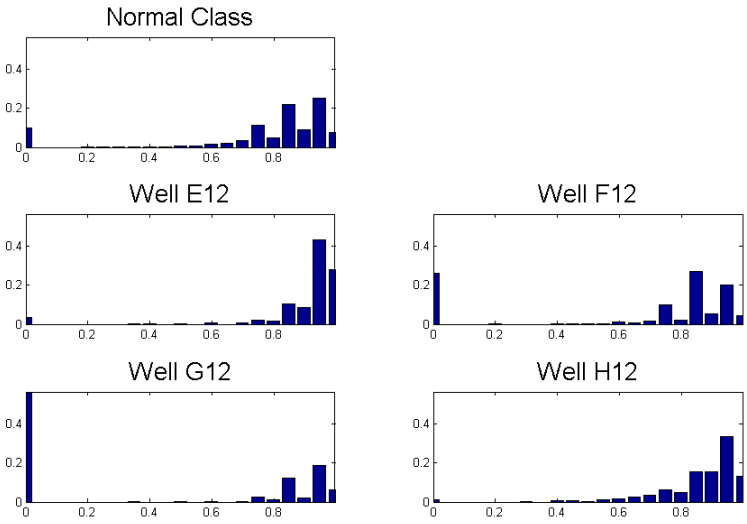


Figure 5.2.10: Histograms for maximum of ecc_{modes} in v2

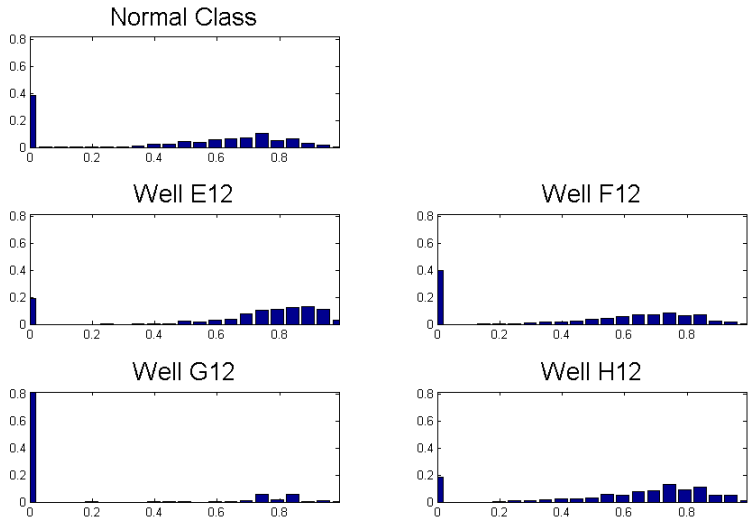


Figure 5.2.11: Histograms for minimum of ecc_{modes} in v1

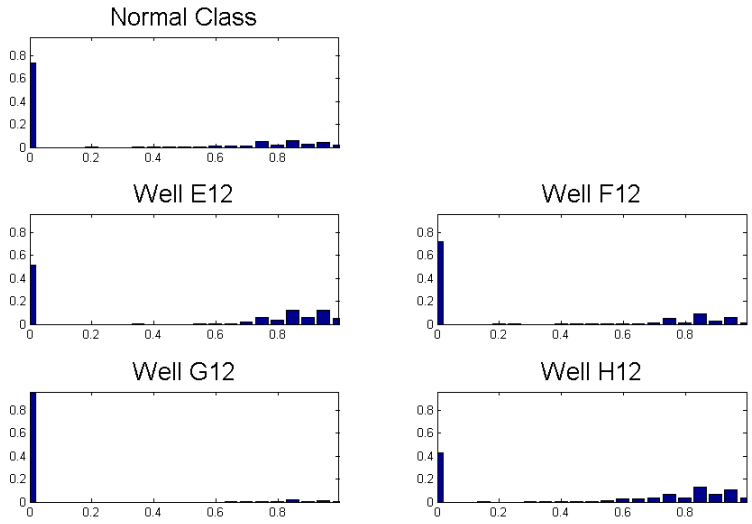


Figure 5.2.12: Histograms for minimum of ecc_{modes} in v2

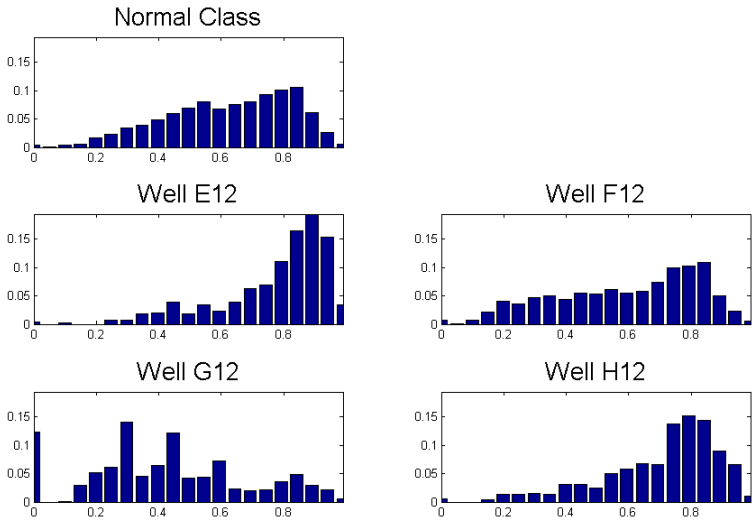


Figure 5.2.13: Histograms for mean of ecc_{modes} in v1

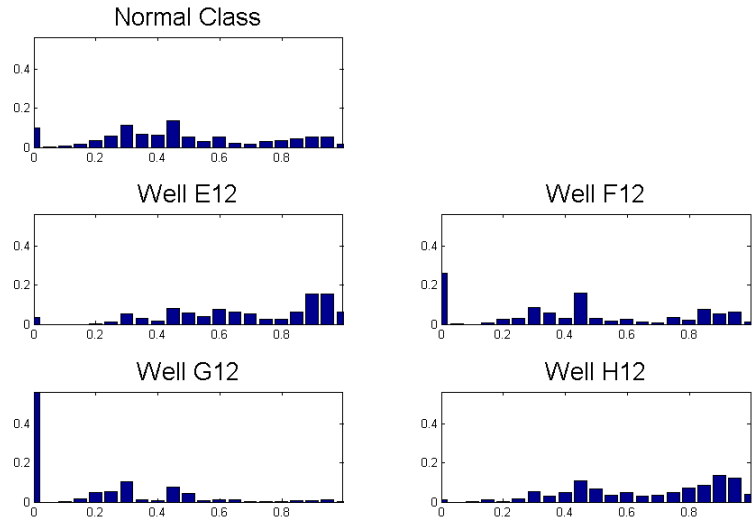


Figure 5.2.14: Histograms for mean of ecc_{modes} in v2

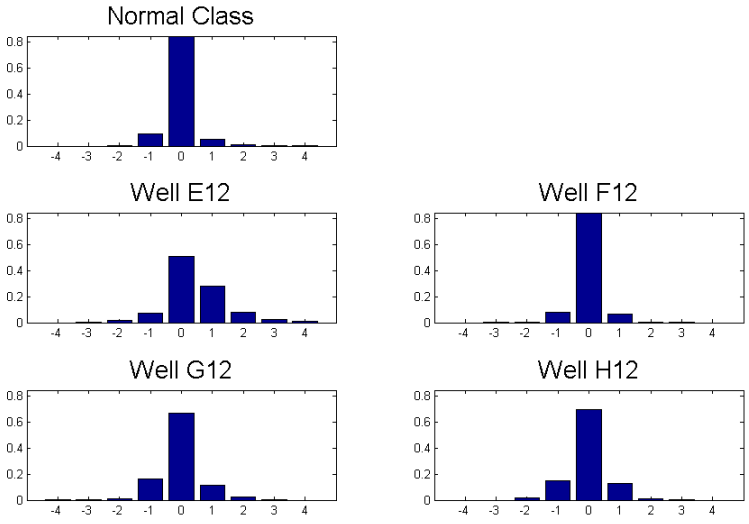


Figure 5.2.15: Histograms for the comparison of $numModes$

$position$	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
$maxH$	49.6476	**121.2238**	64.9143	**101.4476**	90.9571
$modeH$	50.2190	**122.3762**	65.6048	**102.7000**	92.5714
$absMin$	37.9333	77.1048	50.6143	**93.0524**	54.1952

Table 5.8: EMD values for feature $position$

$numOsc$	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
line1	6.8429	13.9762	9.5238	**15.2810**	9.8333
line2	7.8762	**17.5524**	11.4095	**19.2619**	12.7714

Table 5.9: EMD values for feature $numOsc$

Following, we proceed to do the validation of Lines features. Matlab code used to find the maxima and minima in Lines can be found in [16]. We present the histograms of the features expounded in 4.2.5 with the exception of features $normVal$ and $gammaVal$, which have been already tested with mean shift feature $intensity_{modes}$. Next, we measure the relation between the features in Tab. 5.2 and compare this relation in normal class and abnormal wells. We choose the v1 due to the property of bw_{MS}= 6 to merge peaks in the same cluster. In this case, lines have more information for abnormal lines than with v2 while in normal lines the information is almost identical in both variants.

First, we present the results for feature $position$ in line1, being the results for the four lines (line1, line2, line31 and line32) quite similar. EMD values are found in Tab. 5.8 and histograms in Fig. 5.2.17, Fig. 5.2.18 and Fig. 5.2.19. We can see that the capping phenotype of well G12 is slightly detected by the $positions$ of $maxH$ and $modeH$ moving the histograms to the right side while the unfolding phenotype in E12 moves the histograms of $maxH$ and $modeH$ to the left, indicating a the decentralization of the height $position$. In Fig. 5.2.19 we find an interesting result for well G12: Lines have a large part of the $absMin$ locations placed very close to one end of line, which suggests that a large part of those lines have only one oscillation.

For feature $numOsc$, EMD values from line1 and line31 and from line2 and line32 are very similar. For this reason, we are presenting just the EMD values from line1 and line2. As it was to be expected, EMD values are higher for line2 and line32 because these lines were chosen especially for having the highest number of oscillations. EMD values are presented in Tab. 5.9 and their histograms are found in Fig. 5.2.20 and Fig. 5.2.21.

The same line choice is followed with $lengthLine$ and EMD values are presented in Tab. 5.10, with their histograms in Fig. 5.2.22 and Fig. 5.2.23. According to histograms, this feature is pretty discriminant with well G12 although this is not reflected by the EMD values for line1. This feature is normalized by the major axis of the nucleus.

$lengthLine$	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
line1	36.0667	50.0095	**87.2143**	62.5333	62.5333
line2	32.5143	75.7857	44.4190	**78.8619**	54.6000

Table 5.10: EMD values for feature $lengthLine$

$width$	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
$50max$	21.7667	**52.8810**	36.5095	**57.1810**	39.8810
$50mode$	20.8619	**51.8571**	35.4048	**55.8810**	38.6019
$80max$	11.8951	**32.6837**	21.3381	**32.3095**	23.5048
$80mode$	12.5476	**34.5286**	22.6286	**34.9143**	25.4286

Table 5.11: EMD values for feature $width$

For feature $width$ EMD values are very similar for the 4 proposed lines thus we present EMD values obtained for line1. As you can see, EMD values in Tab. 5.11 show that this feature is definitely discriminant for wells E12 and G12. Histograms in Fig. 5.2.24, Fig. 5.2.25, Fig. 5.2.26 and Fig. 5.2.27 confirm it showing that $width$ is smaller for these two wells compared with normal class. Also well F12 presents discriminant histograms despite EMD values is not reflecting it. Moreover, very similar EMD values are obtained from $width50max/width80max$ and from $width50mode/width80mode$, which implies that our mean shift configuration is working quite well because in most of the cases $maxH$ and $modeH$ are the same location. This feature is also normalized by the major axis of the nucleus.

For feature $ratioLine$ we calculate the ratio introduced in Equation 4.2.1 with $valGamma$ of $maxH$ to ensure that we are working with the absolute maximum of the line. EMD values for line1 are presented in Tab. 5.12 and their respective histograms in Fig. 5.2.28.

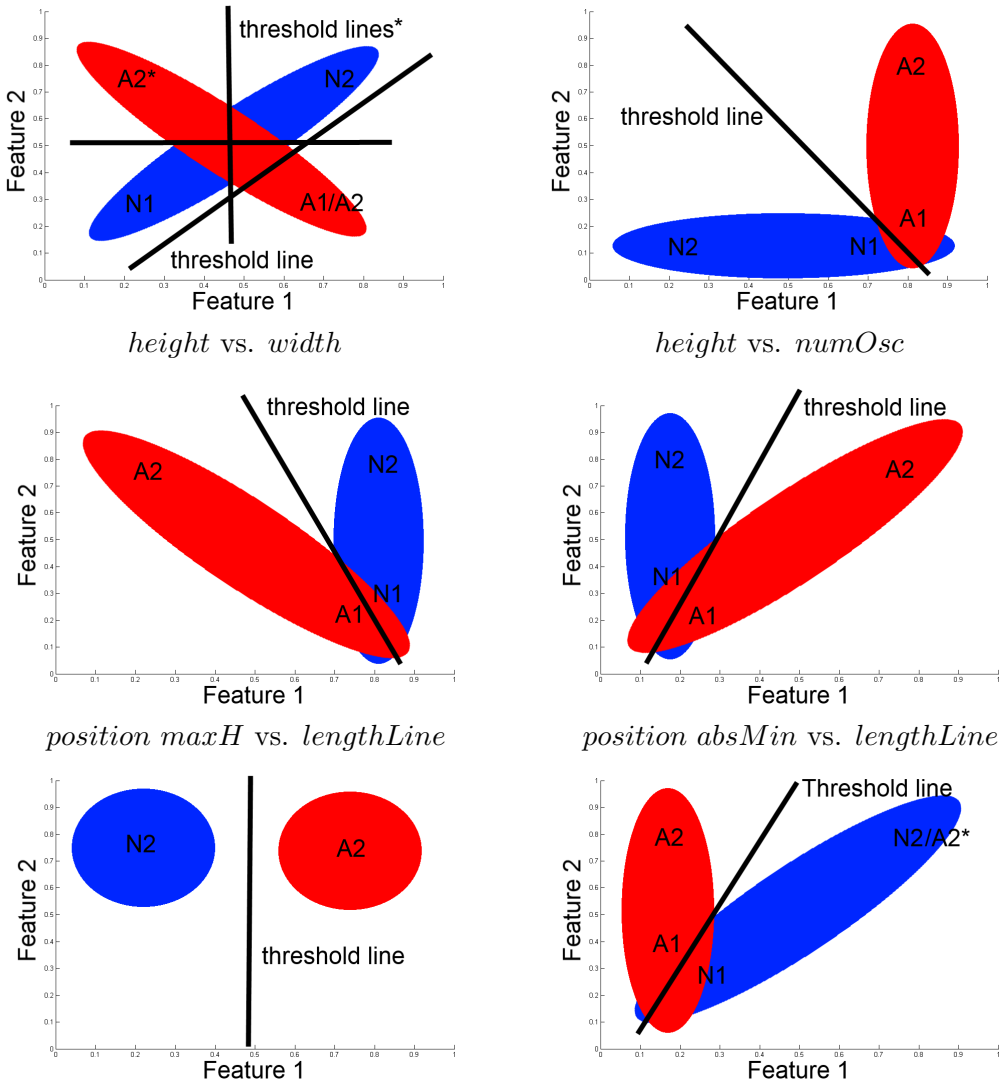
EMD	
Well A12 versus Well B12	5.8381
Normal class versus Well E12	9.7952
Normal class versus Well F12	6.8810
Normal class versus Well G12	12.4857
Normal class versus Well H12	7.2619

Table 5.12: EMD values for feature $ratioLine$

Now that we have already studied the behavior of many of the features proposed for lines individually, we continue with the measures of the relation between features. We present the ideal results for our model of typical normal and abnormal lines in Fig. 5.2.16, being N1 the values from typical normal line 1, N2 the values from typical normal line 2, A1 the values from typical abnormal line 1 and A2 the values from typical abnormal line 2. The asterisk over A2 found in some graphics indicate that the set of values A2 can change of place depending on the $width$ of the line. That is, if the minimum between two maxima is lower than the width placement, then the width will be small and values from abnormal line 2 will be located in A2 without asterisk but if the the minimum is higher, width will be larger and values from abnormal line 2 will be located in A2*. Particularly for $ratioLine$ vs. $width$, we just contemplate the typical lines 1 because the ratio is only calculated for those lines that have more than one oscillation. Following, real results are presented for line 1:

- $height$ vs. $width$ for lines with $numOsc$ = 1: we choose the $gammaVal$ and $normVal$ from $maxH$ to ensure that we work with the higher value in the line and test this comparison with $width50max$ and $width80max$. Graphics are presented in Fig. 5.2.29, Fig. 5.2.30, Fig. 5.2.31 and Fig. 5.2.32.
- $height$ vs. $numOsc$: due to the above result we choose this time just the value $normVal$ from $maxH$ to look for the relationship between the height and the number of oscillations. Graphics are presented in Fig. 5.2.33.
- $ratioLine$ vs. $width$. We have seen that feature $ratioline$ is not very discriminant individually but, as it is directly related with the $width$ of the line, we want to see if results improve by joining these two features. Results are shown in Fig. 5.2.34 and Fig. 5.2.35.
- $position$ vs. $lengthLine$. We test this comparison with $position$ from $maxH$ and from $absMin$ and graphics are shown in Fig. 5.2.36 and Fig. 5.2.37.
- $width$ vs. $lengthLine$. we use the width values from the maximum of the line $width50max$ and $width80max$ to carry out this measure. We present the results in Fig. 5.2.38 and Fig. 5.2.39.

