LEUKO: Computer Decision Support System for Leukemia Diagnosis

Daniela M. Ushizima
UNISANTOS

Motivation

- Leukemia = 30% cancer.
- Pattern recognition: low-cost alternative solution;
- Human visual analysis:
  - 79.3% to 92.3% correct (lymphocytes);
  - 150 cells >> 10 minutes for an hematologist;
- Why lymphoproliferative diseases?
- Software for cell analysis and classification:
  - computer-aided automation;
  - educational purposes;

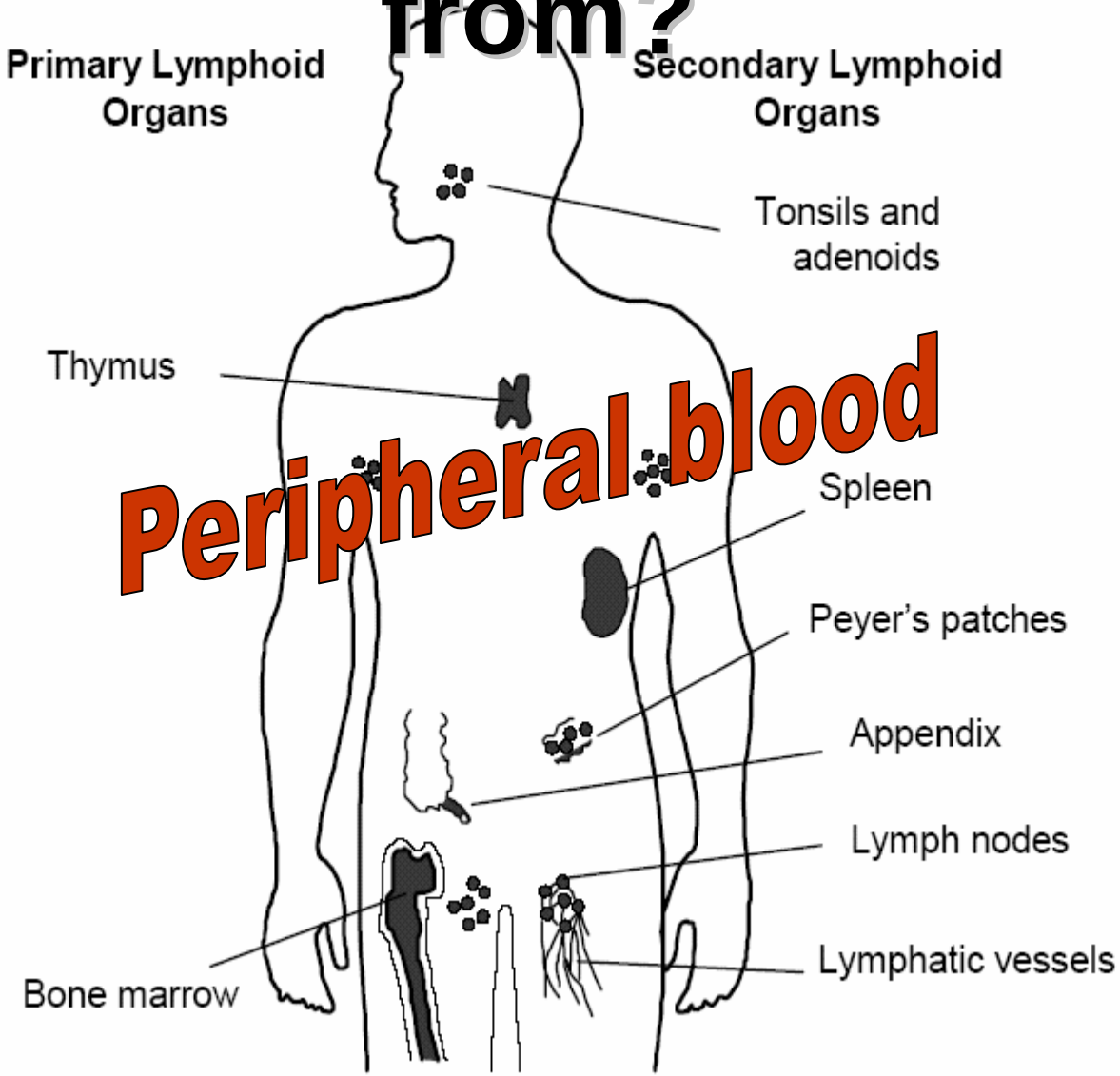
Synopsis

1. Human blood cells
2. Pattern recognition
3. Feature selection
4. Results
5. Latest developments

1

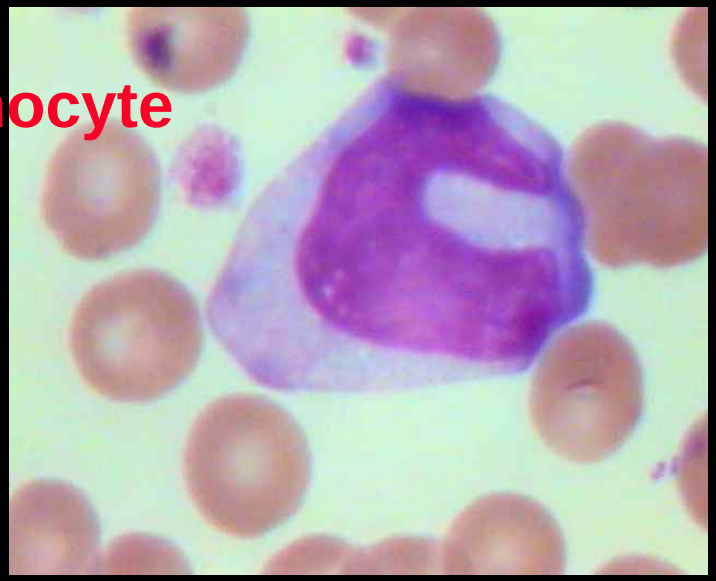
Human blood

1.1. Where leukocytes come from?

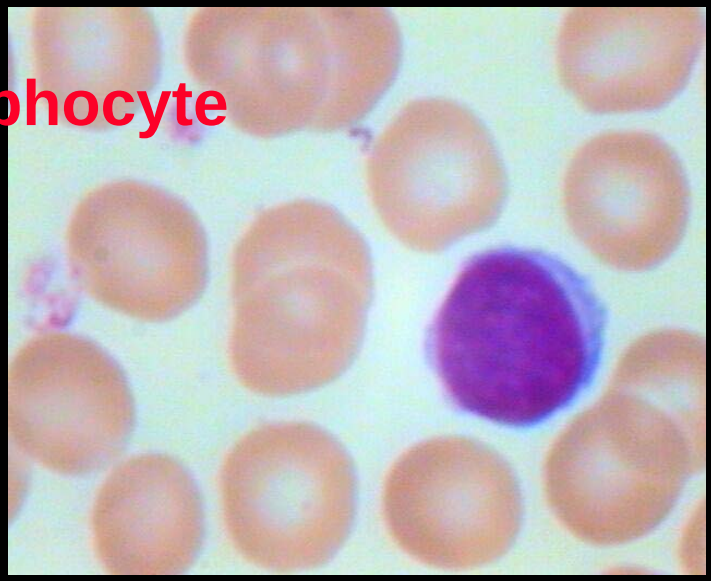


1.2. Normal leukocytes

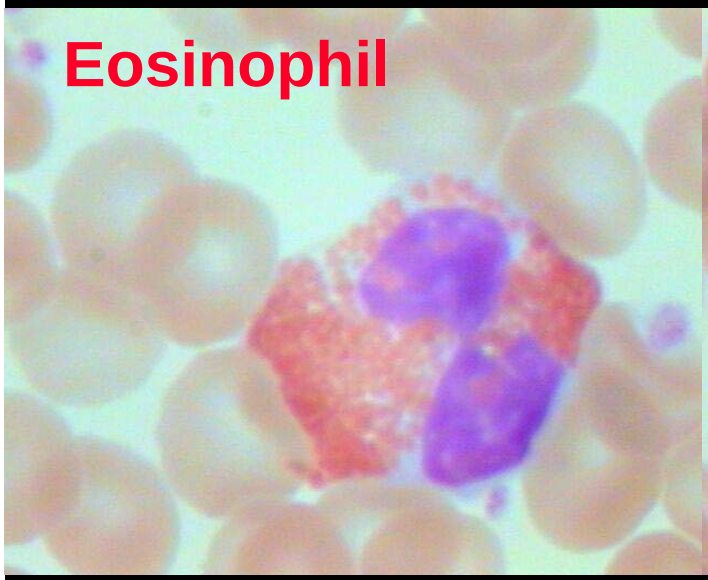
Monocyte