You can see that the reality is far from the ideal models. There are a widespread of profiles within each normal and abnormal well and modeling them is quite difficult. Even so, there are some results that approximate the model for well G12, like in Fig. 5.2.37 and Fig. 5.2.38. Also some hopeful results are obtained for well F12 like in Fig. 5.2.30 or Fig. 5.2.38.



height vs. *width*

height vs. *numOsc*

position maxH vs. *lengthLine*

position absMin vs. *lengthLine*

ratioLine vs. *width*

width vs. *lengthLine*

Figure 5.2.16: Ideal results obtained with our model lines by comparison of features. In each image, blue spot corresponds to normal samples and red spot corresponds to abnormal samples

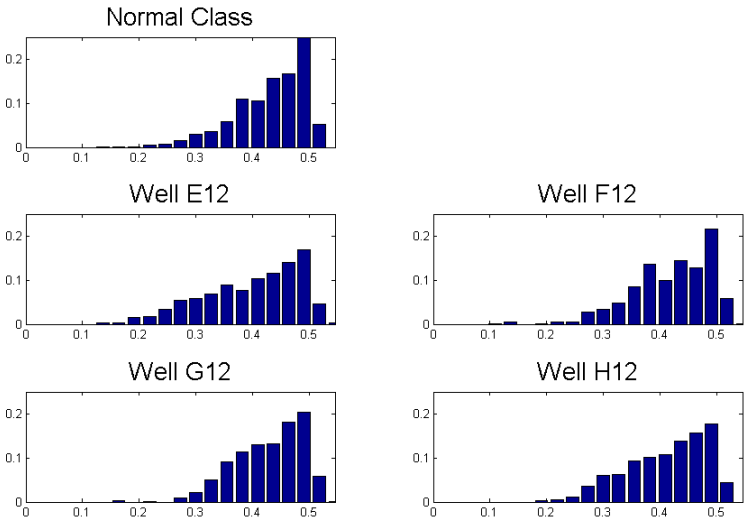


Figure 5.2.17: Histograms of feature *position* for $maxH$

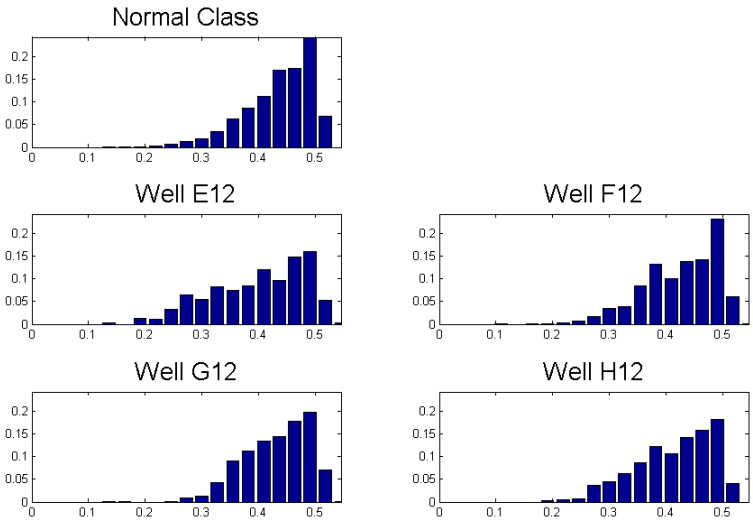


Figure 5.2.18: Histograms of feature *position* for $modeH$

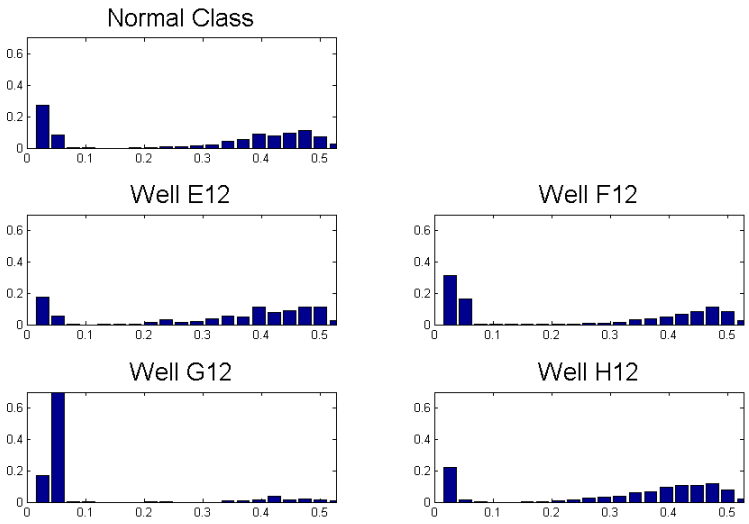


Figure 5.2.19: Histograms of feature *position* for $absMin$

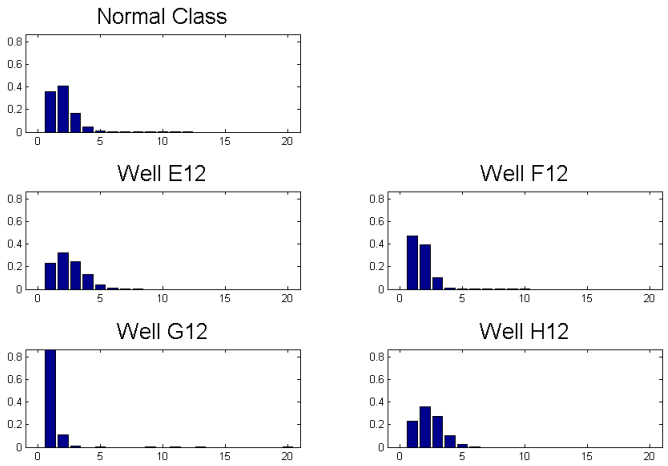


Figure 5.2.20: Histograms of feature $numOsc$ in line1

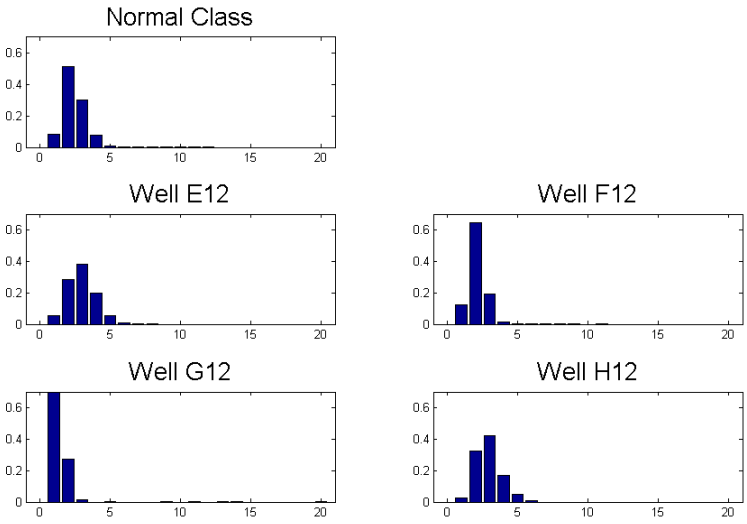


Figure 5.2.21: Histograms of feature $numOsc$ in line2

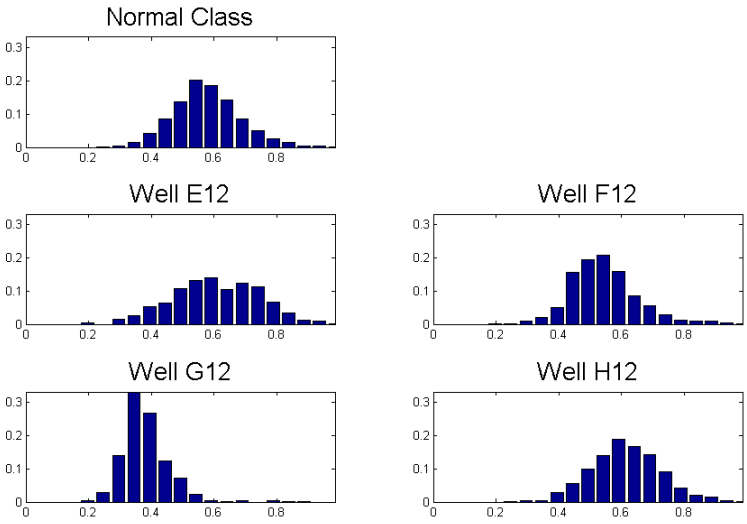


Figure 5.2.22: Histograms of feature $lengthLine$ in line1

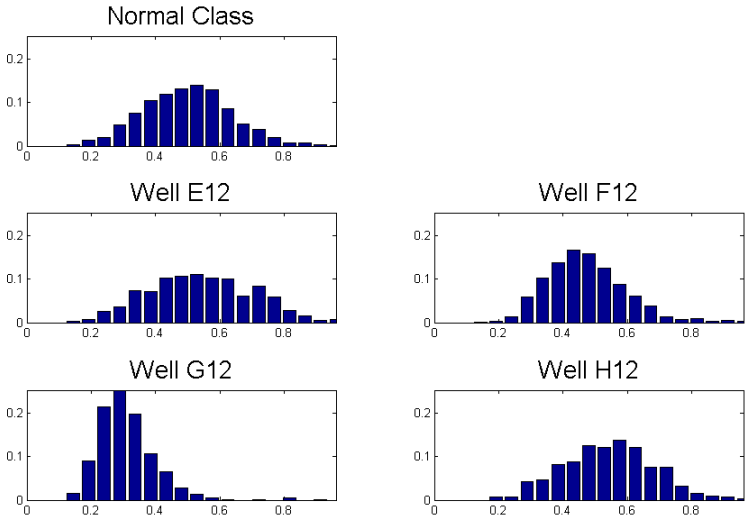


Figure 5.2.23: Histograms of feature *lengthLine* in line2

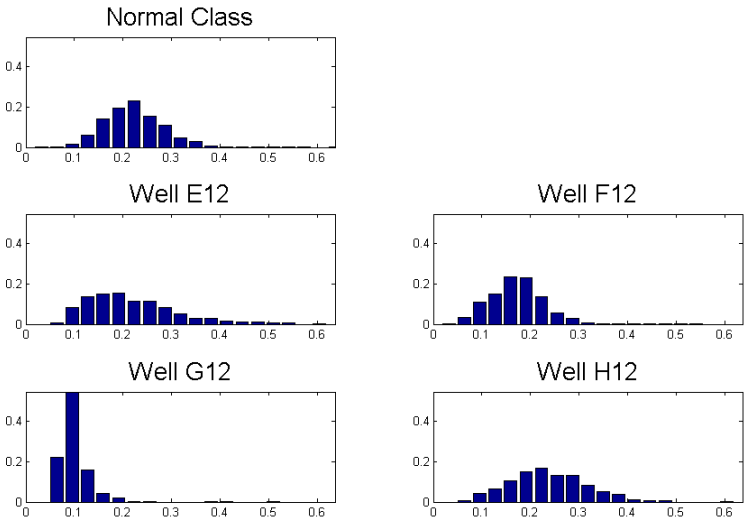


Figure 5.2.24: Histograms of feature $width50max$

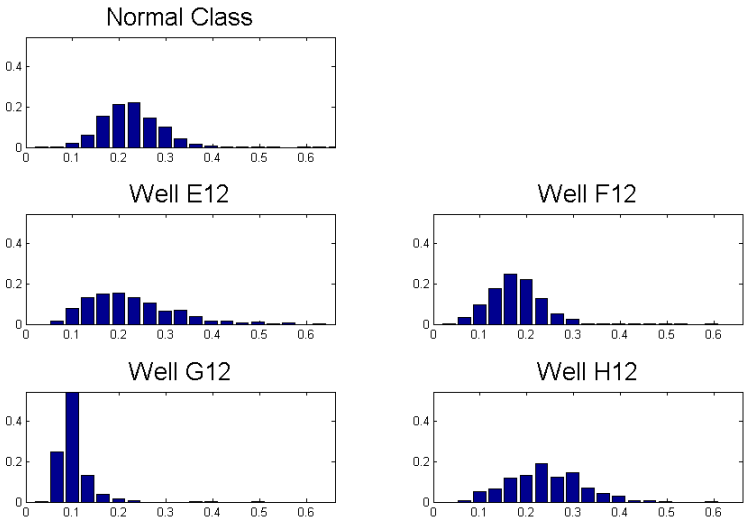


Figure 5.2.25: Histograms of feature $width50mode$

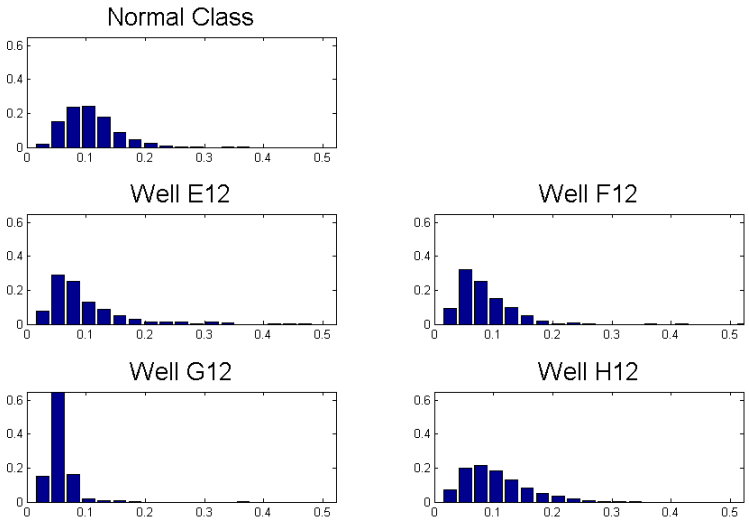


Figure 5.2.26: Histograms of feature $width80max$

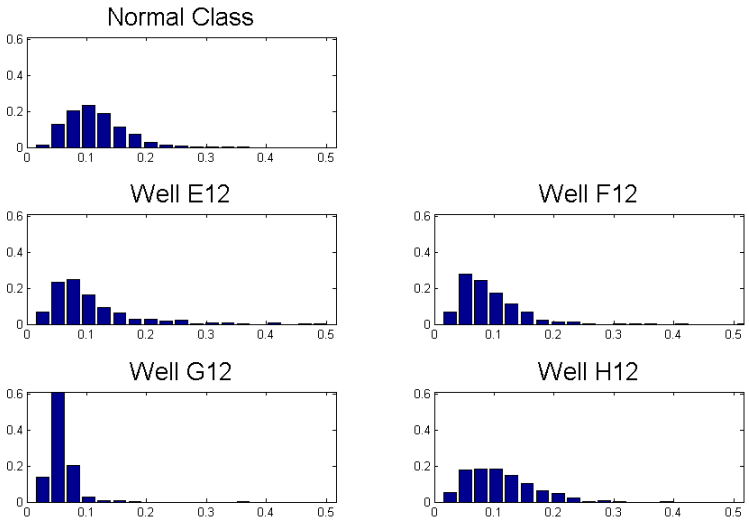


Figure 5.2.27: Histograms of feature $width80mode$

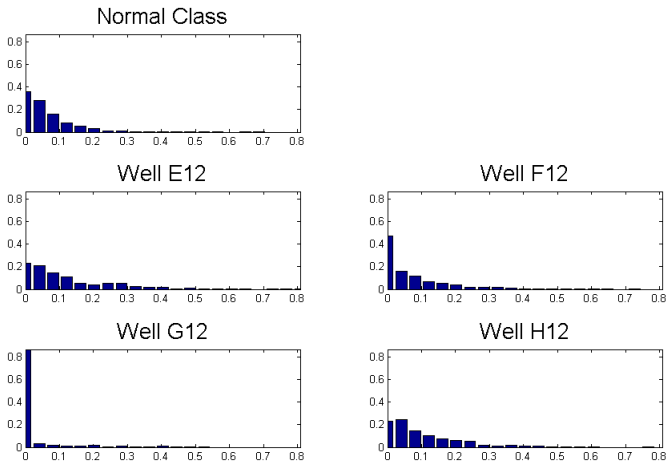


Figure 5.2.28: Histograms of feature $ratioLine$

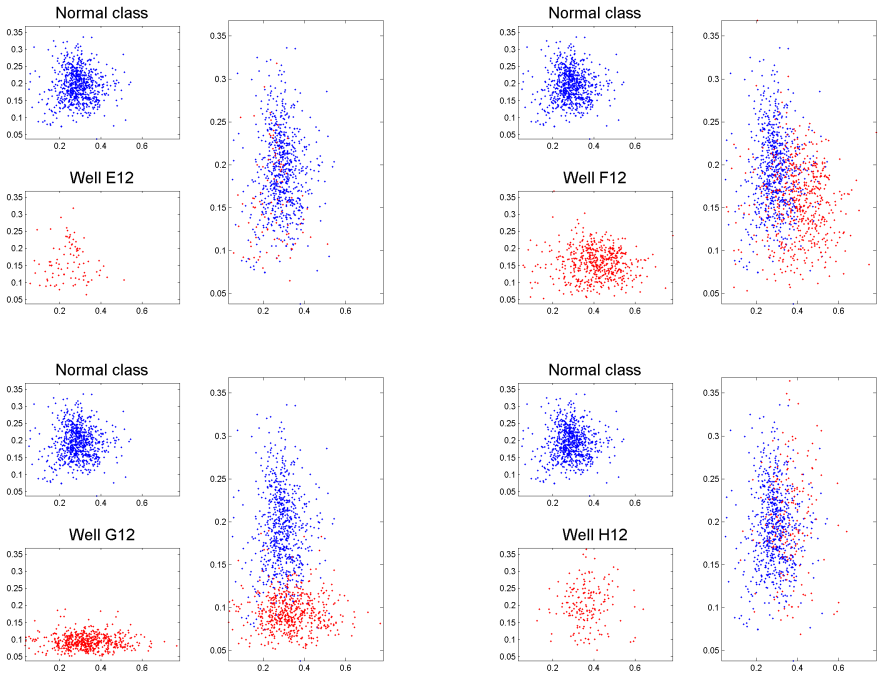


Figure 5.2.29: $normVal$ of $maxH$ vs. Width50max for lines with $numOsc = 1$

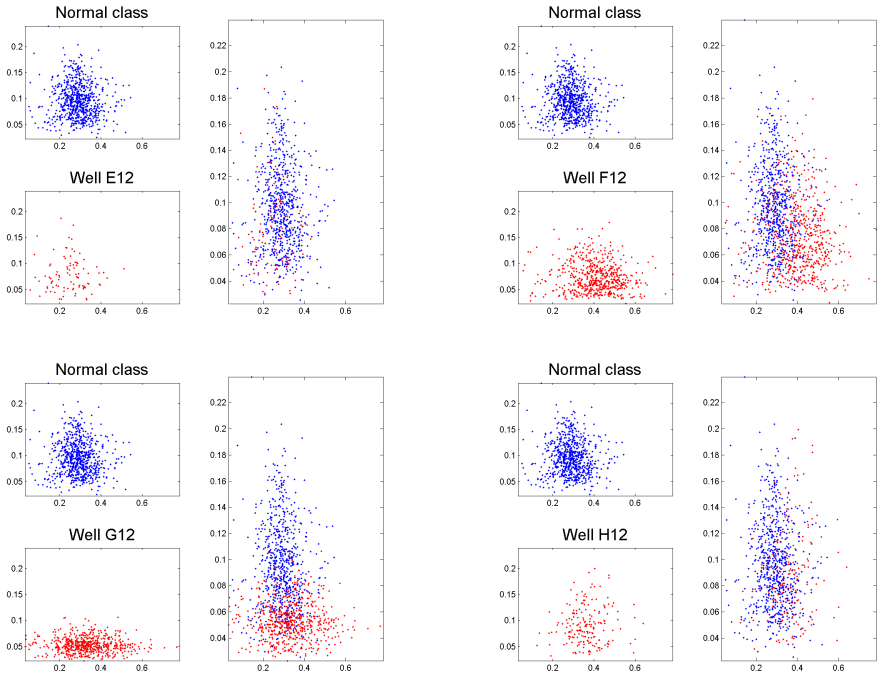


Figure 5.2.30: $normVal$ of $maxH$ vs. $width80max$ for lines with $numOsc = 1$

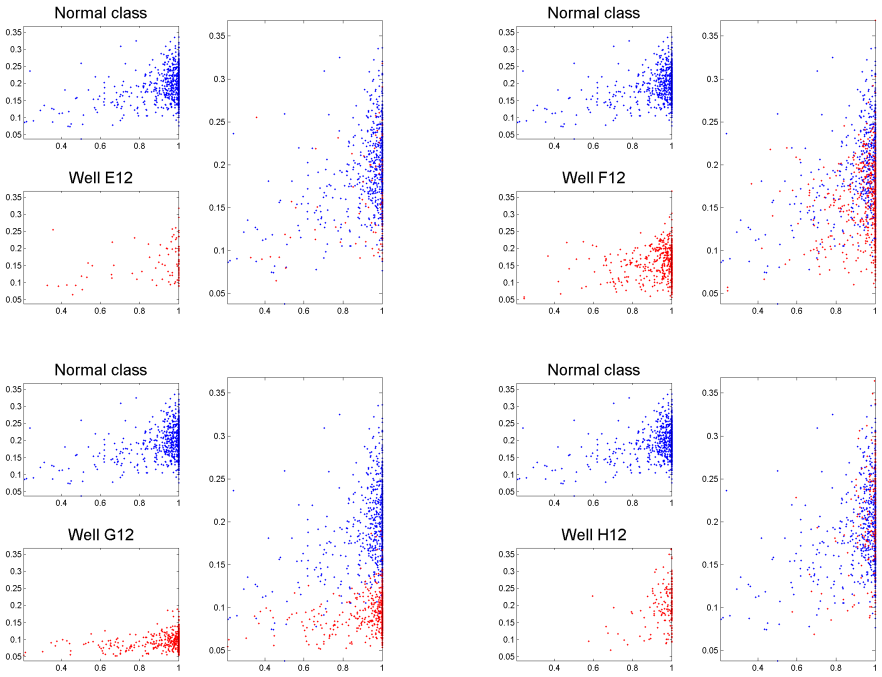


Figure 5.2.31: $gammaVal$ of $maxH$ vs.$width50max$ for lines with $numOsc = 1$

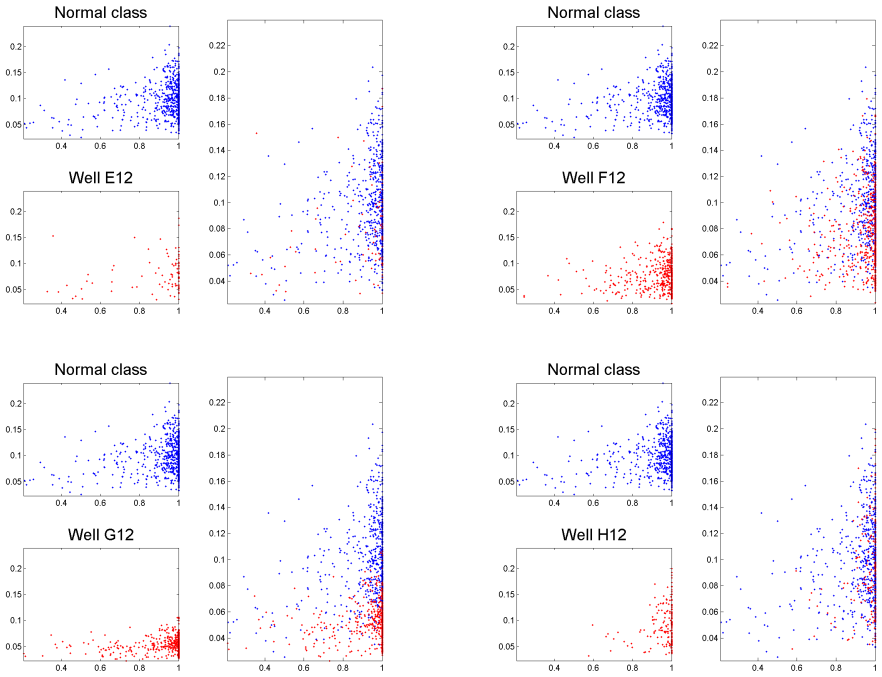


Figure 5.2.32: $gammaVal$ of $maxH$ vs. $width80max$ for lines with $numOsc = 1$

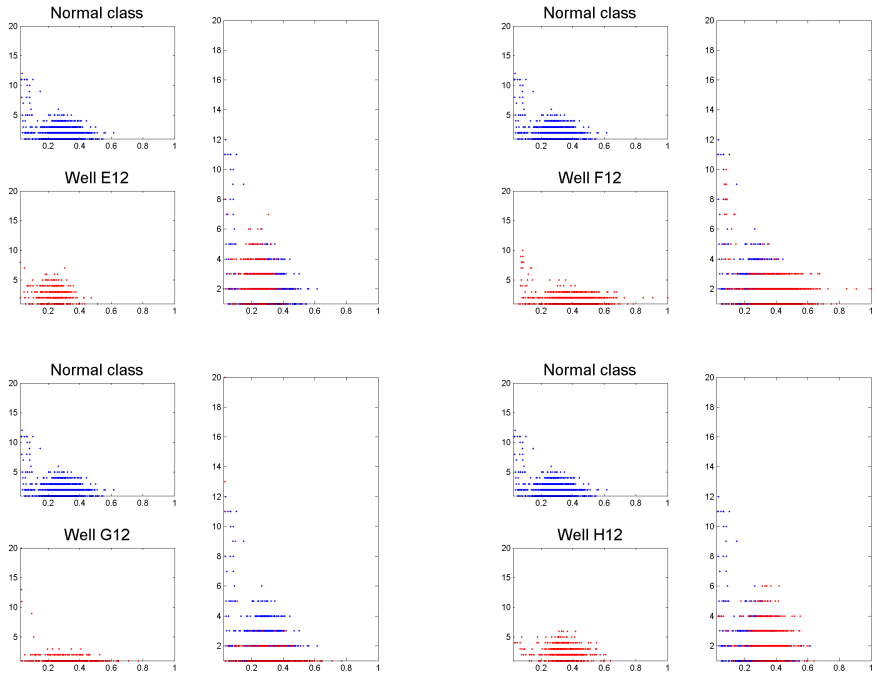


Figure 5.2.33: $normVal$ of $maxH$ vs. $numOsc$

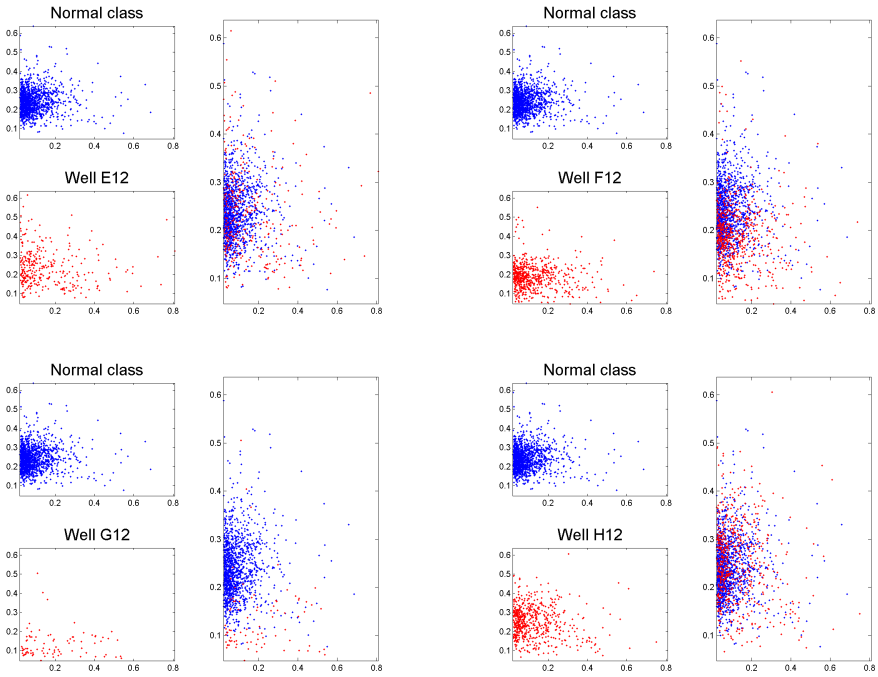


Figure 5.2.34: $ratioLine$ vs. $width50max$

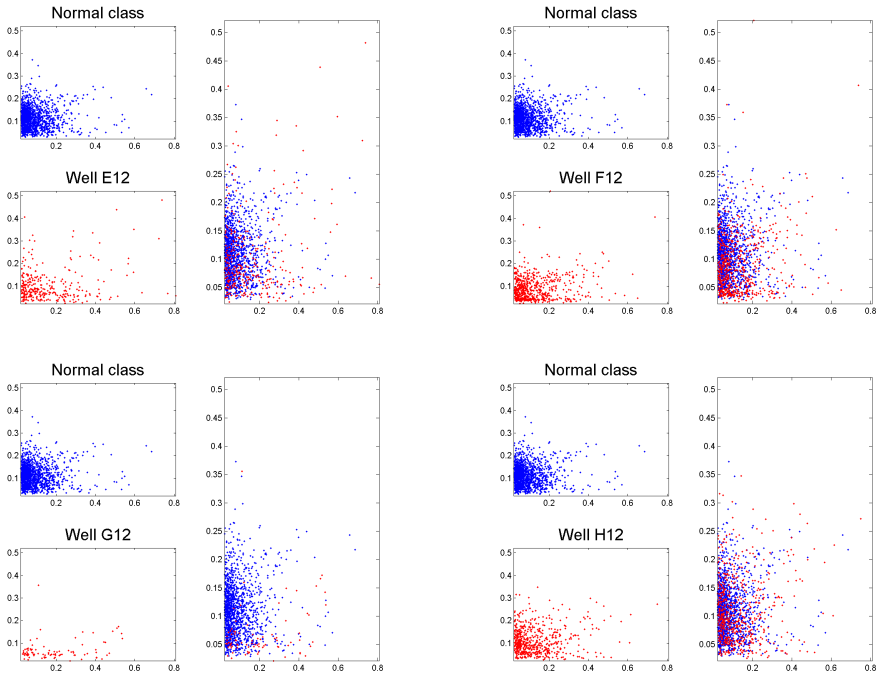


Figure 5.2.35: $ratioLine$ vs. $width80max$

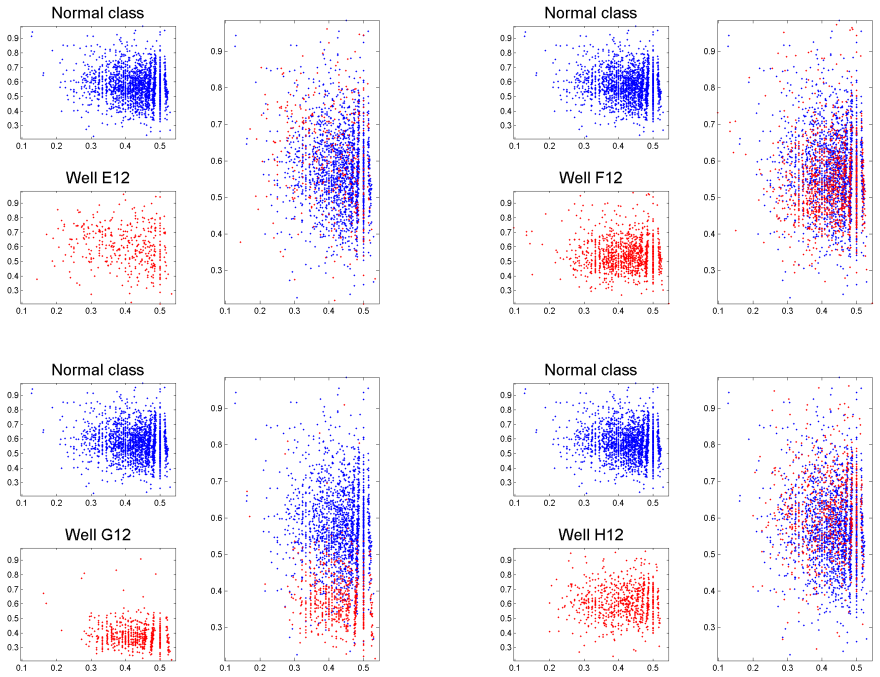


Figure 5.2.36: $position$ of $maxH$ vs. $lengthLine$

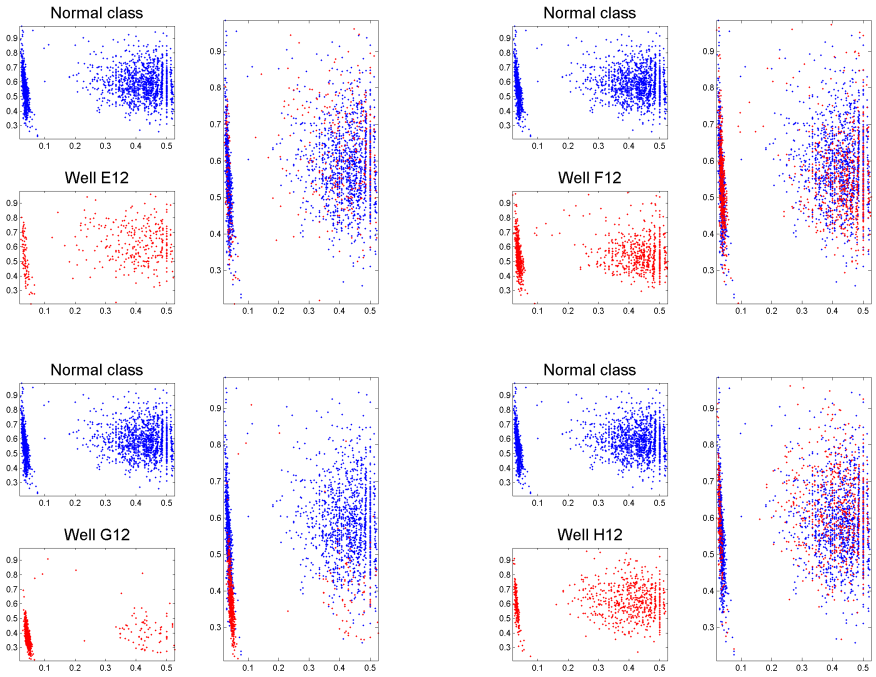


Figure 5.2.37: *position* of *absMin* vs. *lengthLine*

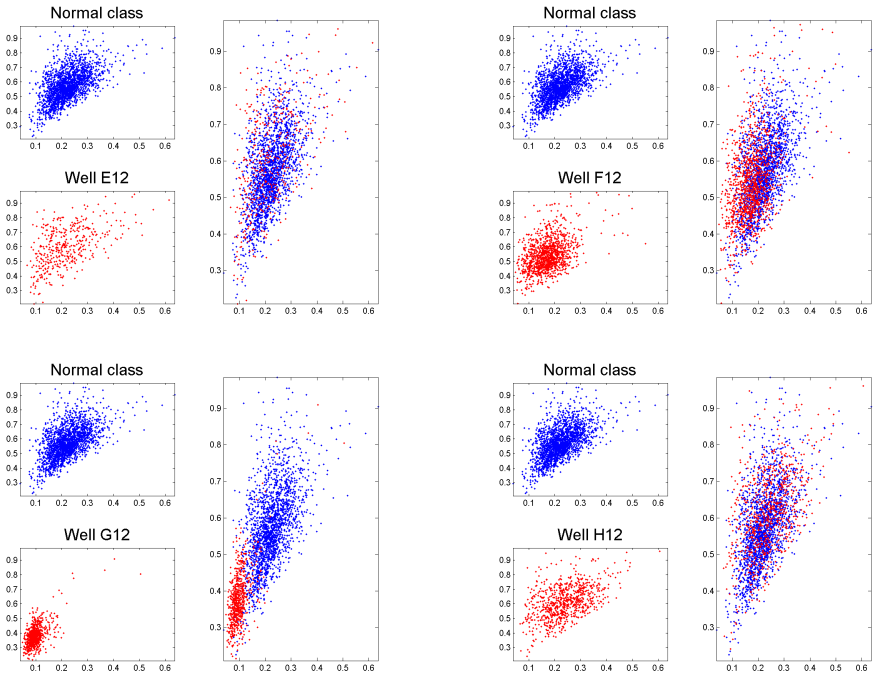


Figure 5.2.38: *width*50*max* vs. *lengthLine*

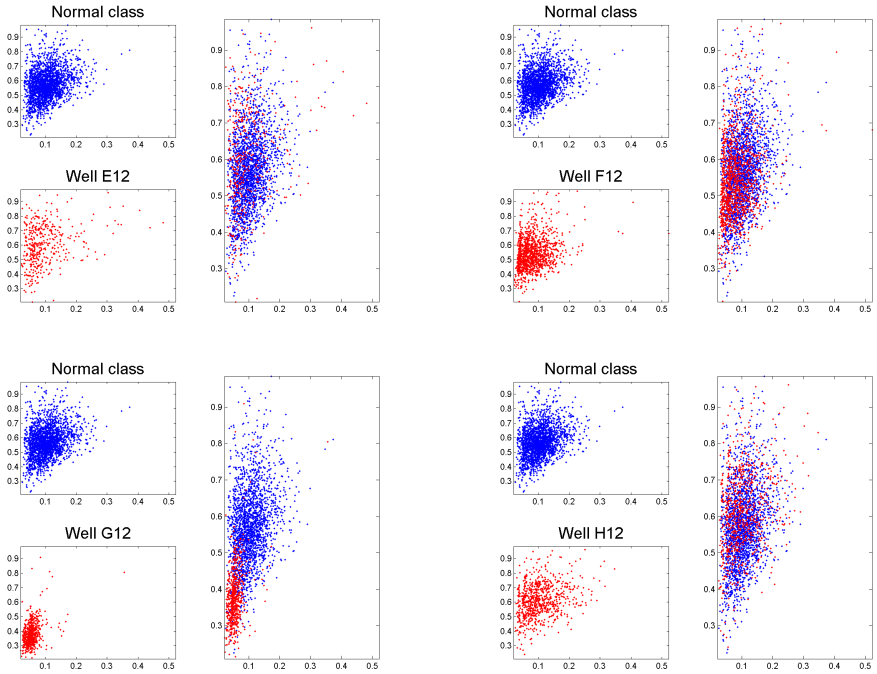


Figure 5.2.39: $width80max$ vs. $lengthLine$

EMD	
Well A12 versus Well B12	51.2500
Normal class versus Well E12	112.0278
Normal class versus Well F12	68.8333
Normal class versus Well G12	82.4167
Normal class versus Well H12	51.2500

Table 5.13: EMD values for feature $numRegions$

$intensity_{region}$	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
maximum	22.9286	34.2381	37.9190	42.0619	30.4524
minimum	16.8905	**36.0143**	**31.7810**	23.0952	23.7000
mean	21.1667	**44.5095**	38.0667	37.3714	28.9571

Table 5.14: EMD values for feature $intensity_{regions}$

5.2.3 Edge detection

For feature $numRegions$ we obtain the same results than for mean shift $numModes$. EMD values are presented in Tab. 5.13 and their histograms in Fig. 5.2.40.

EMD values for feature $intensity_{regions}$ are presented in Tab. 5.14 and their histograms in Fig. 5.2.41, Fig. 5.2.42 and Fig. 5.2.43. This feature is quite discriminant with wells E12 and F12, chiefly with the histograms of the minimum and the mean. Despite EMD values do not reflect it, well G12 is also discriminated for this feature, concretely with the minimum histogram. Also well H12 presents an interesting result in the minimum and mean histograms.

For feature $distance_{regions}$, we obtain wide histograms, which lead to high EMD values although some of those histograms are not as discriminant as the EMD value suggests. EMD values are expounded in Tab. 5.15 and their histograms are found in Fig. 5.2.44 and Fig. 5.2.45. We can observe a good discrimination of well G12 and also a slight discrimination of wells E12 and H12.