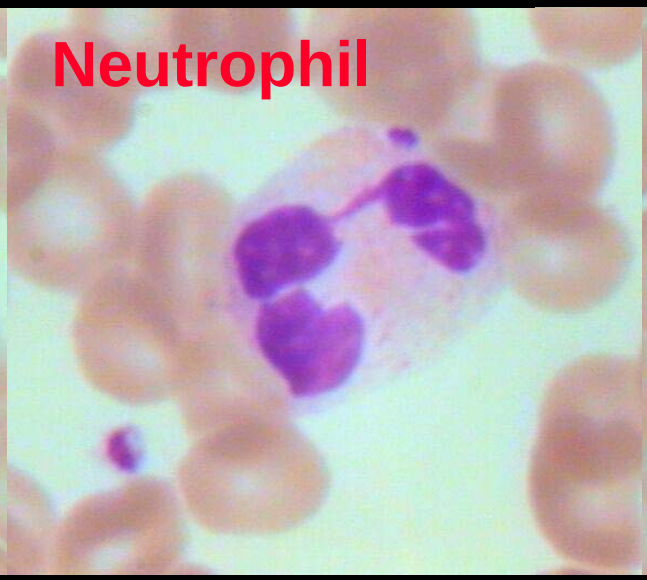
Lymphocyte



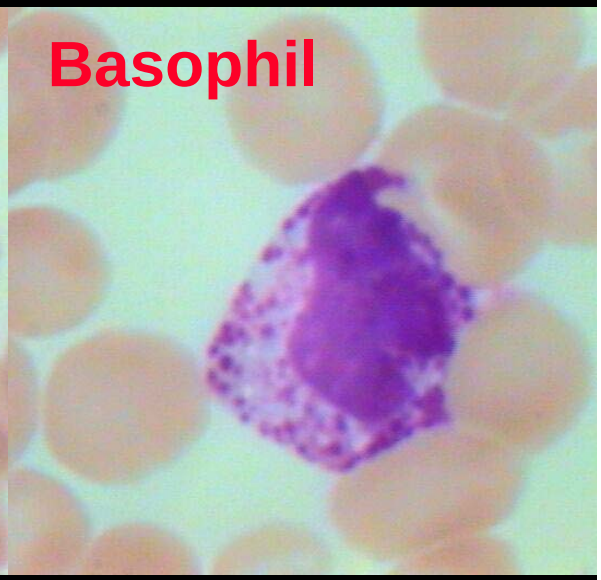
Eosinophil



Neutrophil

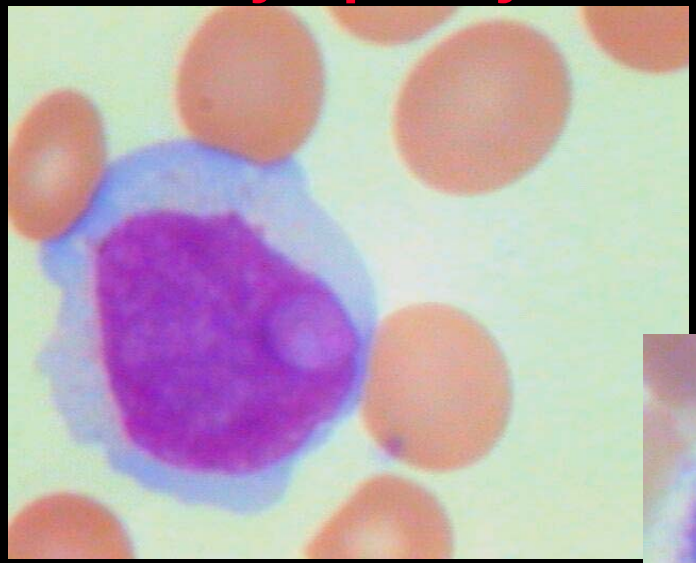


Basophil

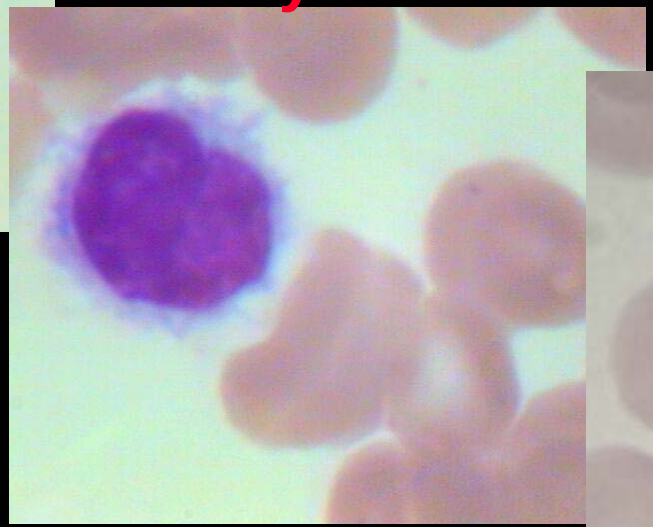


1.3. Malignant cells

Prolymphocyte



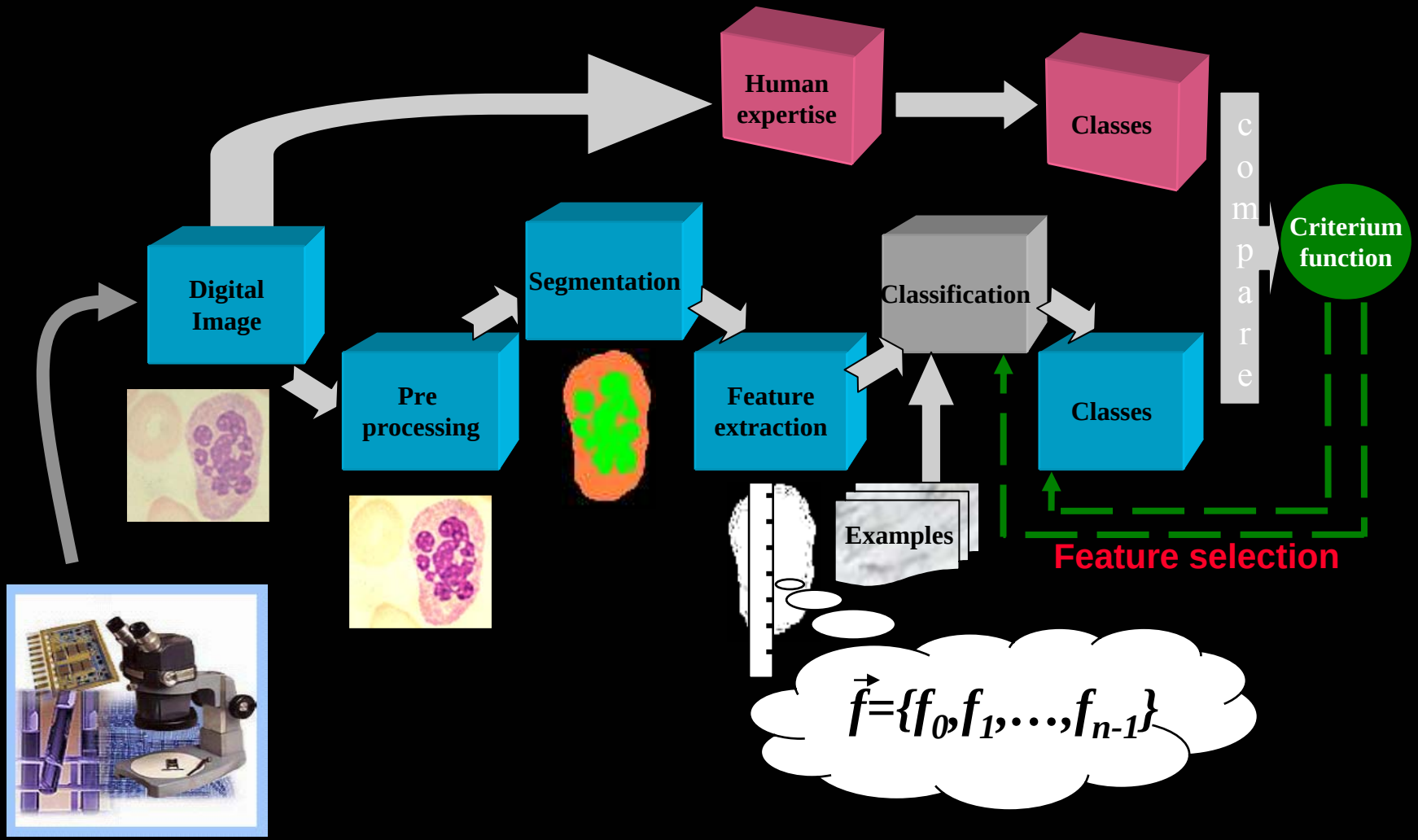
Hairy-cell



CLL

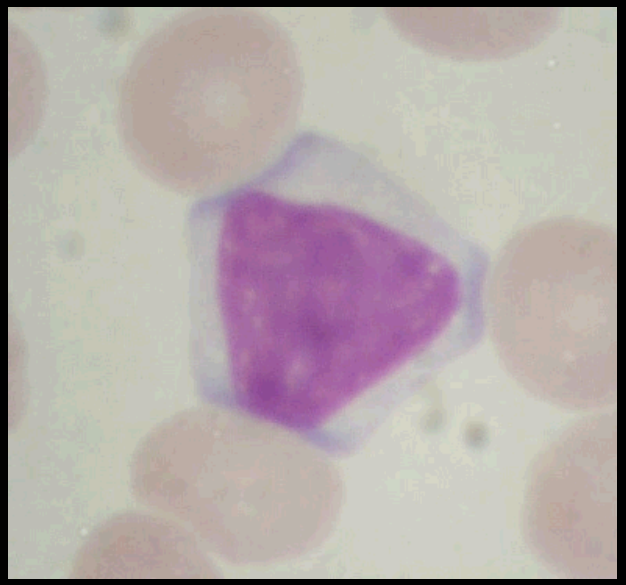
2
Pattern Recognition

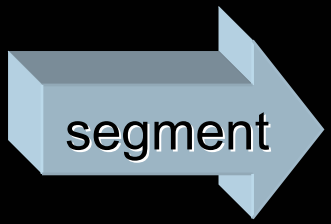
2.1. Steps in PR design

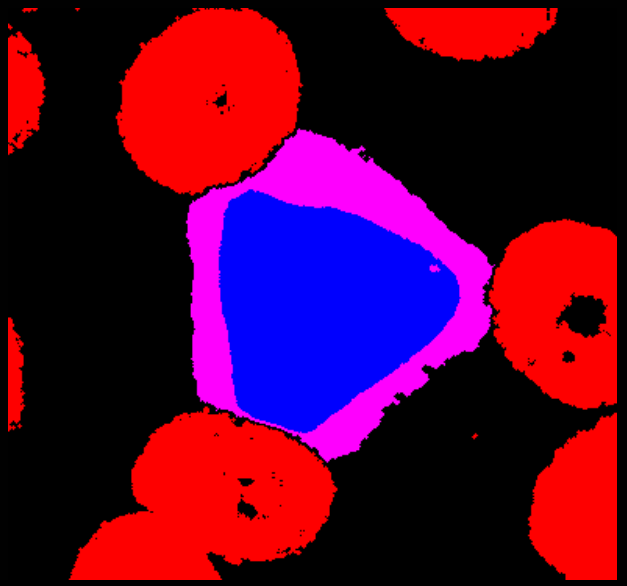


2.2. Segmentation

- Regions of interest: nucleus, cytoplasm, red blood cells and background



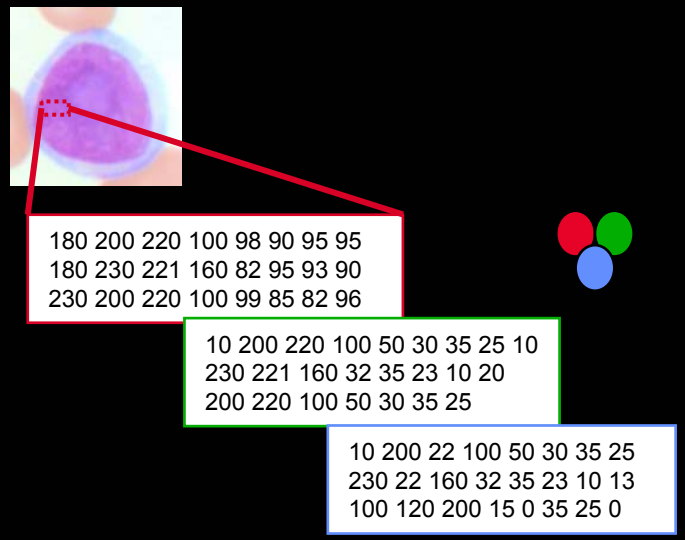


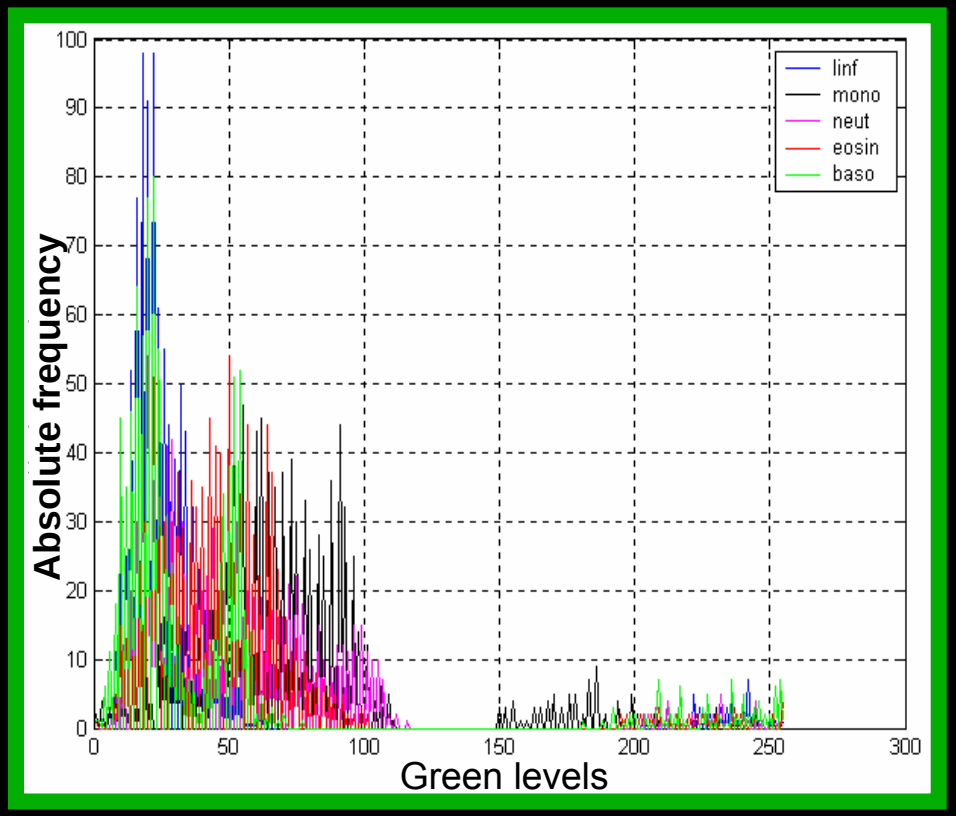


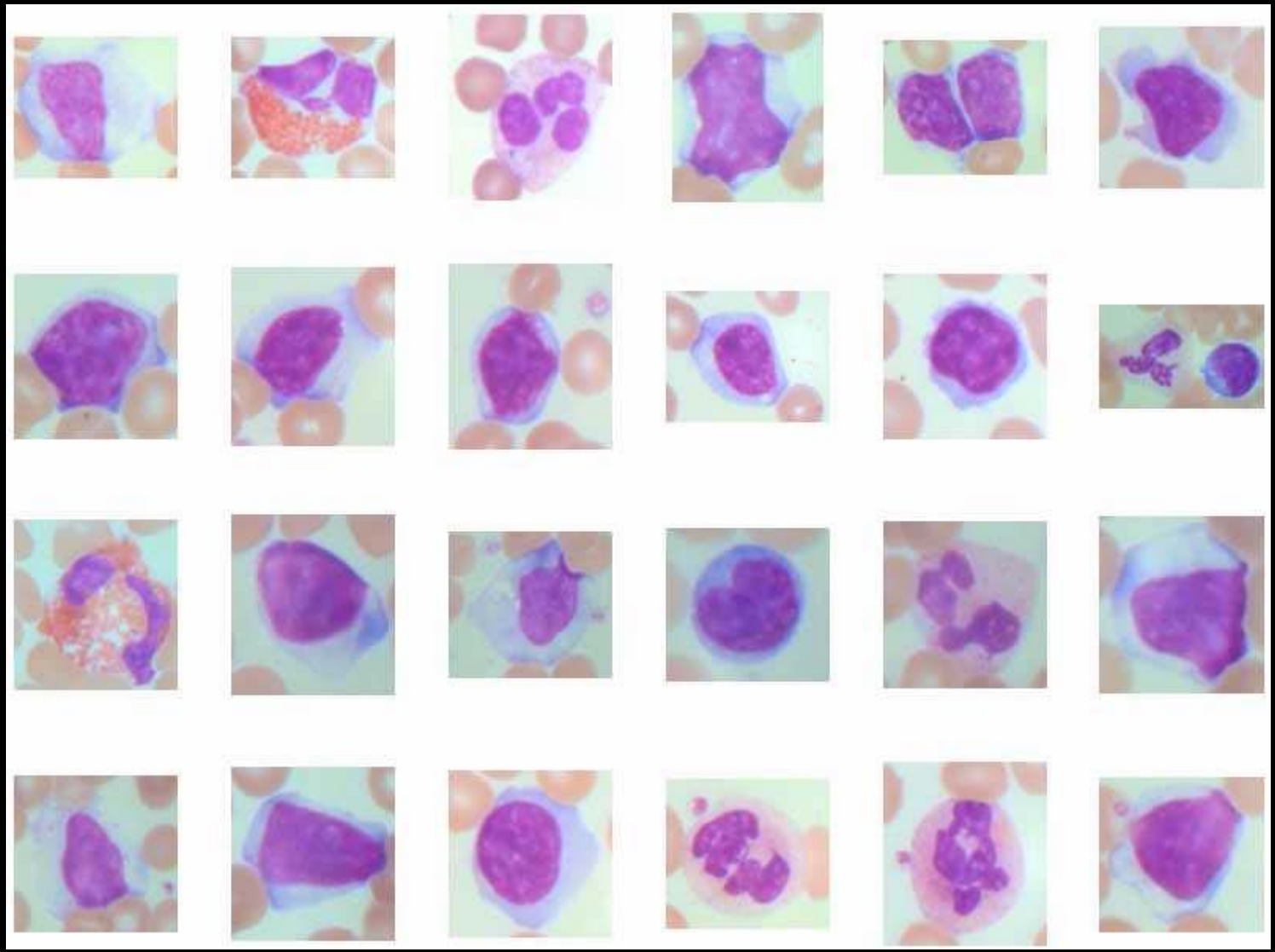
2.2.1. Methods

- G channel thresholding: where is the nucleus?
- Color segmentation: where are the leukocyte and RBCs?