$distance_{regions}$	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
maximum	27.4333	**62.9905**	39.0048	**47.7095**	**53.6571**
minimum	24.5619	**60.8333**	33.2238	**54.4048**	**49.4905**

Table 5.15: EMD values for feature $distance_{regions}$

$distanceCentroid_{regions}$	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
maximum	8.5048	**20.9619**	11.3952	**21.1857**	**16.0905**
second maximum	7.6048	**19.4831**	10.6619	**19.8381**	14.5905
mean	9.0095	**22.3381**	11.5095	**23.2571**	**17.4905**

Table 5.16: EMD values for feature $distanceCentroid_{regions}$

Feature $distanceCentroid_{regions}$ end up being a quite discriminant feature for wells E12 and G12. EMD values are shown in Tab. 5.16 and their respective histograms in Fig. 5.2.46, Fig. 5.2.47 and Fig. 5.2.48. Both features $distance_{regions}$ and $distanceCentroid_{regions}$ are normalized by the major axis of the nucleus.

Finally, we proceed to assess shape features from edge detection. A table collecting all the EMD values is presented in Tab. 5.17 and the respective histograms are found in Fig. 5.2.49, Fig. 5.2.50, Fig. 5.2.51, Fig. 5.2.52, Fig. 5.2.53, Fig. 5.2.54, Fig. 5.2.55, Fig. 5.2.56, Fig. 5.2.57, Fig. 5.2.58 and Fig. 5.2.59. We find out that wells E12 and G12 are the main discriminated wells with the shape features. In some cases, also H12 have good results like in $MA_{regions}$ or the ratio $areasR$ where it is evidenced the existence of the largest spot in H12. Feature $MA_{regions}$ is normalized by the major axis of the nucleus.

At last, the comparison between $numRegions$ and $numModes$ is carried out with the two variants of mean shift configurations. EMD values are found in Tab. 5.18 and their histograms in Fig. 5.2.60 for v1 and Fig. 5.2.61 for v2. In G12 histograms we always find negative values because there are more modes in edge detection than in mean shift in both configurations. This is because it is difficult to mean shift to find the small and heterogeneous spots in G12. Due to the unfolding phenotype, in the histograms of E12 we have positive values with v2 (bw_{MS} = 10) and negative values with v1 (bw_{MS} = 6) because with bw_{MS} = 10 we find more values than with edge detection while with bw_{MS} = 6 we find less values, being edge detection the midpoint.