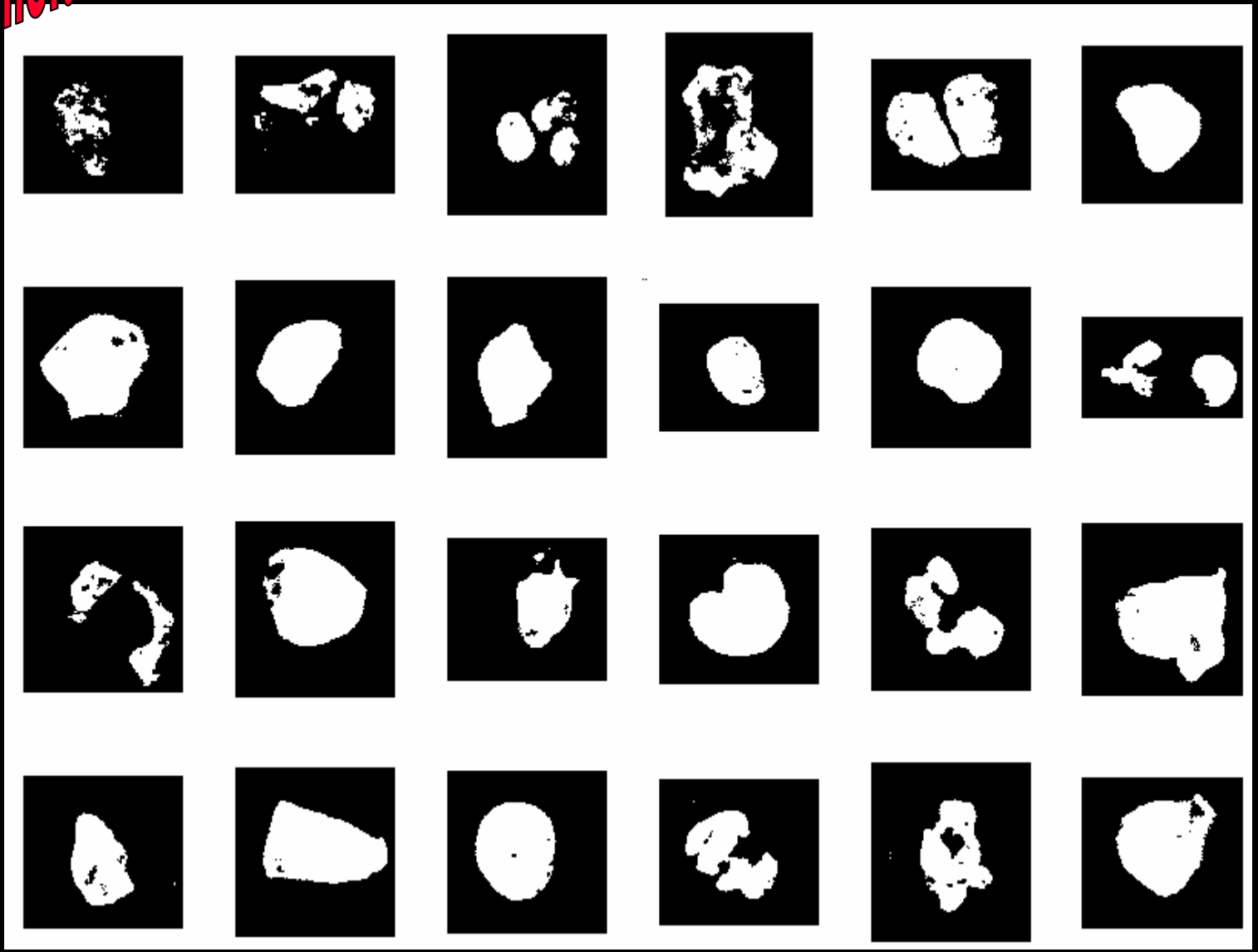
2.2.2. Thresholding





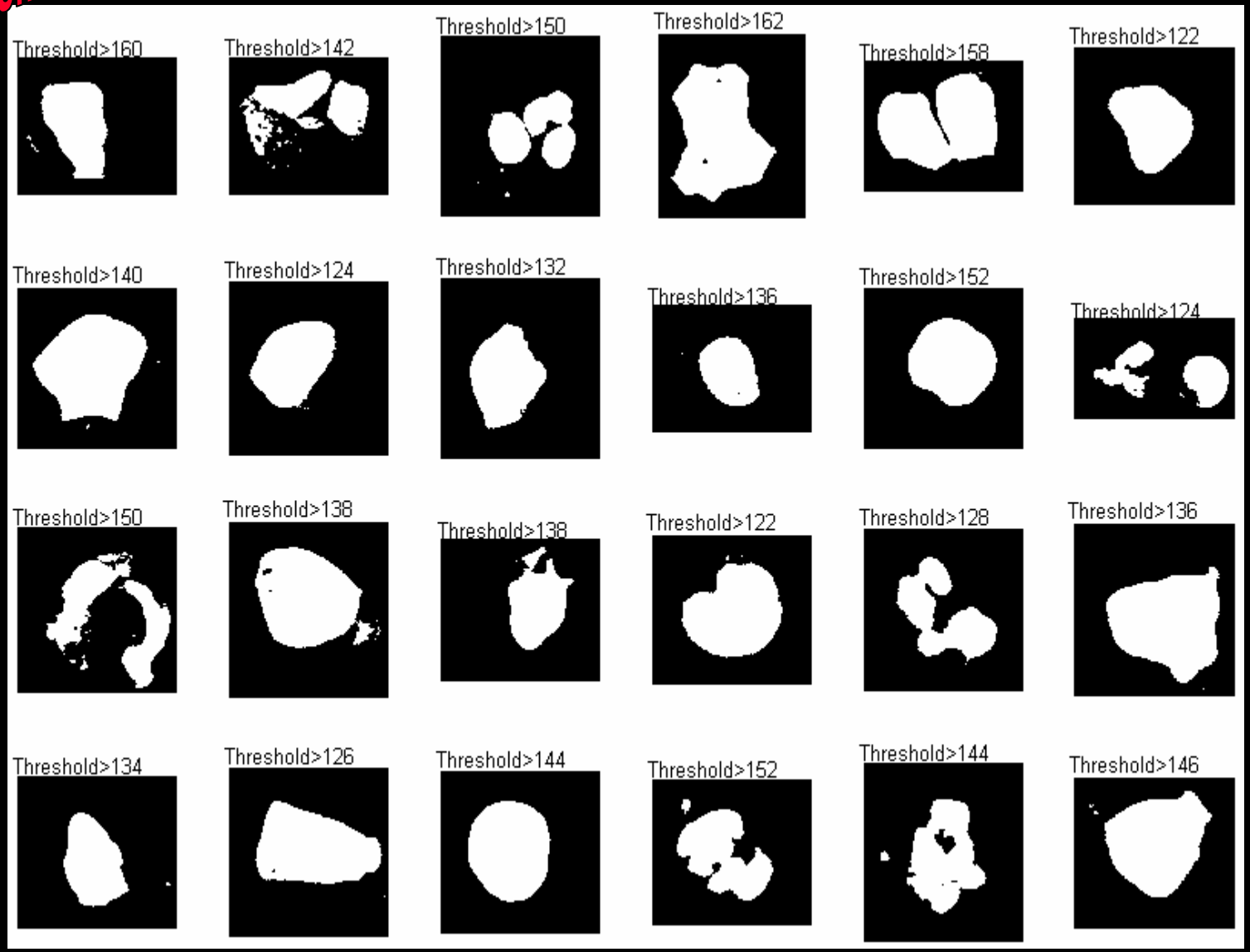


Fixed threshold



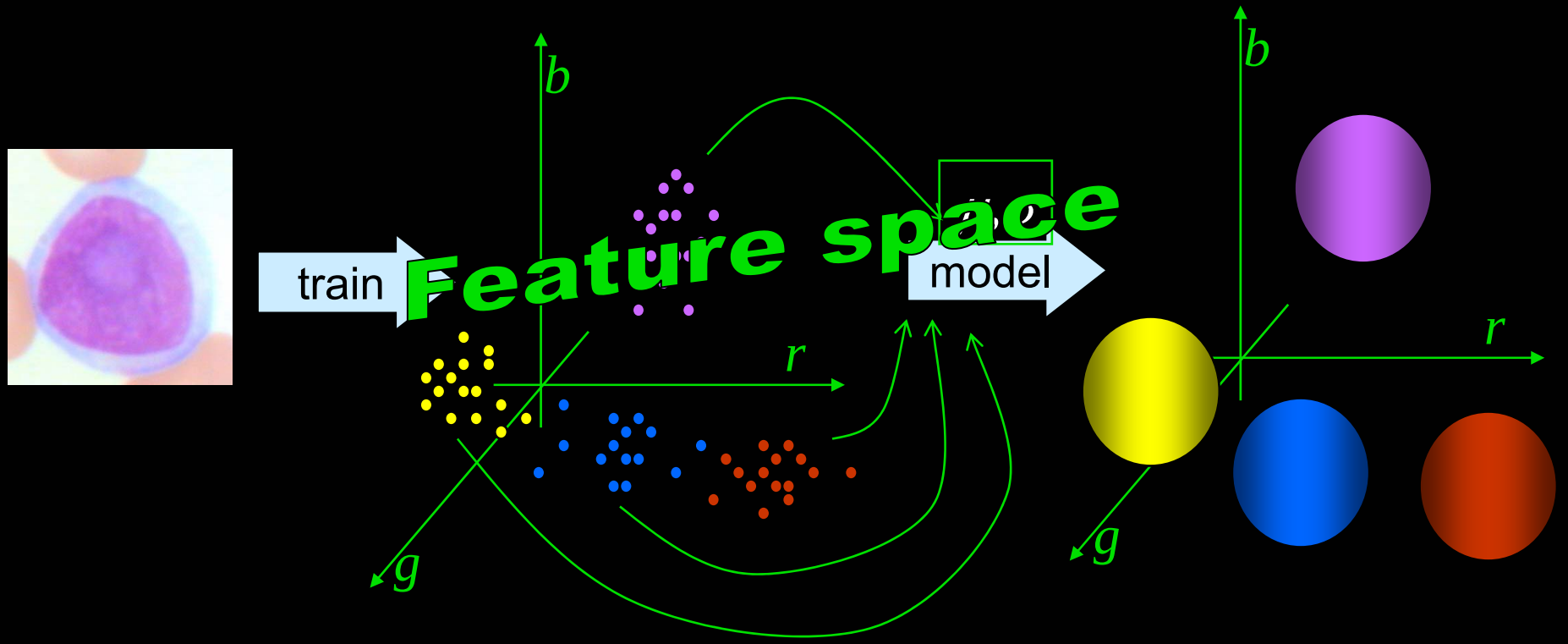
Katz (2000)