$MA_{regions}$	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
maximum	21.3952	**50.6571**	34.1286	**55.5429**	**36.0190**
minimum	12.8476	**31.5810**	19.6095	**33.8429**	22.5670
mean	16.5619	**40.3381**	24.5571	**43.8143**	29.5048
biggest	21.1667	**50.3667**	33.9905	**55.0571**	**35.6048**

$ecc_{regions}$	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
maximum	43.7048	**103.6857**	58.0095	92.9190	79.1095
minimum	28.9476	**68.4667**	39.9095	**75.3619**	**50.1048**
biggest	39.3286	**93.7857**	59.0476	**84.4048**	71.2000

$R_{regions}$	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
maximum	54.2143	**131.9524**	68.0571	105.4333	99.8857
minimum	51.1762	**125.4143**	62.3286	101.7095	92.8286
biggest	48.6714	**119.3952**	59.5333	96.3952	88.7762

	$areasR$
Well A12 versus Well B12	26.3333
Normal class versus Well E12	**68.3810**
Normal class versus Well F12	27.6190
Normal class versus Well G12	**56.1952**
Normal class versus Well H12	**46.4286**

Table 5.17: EMD values for shape features in edge detection

Variants	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
v1	88.0222	**211.5778**	113.0444	**198.4667**	158.8222
v2	82.9455	**201.6182**	105.6909	**181.3091**	147.9091

Table 5.18: EMD values for the comparison of $numRegions$ with $numModes$

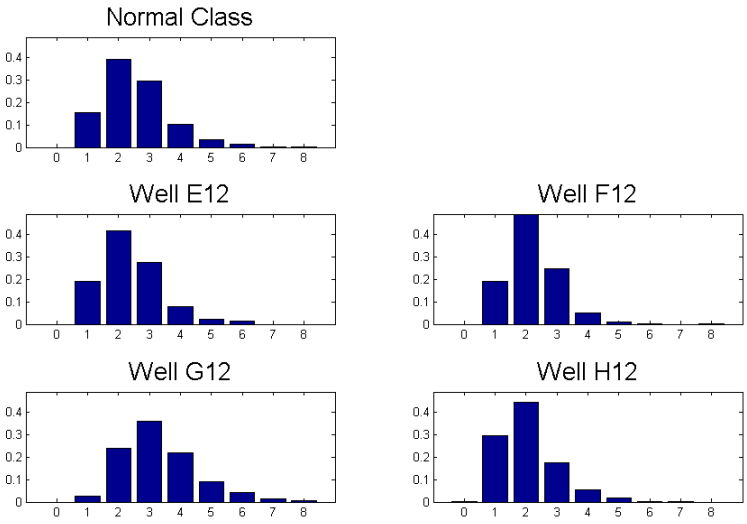


Figure 5.2.40: Histograms for feature $numRegions$

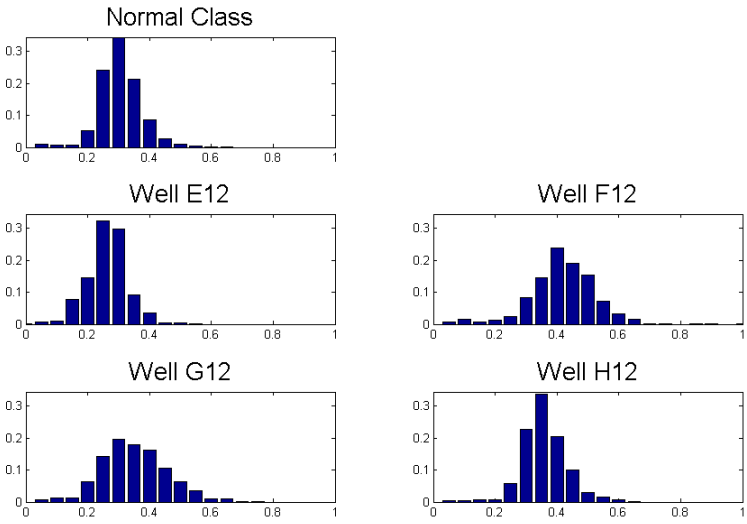


Figure 5.2.41: Histograms for maximum of feature $intensity_{regions}$

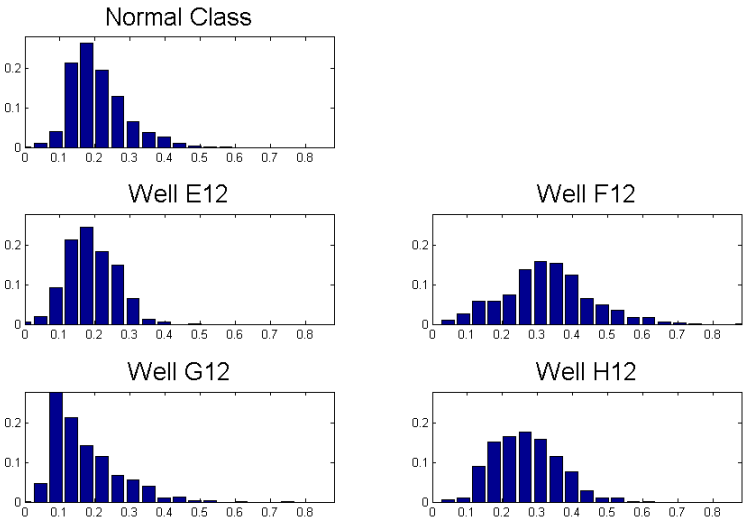


Figure 5.2.42: Histograms for minimum of feature $intensity_{regions}$

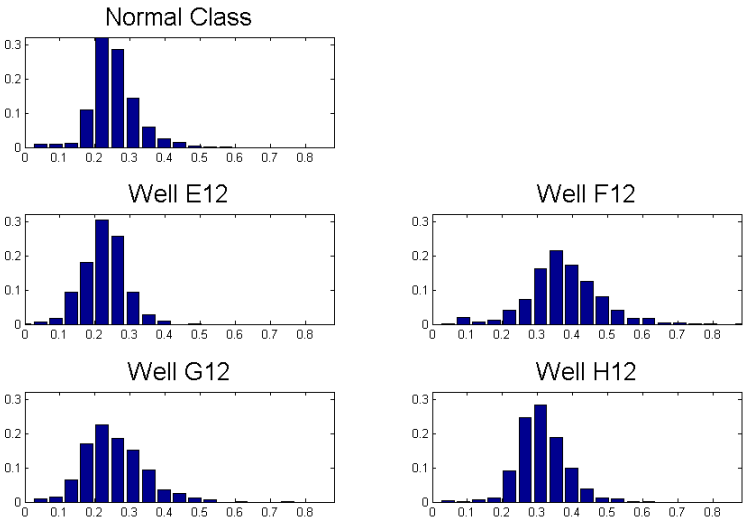


Figure 5.2.43: Histograms for mean of feature $intensity_{regions}$

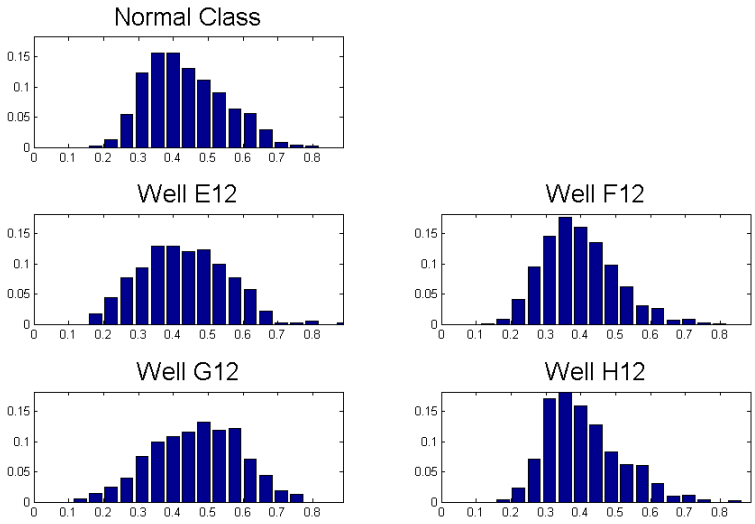


Figure 5.2.44: Histograms for maximum of feature $distance_{regions}$

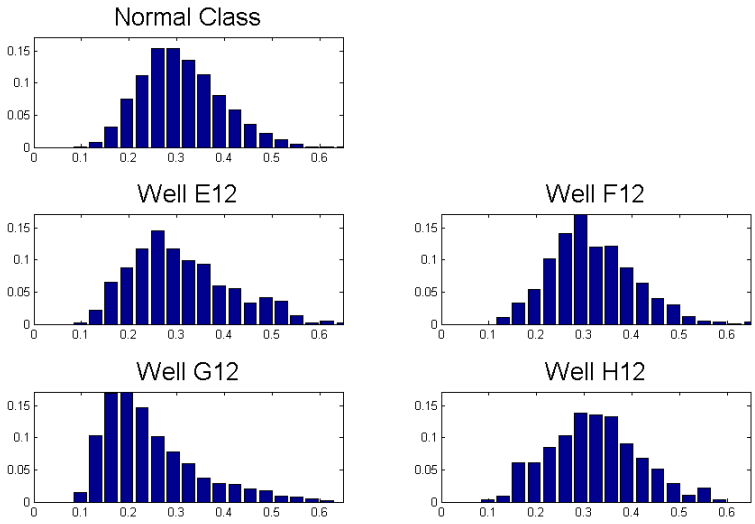


Figure 5.2.45: Histograms for minimum of feature $distance_{regions}$

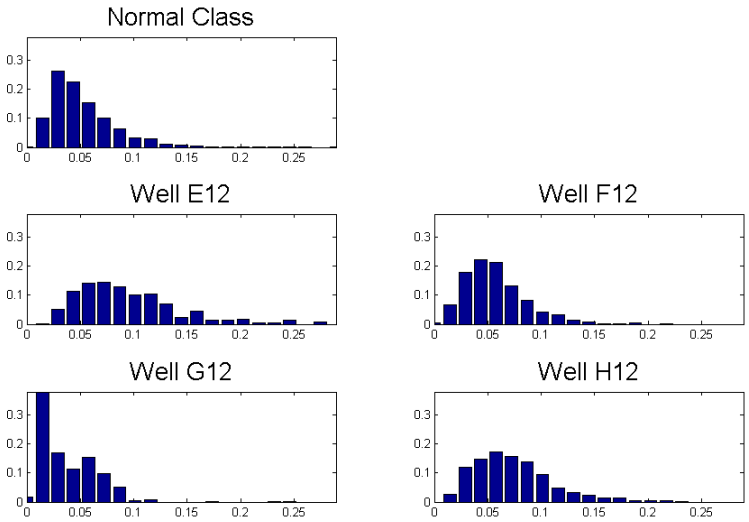


Figure 5.2.46: Histograms for maximum of feature $distanceCentroid_{regions}$

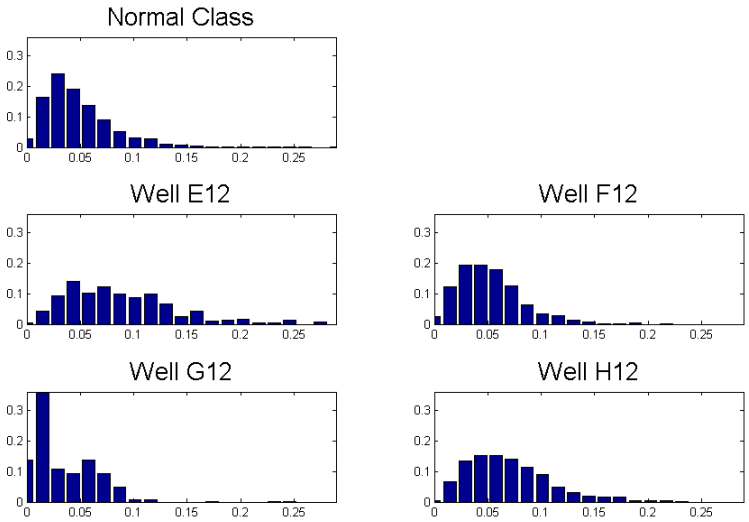


Figure 5.2.47: Histograms for second maximum of feature $distanceCentroid_{regions}$

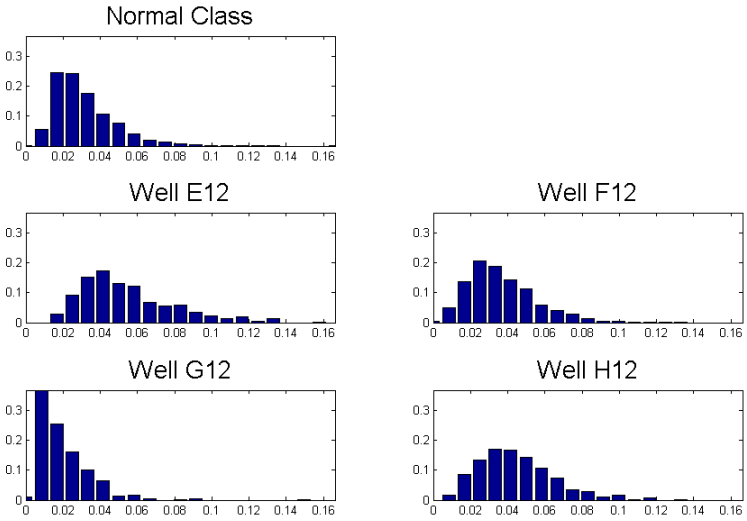


Figure 5.2.48: Histograms for mean of feature $distanceCentroid_{regions}$

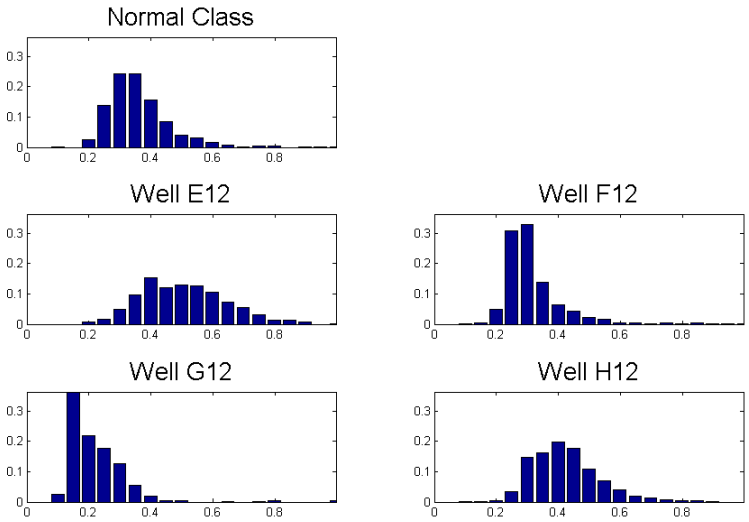


Figure 5.2.49: Histograms for maximum of $MA_{regions}$

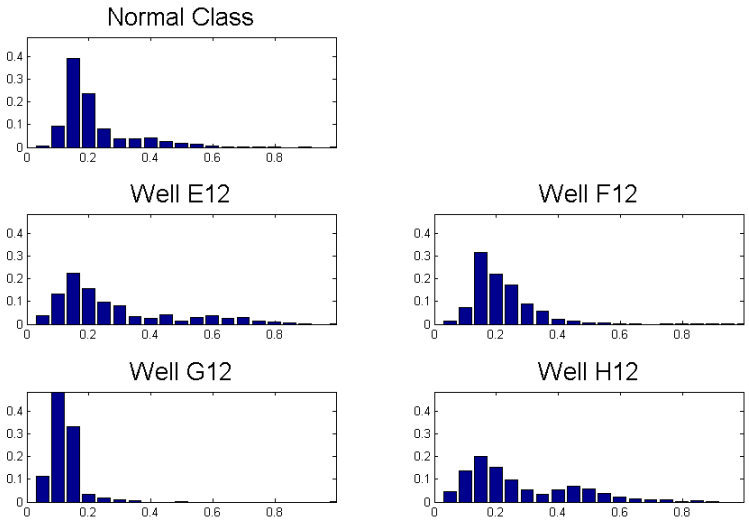


Figure 5.2.50: Histograms for minimum of $MA_{regions}$

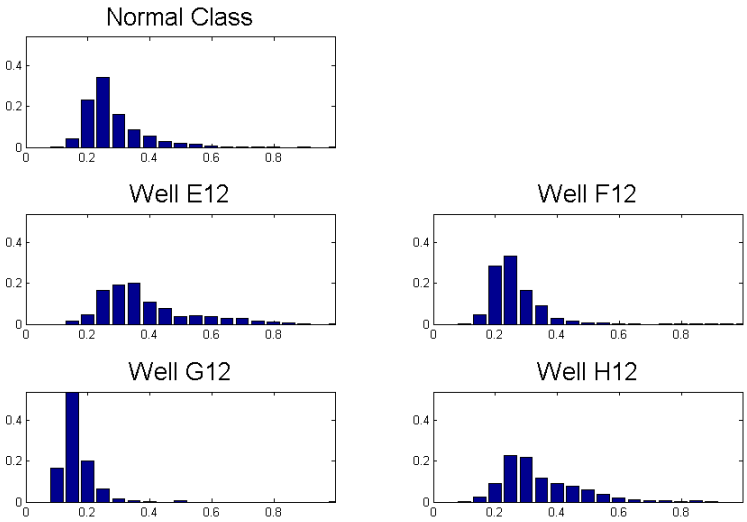


Figure 5.2.51: Histograms for mean of $MA_{regions}$

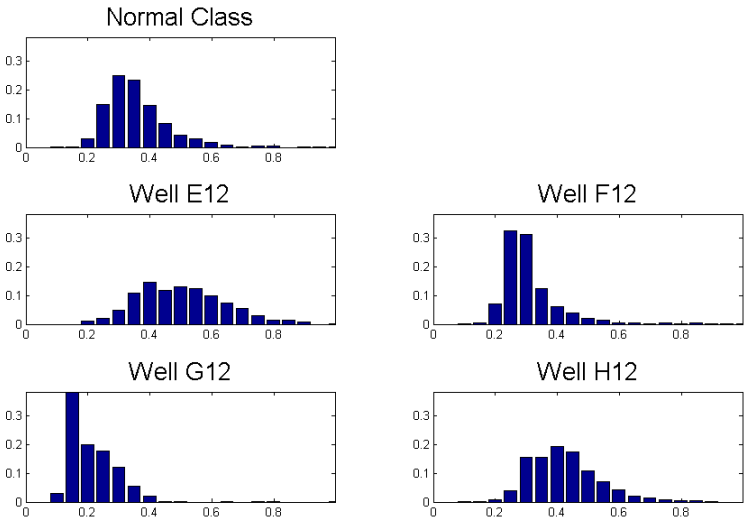


Figure 5.2.52: Histograms of $MA_{regions}$ for the biggest region in the nucleolus