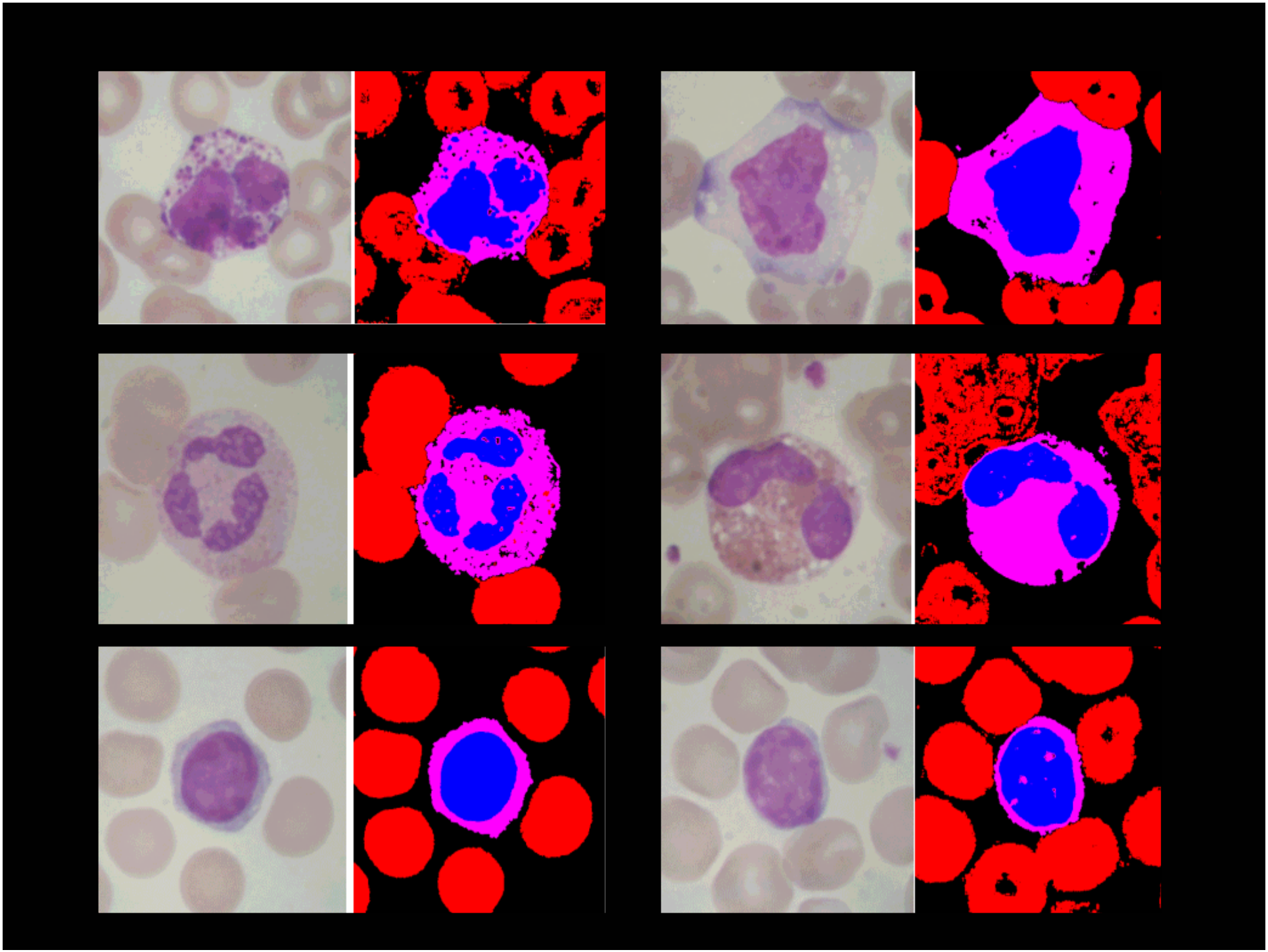
Dynamic threshold



Sabino et al (2003)

2.2.3. Color segmentation





2.3. Feature extraction

- Image properties;
- Nucleus and cytoplasm attributes;
- Shape;
- Size;
- Color;
- Texture.

Traditional measures

Subjective descriptions

Id	Feature description	Id	Feature description
1	N perimeter	32	N energy Blue
2	N area	33	N entropy Blue
3	N max area	34	N/C ratio
4	N num elements	35	C inertia1
5	N factor form	36	C entropy1
6	N num concavity	37	C energy1
7	N num convexity	38	C correlation1
8	N bending energy	39	C homogeneity1
9	N num holes	40	C inertia2
10	N area holes	41	C entropy2
11	N inertia	42	C energy2
12	N entropy	43	C correlation2
13	N energy	44	C homogeneity2
14	N correlation	45	C mean Red
15	N homogeneity	46	C std Red
16	N mean Red	47	C skewness Red
17	N std Red	48	C kurtosis Red
18	N skewness Red	49	C energy Red
19	N kurtosis Red	50	C entropy Red
20	N energy Red	51	C mean Green
21	N entropy Red	52	C std Green
22	N mean Green	53	C skewness Green
23	N std Green	54	C kurtosis Green
24	N skewness Green	55	C energy Green
25	N kurtosis Green	56	C entropy Green
26	N energy Green	57	C mean Blue
27	N entropy Green	58	C std Blue
28	N mean Blue	59	C skewness Blue
29	N std Blue	60	C kurtosis Blue
30	N skewness Blue	61	C energy Blue
31	N kurtosis Blue	62	C entropy Blue

N = nucleus, C = cytoplasm

Shape and size

Texture

Color

2.3.1. Shape and size

- Nucleus-cytoplasm ratio (*NC*):
  - $NC=A_{nuc}/(A_{nuc}+Area_{cit})$
- Factor form=compactness measure (*C*):
  - $C=P^2/A$
- Curvature (*k*):

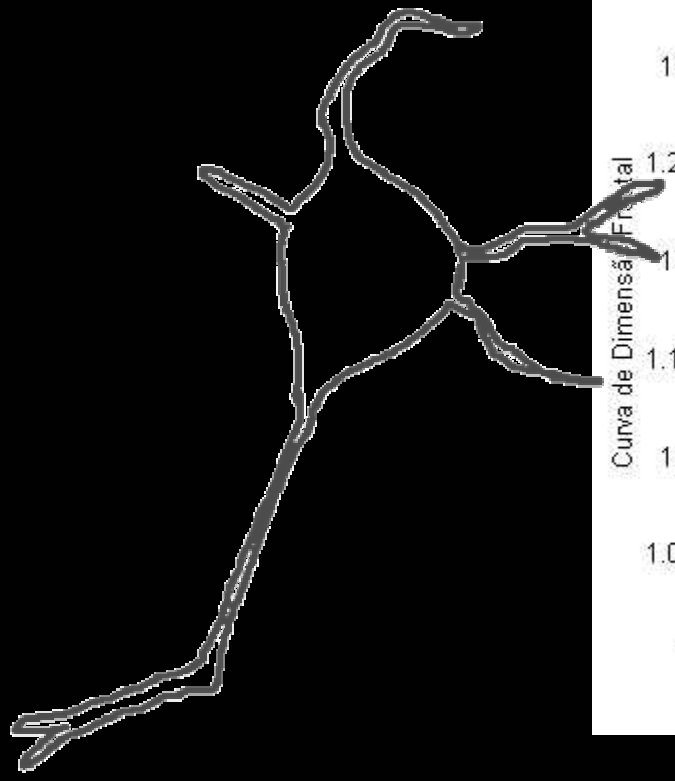
$$k(t)=\frac{x'(t)y''(t)-x''(t)y'(t)}{[x'(t)^2+y'(t)^2]^{3/2}}$$