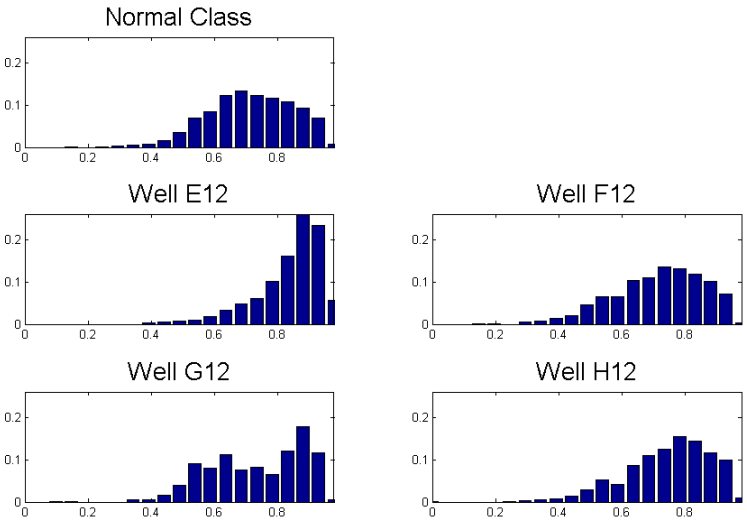


Figure 5.2.53: Histograms for maximum of $ecc_{regions}$

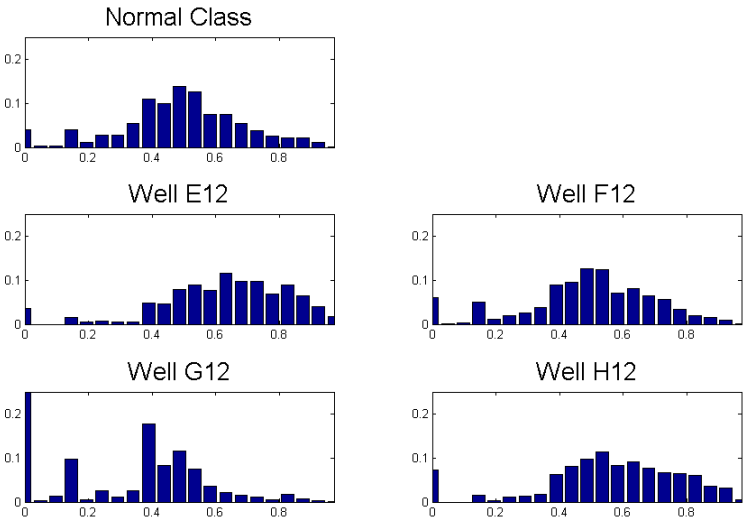


Figure 5.2.54: Histograms for minimum of $ecc_{regions}$

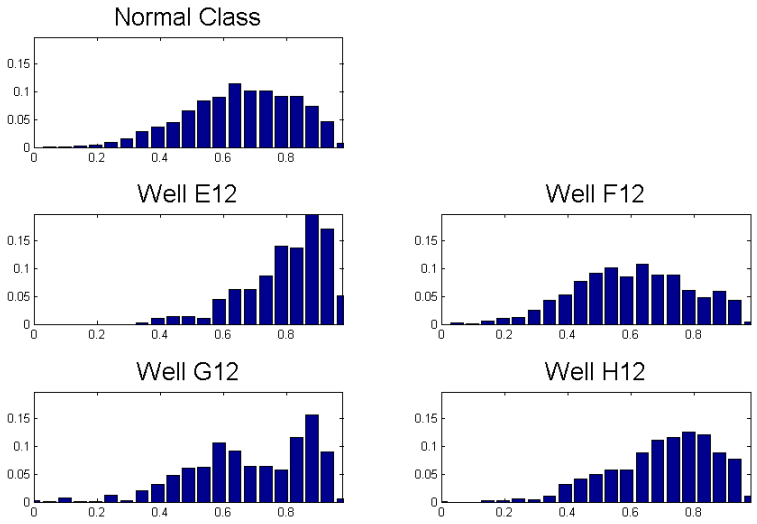


Figure 5.2.55: Histograms of $ecc_{regions}$ for the biggest region in the nucleolus

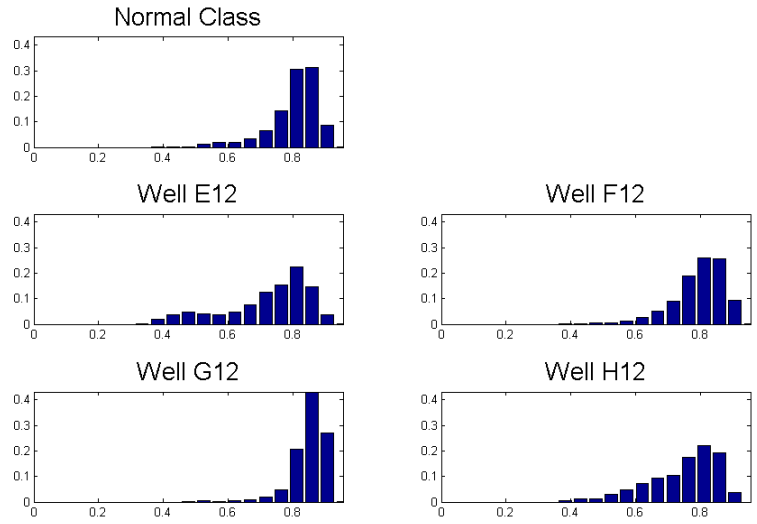


Figure 5.2.56: Histograms for maximum of $R_{regions}$

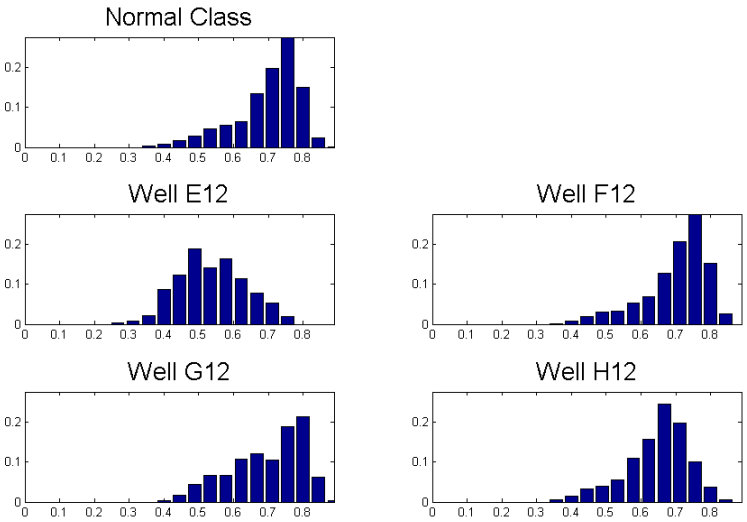


Figure 5.2.57: Histograms for minimum of $R_{regions}$

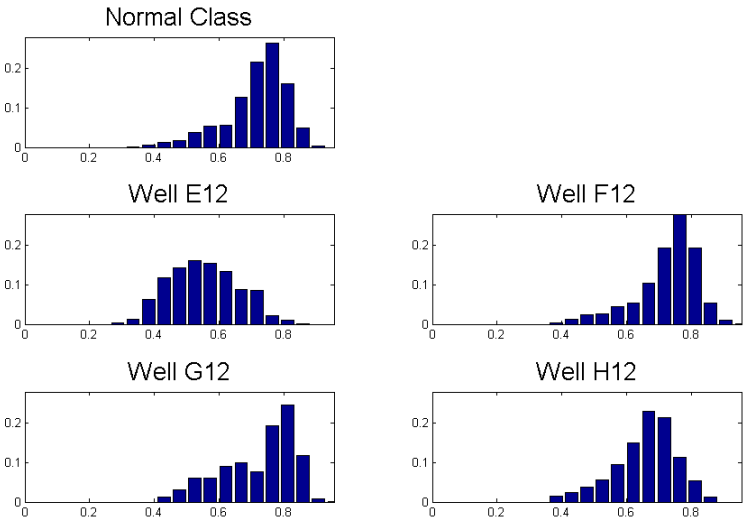


Figure 5.2.58: Histograms of $R_{regions}$ for the biggest region in the nucleolus

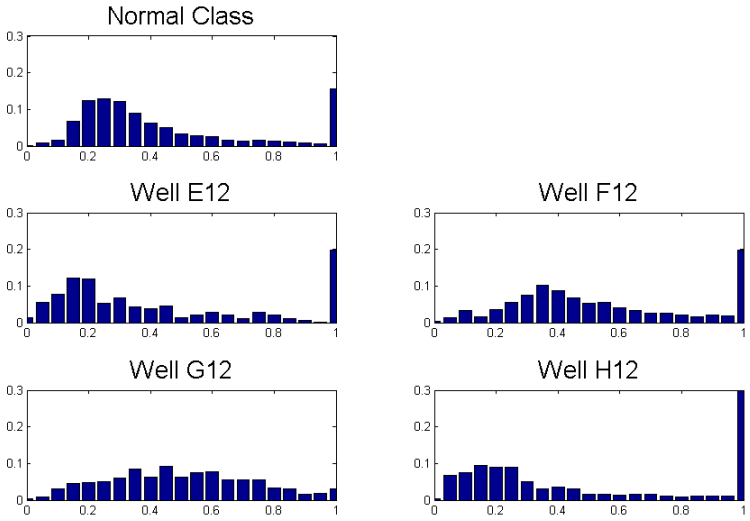


Figure 5.2.59: Histograms for feature $areasR$

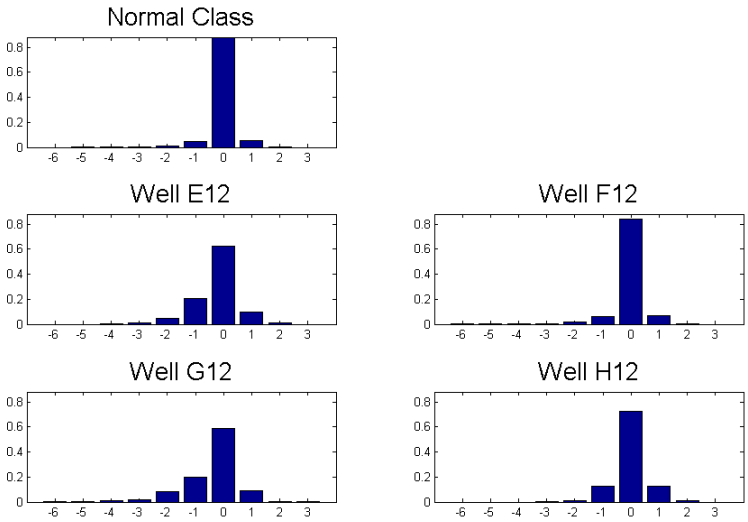


Figure 5.2.60: Histograms for the comparison of *numRegions* with *numModes* in v1

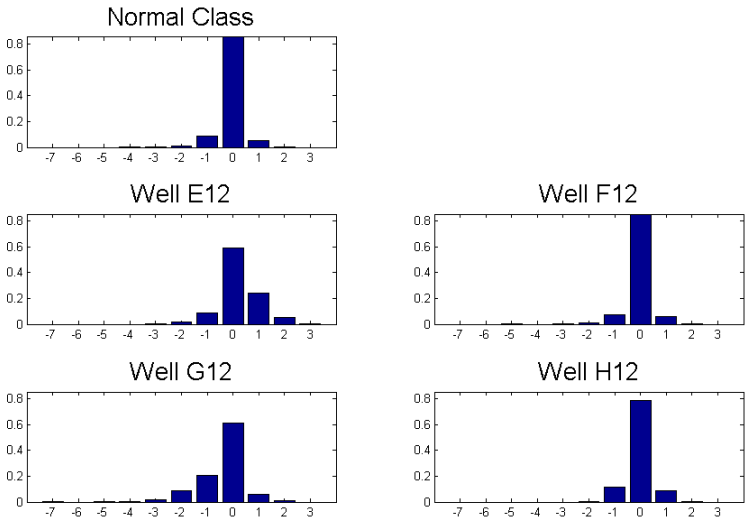


Figure 5.2.61: Histograms for the comparison of *numRegions* with *numModes* in v2

Norm image					
Threshold	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
mean	23.7429	**54.0190**	35.6429	**53.9762**	41.5667
0.05	26.2048	**55.5762**	**63.0762**	**59.3381**	**52.5810**
0.1	14.7762	**35.6714**	23.0000	**37.7190**	**31.9952**
Gamma image					
Threshold	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
mean	24.3190	**57.7667**	38.7619	**58.8571**	**46.4476**
0.05	18.8190	**42.3905**	30.1952	**48.4095**	**39.7095**
0.1	14.3762	**33.8810**	23.8762	**38.3429**	29.0048

Table 5.19: EMD values for feature $OccupationTH$

EMD EMD values	
Well A12 versus Well B12	10.8095
Normal class versus Well E12	26.8571
Normal class versus Well F12	18.7381
Normal class versus Well G12	**28.0857**
Normal class versus Well H12	21.3476

Table 5.20: EMD values for feature $OccupationED$

5.2.4 Area ratios

We have tested the feature $OccupationTH$ with different fixed thresholding values from 0.025 to 0.80 and also with an adjustable value that changes for each cell: the mean intensity in the nucleolus image within the nucleus area. We have tested it with norm image and gamma image with a gamma value of 1.5. EMD values for each well are listed in Tab. 5.19. Histograms are presented in Fig. 5.2.63, Fig. 5.2.64, Fig. 5.2.62, Fig. 5.2.66, Fig. 5.2.67 and Fig. 5.2.65. As you will notice, best histograms are obtained for well G12, which is highly discriminated by different thresholds. Wells E12 and H12 can also be discriminated by this feature, specially with norm image and a threshold on 0.05.

In Tab. 5.20 we present the EMD values obtained for the feature $OccupationED$, which are not very hopeful with the exception of well G12.

Feature $OccupationRegion$ end up being the weaker area ratio leading to very wide

Nucleolus	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
maximum	41.0190	98.4095	51.8810	**99.5571**	66.7000
ratio	26.3333	**68.3810**	27.7095	**56.1857**	**46.4286**

Nucleus	A12 vs. B12	N vs. E12	N vs. F12	N vs. G12	N vs. H12
maximum	7.3238	**17.7095**	12.2429	**19.7143**	15.4000
ratio	26.3333	**68.3810**	27.6190	**56.1905**	**46.4286**

Table 5.21: EMD values for feature $OccupationRegion$ in the nucleolus and in the nucleus

histograms with low values in each bin. However, we can extract some information from the histograms specially about the behavior of H12 in the ratio of $OccupationRegion$, which histogram is accumulated at the ends, evidencing the existence of the large spot that we have mentioned several times.

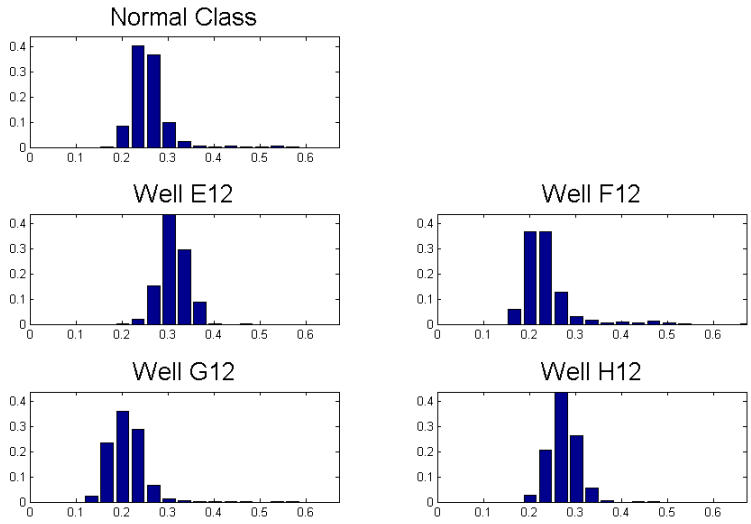


Figure 5.2.62: Histograms for feature *occupationTH* over norm image with threshold = mean

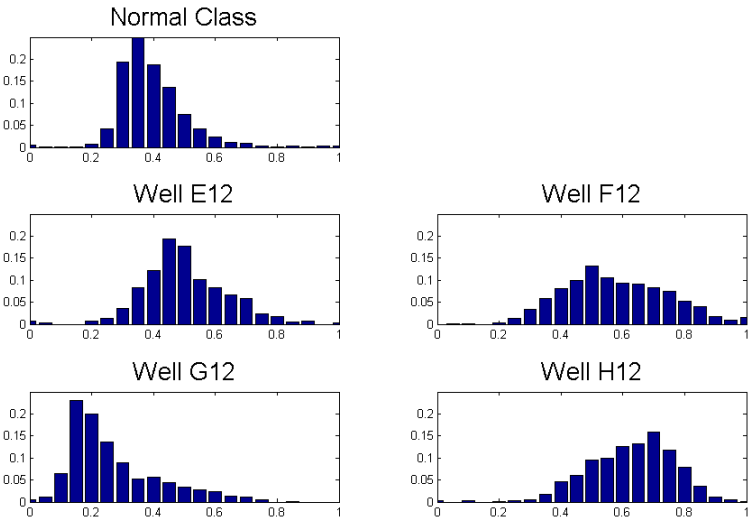


Figure 5.2.63: Histograms for feature *occupationTH* over norm image with threshold = 0.05

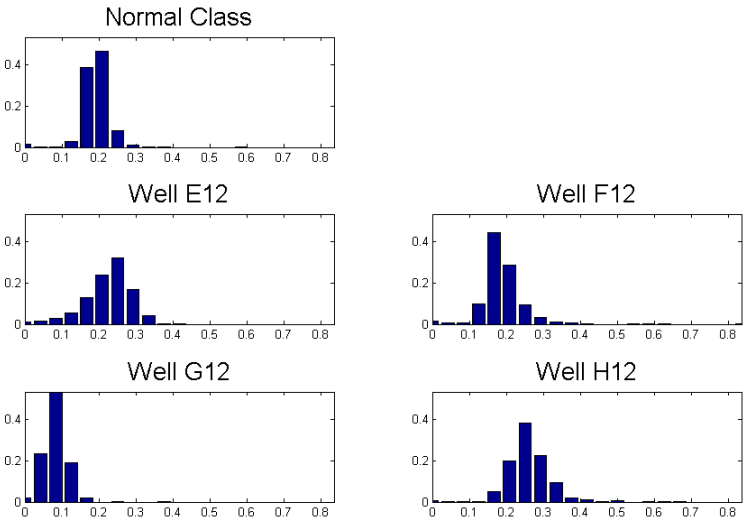


Figure 5.2.64: Histograms for feature *occupationTH* over norm image with threshold = 0.1

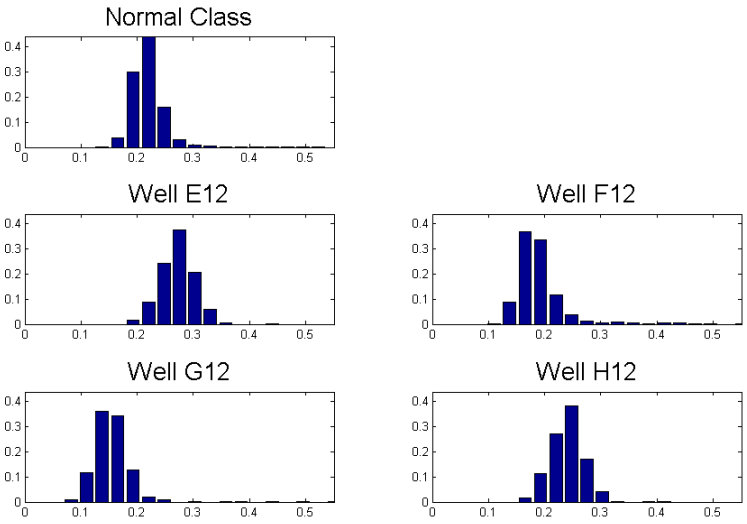


Figure 5.2.65: Histograms for feature *occupationTH* over gamma image with threshold = mean

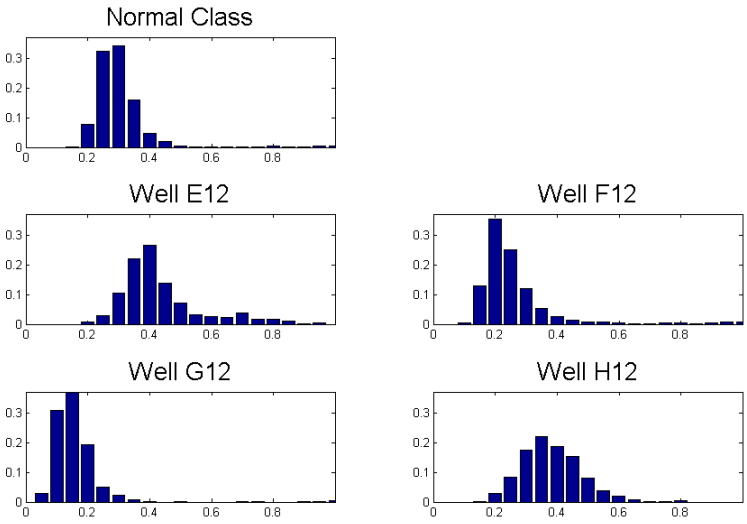


Figure 5.2.66: Histograms for feature *occupationTH* over gamma image with threshold = 0.05

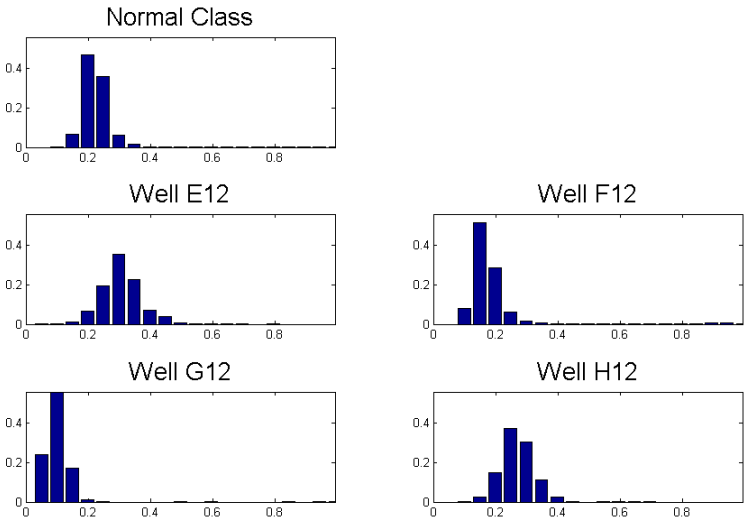


Figure 5.2.67: Histograms for feature *occupationTH* over gamma image with threshold = 0.1

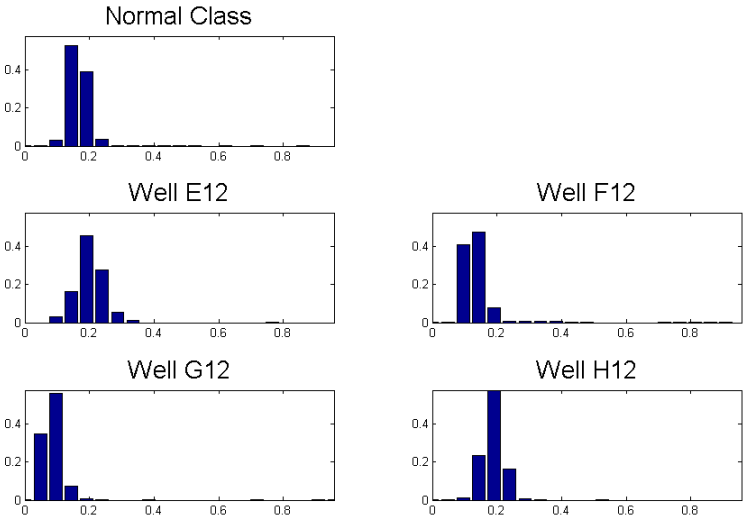


Figure 5.2.68: Histograms for feature *OccupationED*

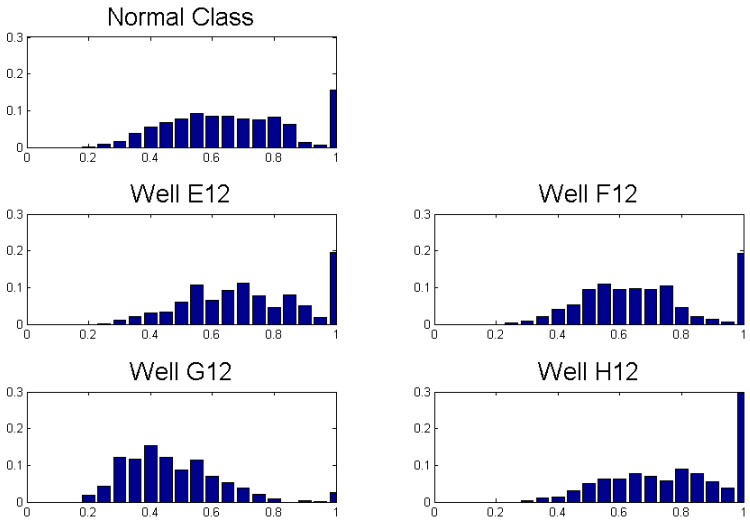


Figure 5.2.69: Histograms for maximum of *OccupationRegion* within the nucleolus

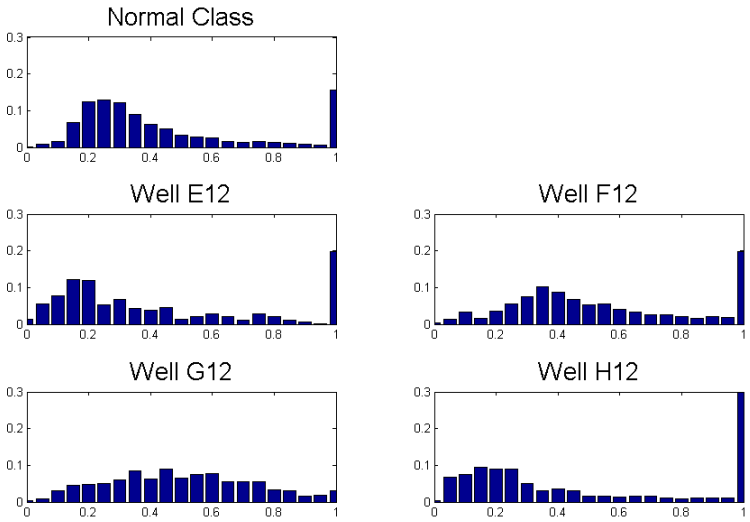


Figure 5.2.70: Histograms for the ratio of *OccupationRegion* within the nucleolus

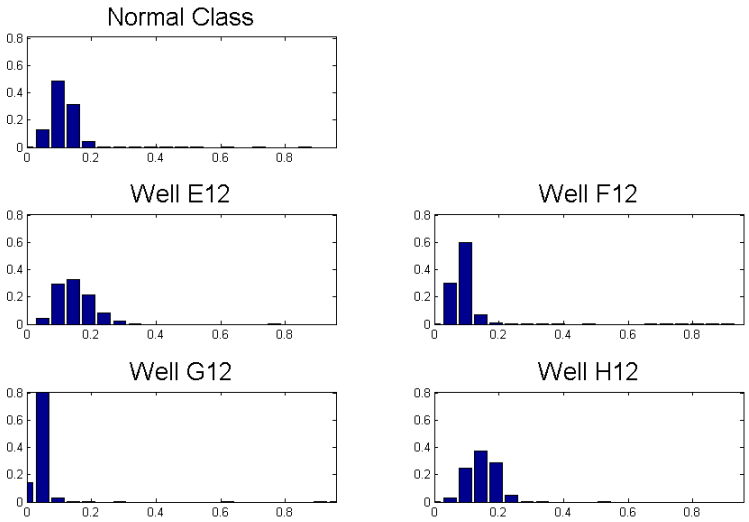


Figure 5.2.71: Histograms for maximum of *OccupationRegion* within the nucleus

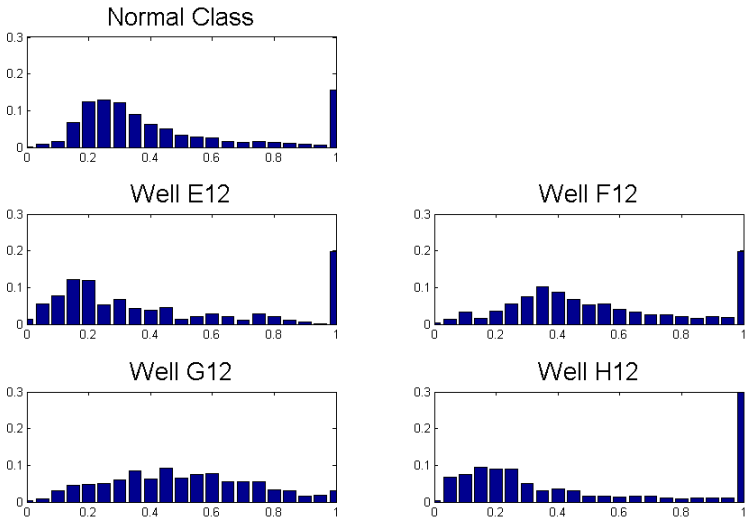


Figure 5.2.72: Histograms for ratio of *OccupationRegion* within the nucleus

6 Future work

On the on hand, we still have some ideas to continue on the same path and sharpen the profiles of certain wells. On the other hand, we also have a new proposal to look at completely different features.