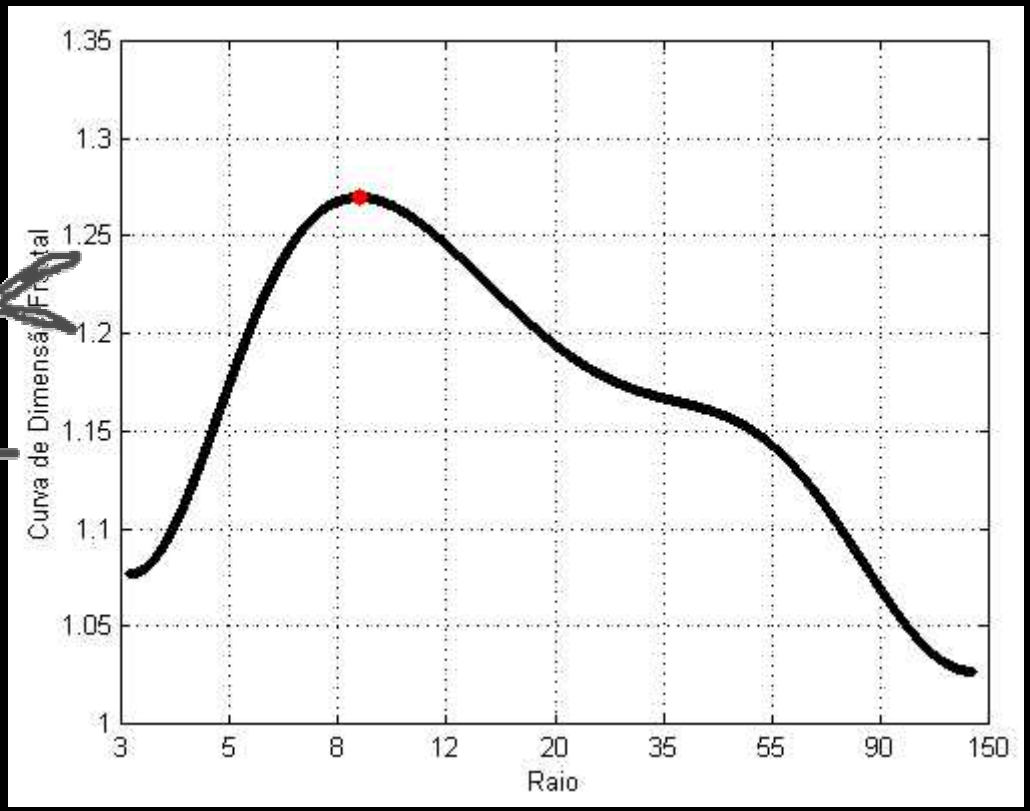
  - Concave and convexity points;
  - Bending Energy (*B*):

$$B=\frac{1}{P}\sum_{t=1}^{P}k^2(t)$$

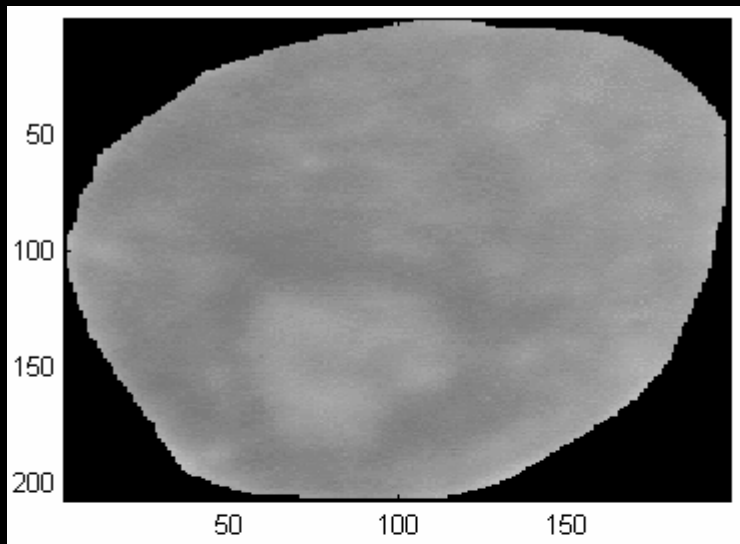
2.3.2. Fractal dimension

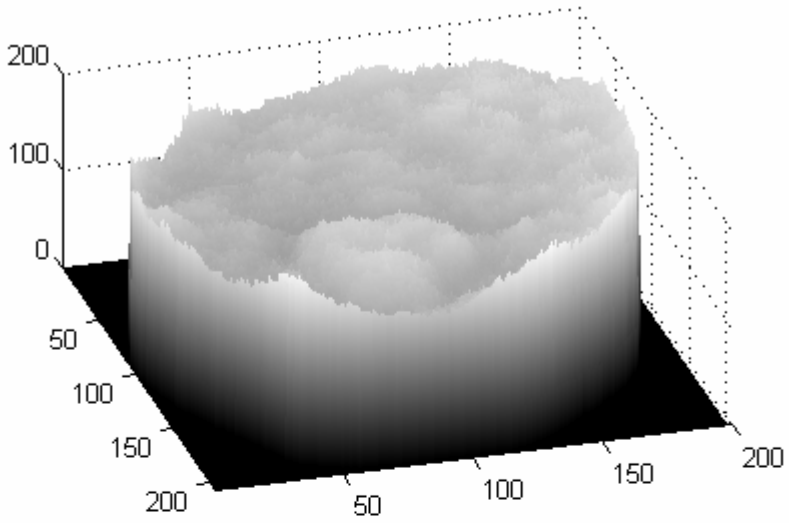
- Costa: multiscale fractal dimension (2D) curve and its peak (FD_{max}) for complexity description of neurons;