The first block is make up by the following proposals:

- Applying lines to edge detection, which will guarantee the correct cutting of the line before the feature extraction. As have been mentioned, edge detection finds spreading like in Fig. 4.2.5 in regions affected by unfolding phenotype, which leave different possibilities of measuring in lines. For example, these lines will be dashed lines most probably and some of them will have modes placed outside the region.
- Circles: we also have considered the possibility of applying concentric circles centered on the modes from mean shift or edge detection and repeat the same procedure as in lines, by comparing the circles displaying their intensities. The main idea is that concentric circles in a typical normal mode will be similar while concentric circles in a typical abnormal mode will be quite different. The purpose is to measure those differences to discriminate normal wells from abnormal wells.

Finally, next step consists on the classification of the phenotypes based on the proposed features to clusterize the wells in different groups, based on some measures of resemblance. The more varied the set of features is, the more exhaustive and accurate the classification is, due to classification is based on a metric of similarity between samples calculated from the set of features. Thus two samples from abnormal cultures with similar features will have a similar metric and will be assigned to the same cluster while two samples from cultures with dissimilar metric will belong to different clusters.

In the second block, we propose the implementation of the similarity measure for texture images based on the Gray Level Aura Matrices (GLAM) defined in [13], where texture features are represented by Aura Matrices, which have been introduced in Section 4.2.6, calculated at multiresolutions of images. In the proposal, the classification of texture images is done by similarity learning implemented with the Support Vector Machine (SVM), which is a statistical learning algorithm of pattern recognition.

7 Conclusions

The main objectives of the thesis have been fulfilled providing an automatic segmentation of the nucleolus and characterizing this segmented nucleolus in terms of mean shift fundamental structures, comprised by modes, clusters and lines; and also in terms of edge detection fundamental structures, comprised by regions and modes. Besides, we have defined a broad and varied range of features and have assessed each feature studying the behavior of each abnormal well (C12, D12, E12, F12, G12, H12) over the particular feature and comparing this behavior with the behavior of the normal class, set up by wells A12 and B12.

Furthermore, we have defined a set of thresholding values that conform the criteria to evaluate whether a feature is discriminant according to a reference value, which is calculated with the Earth Mover's distance and depends on the type of distribution observed and the range of the feature studied.

We have designed a model for typical pathological Lines distinguishing between normal and abnormal typical lines and have defined a 2-D metric to relate features in order to separate the different line models belonging to the same class, which are merged in the histograms unfortunately.

After all, we keep thinking that trying to model nucleoli based on their images extracted from fluorescence microscopy is extremely complicated due to the complexity of those nucleolus and, consequently, the existence of a widespread of profiles within each normal and abnormal wells themselves.

Let's summarize the most significant features that have been tested along the memory. Firstly, the intensity feature allows us to make an early classification, with a doubtless discrimination presented for abnormal wells C12 and D12. Following, we have evaluated mean shift features, including lines, and edge detection features. We have found that shape features from both mean shift and edge detection, together with the width and length values obtained from Lines and the measures of the distances between the physical centers and the modes in edge detection, enable the discrimination against normal class with wells E12 and G12. Adding the results from the area ratios, we can affirm that wells E12 and G12 can be classified with the features proposed in this memory. Wells F12 and H12 present the closer appearance to normal nucleoli of our database leading to a greater difficulty to discriminate them. Even so, we have found some features that present a discrimination more or less competent for these two remaining wells. For example, for well H12, the ratio between the minimum and maximum area of the regions obtained with edge detection reveals the existence of that big spot that we have mentioned several times. Also the occupation ratio with thresholding is

pretty discriminative with well H12. Meanwhile, some measures from the intensity of the modes in edge detection provide high EMD values for F12 as well as some shape features from mean shift and also the occupation ratio with thresholding. At last, the relation between the height and the width or between the width and the length of Line in well F12 show some slightly successful results.

To be honest, we should also mention that there are some disappointing features that have not meet the expectations. For instance, we do not recommend to work with the number of modes because practically any information is obtained whatever the chosen configuration of mean shift, or the $ratioLine$, which is also useless with norm image as much as gamma image.

Bibliography

[1] B. Chazotte, "Labeling Nuclear DNA Using DAPI", Cold Spring Harb Protoc (January 2011) 80–82.
URL http://cshprotocols.cshlp.org/content/2011/1/pdb.prot5556.full.pdf+html

[2] P. H. PhD, B. Bishop, "Automated Tissue Culture Cell Fixation and Staining in Microplates", Bioket Instruments Inc.
URL http://www.biotek.com/assets/tech_resources/Cell_Staining_App_Note.pdfl

[3] Biology online dictionary (2001).
URL http://www.biology-online.org/dictionary/Main_Page

[4] S. Snaar, K. Wiesmeijer, A. G. Jochemsen, H. J. Tanke, R. W. Dirks, "Mutational Analysis of Fibrillarin and Its Mobility in Living Human Cells", J Cel Biol 2000 151 (October 2000) 653–662.
URL http://jcb.rupress.org/content/151/3/653.full.pdf+html

[5] G. J. Phillips, "Green fluorescent protein - a bright idea for the study of bacteria protein localization", FEMS Microbiology Letters 204 (2001) 9–18.
URL http://onlinelibrary.wiley.com/doi/10.1111/j.1574-6968.2001.tb10854.x/pdf

[6] F. Collins, "HeLa Cells: A New Chapter in An Enduring Story", National Institutes of Health (NIH) Director's blog.
URL http://directorsblog.nih.gov/2013/08/07

[7] J. S. Andersen, C. E. Lyon, A. H. Fox, A. K. Leung, Y. W. Lam, H. Steen, M. Mann, A. I. Lamond, "Directed Proteomic Analysis of the Human Nucleolus", Current Biology, Vol.12 (January 2002) 1–11.
URL http://www.lamondlab.com/pubpdf/02/cb02.pdf

[8] T. N. A. at Karolinska Institutet, "Advanced Information on the Nobel Prize in Physiology or Medicine".
URL http://web.archive.org/web/20070102190732/http://nobelprize.org/nobel_prizes/medicine/laureates/2006/adv.pdf

[9] S. E. University, Austin, Texas, Ribosomes, Department of Chemistry and Biochemistry: Course materials.
URL http://www.cs.stedwards.edu/chem/Chemistry/CHEM43/CHEM43/Ribosomes/Ribosome.HTML

[10] Transfection, Promega Corporation: Protocols and Applications Guide.

URL http://www.promega.es/~/media/files/resources/paguide/a4/chap12a4.pdf?la=es-es

[11] GeneCards, Gene summary (2001).
URL http://www.genecards.org/

[12] D. Comaniciu, P. Meer, "Mean shift: A robust approach toward feature space analysis", PAMI (May 2002) 603–619.

[13] X. Qin, Y.-H. Yang, "Similarity Measure and Learning with Gray Level Aura Matrices (GLAM) for Texture Image Retrieval", Computer Vision and Pattern Recognition CVPR, Vol.1 (2004) 326–333.

[14] E. Levina, P. Bickel, "The Earth Mover's Distance is the Mallows Distance: Some Insights from Statistics", International Conference on Computer Vision ICCV, Vol.2 (July 2001) 251–256.
URL http://dept.stat.lsa.umich.edu/~elevina/EMD.pdf

[15] M. Inc., EMD code.
URL http://www.mathworks.com/matlabcentral/fileexchange/22962

[16] M. Inc., Extrema code.
URL http://www.mathworks.com/matlabcentral/fileexchange/12275

Annex

Set of sample nucleolus

We expose a set of random nucleolus images normalized as gamma image from wells A12, B12, E12, F12, G12 and H12.

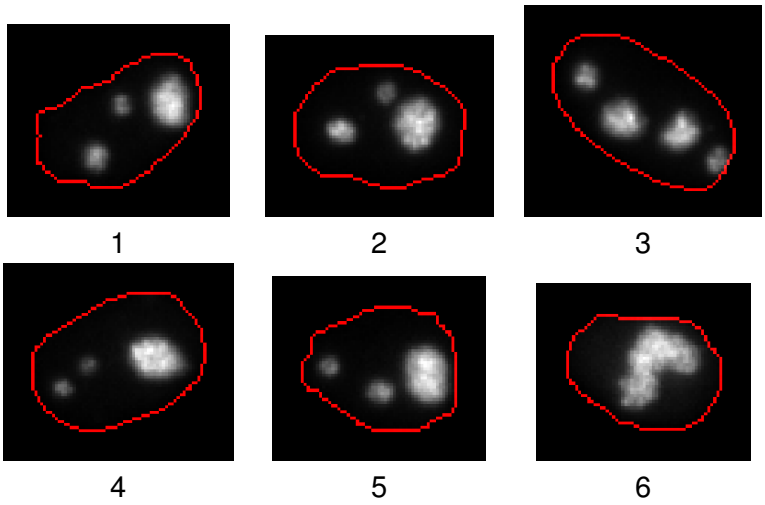


Figure 7.0.1: Well A12

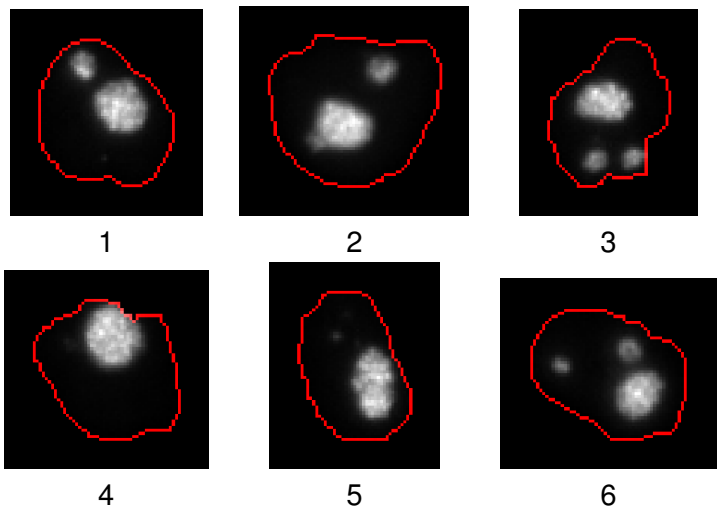


Figure 7.0.2: Well B12

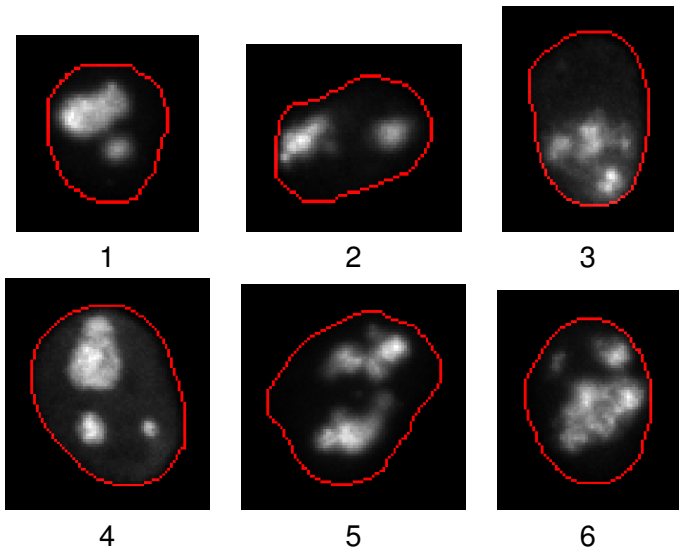


Figure 7.0.3: Well E12

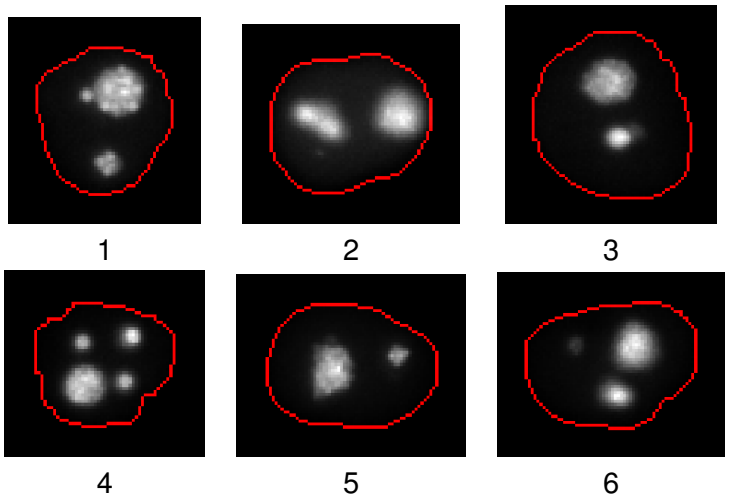


Figure 7.0.4: Well F12

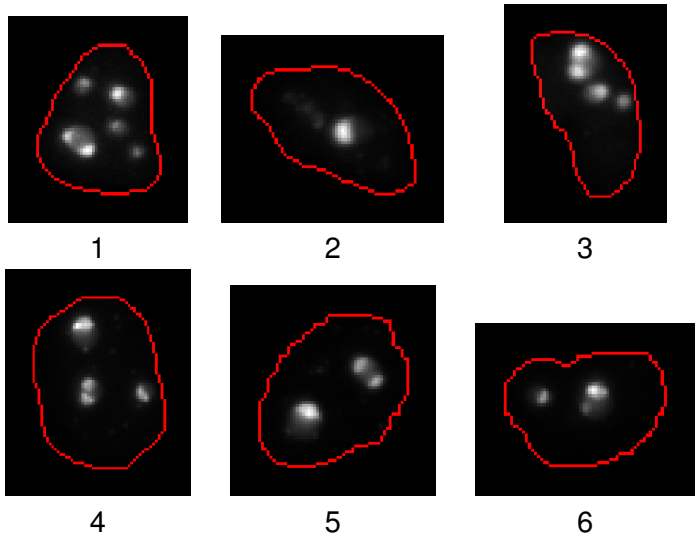


Figure 7.0.5: Well G12

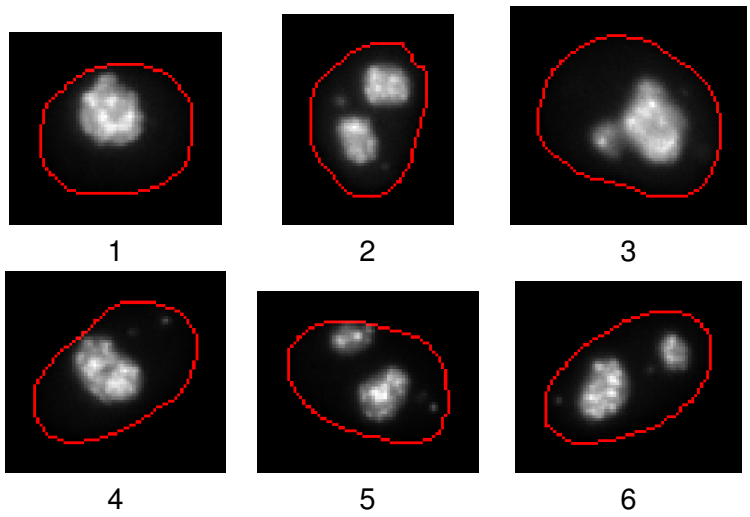


Figure 7.0.6: Well H12