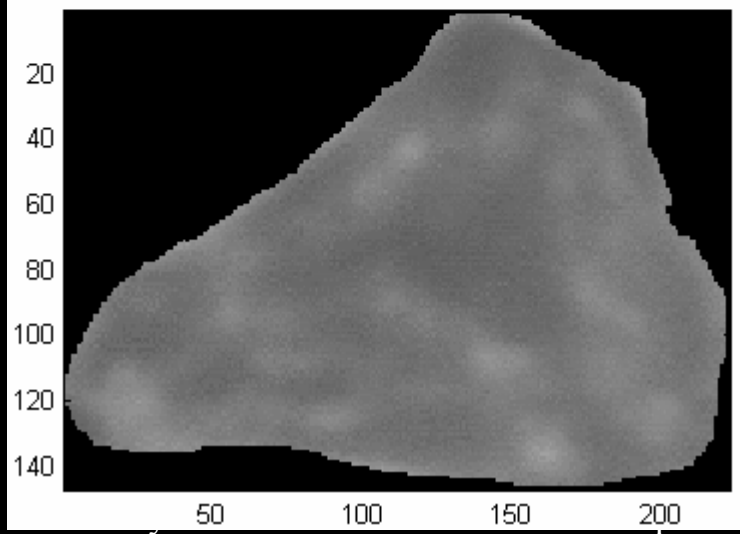


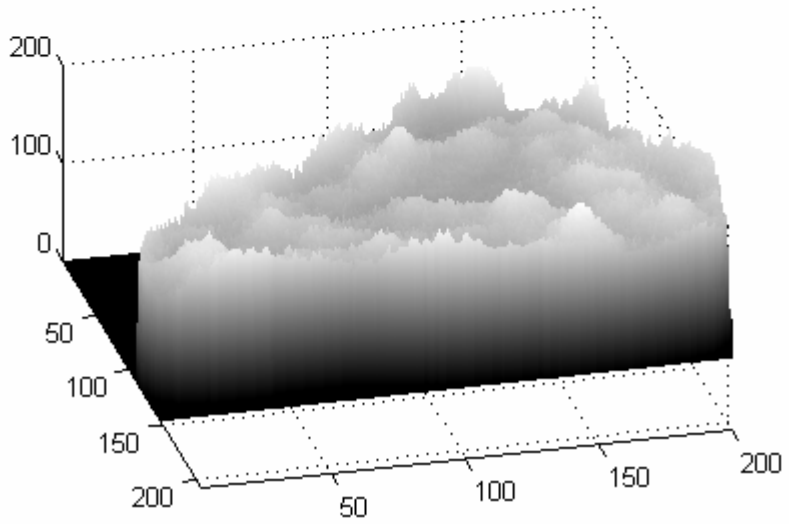


Spatial distribution of the graylevels

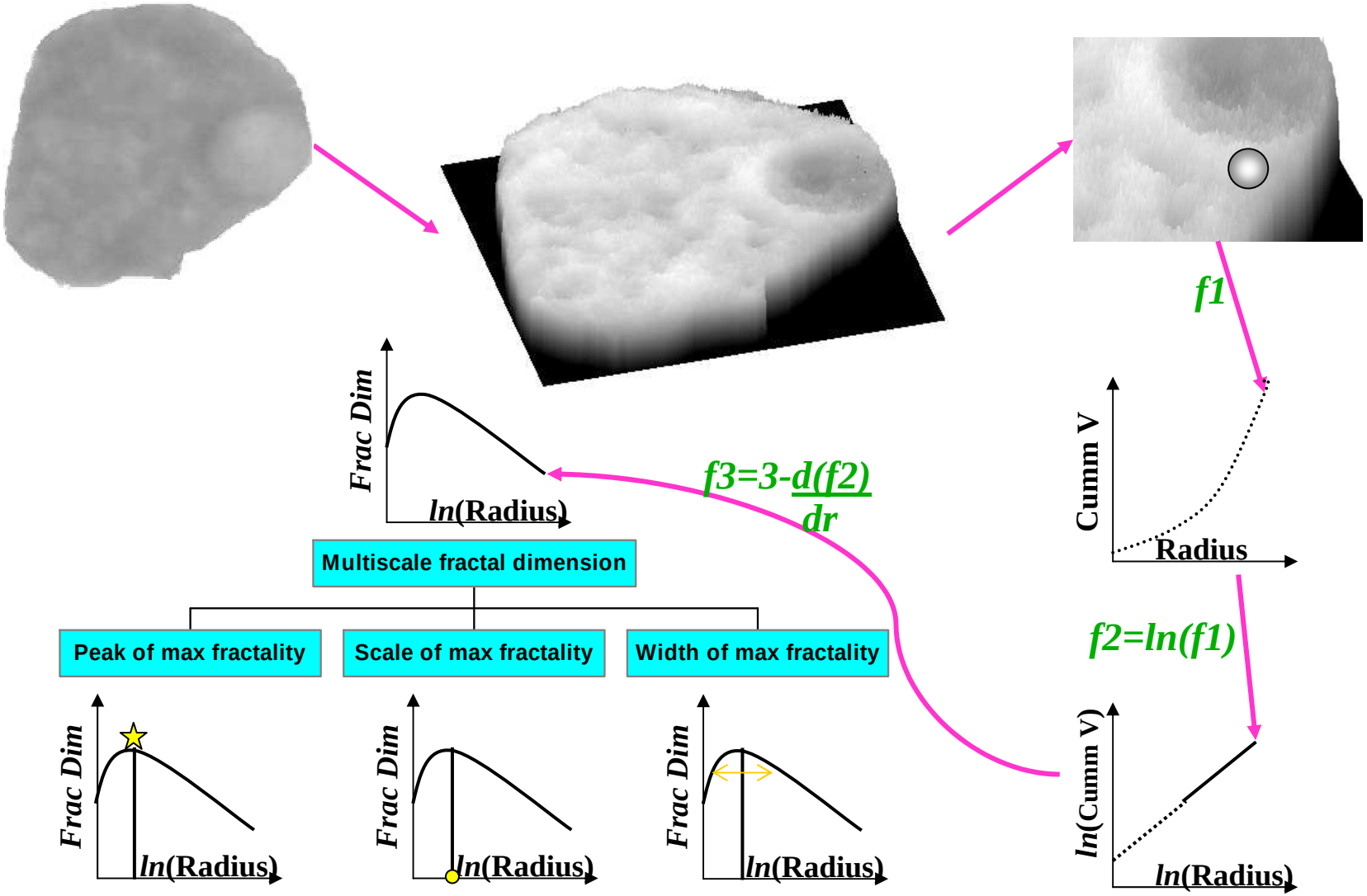






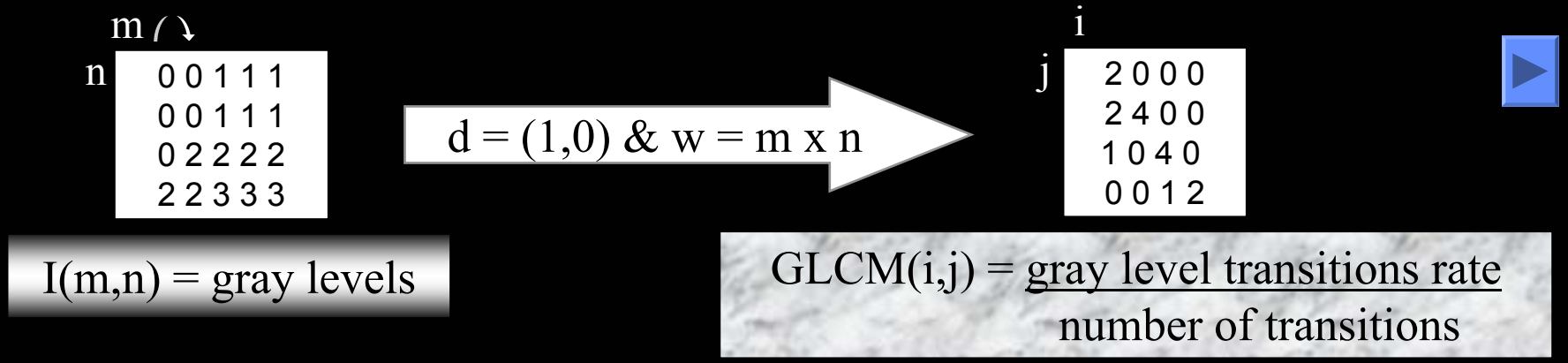


2.3.2. FD_{max} - the method



2.3.2. Gray Level Coocurrence Matrix

- Joint probability distribution of 2 pixels given a *displacement (direction+distance)* and a *window*;



- Spatial distribution and spatial dependence.

2.3.2. GLCM

$$Inertia = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)(i-j)^2$$

$$Entropy = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\log(p(i,j))$$

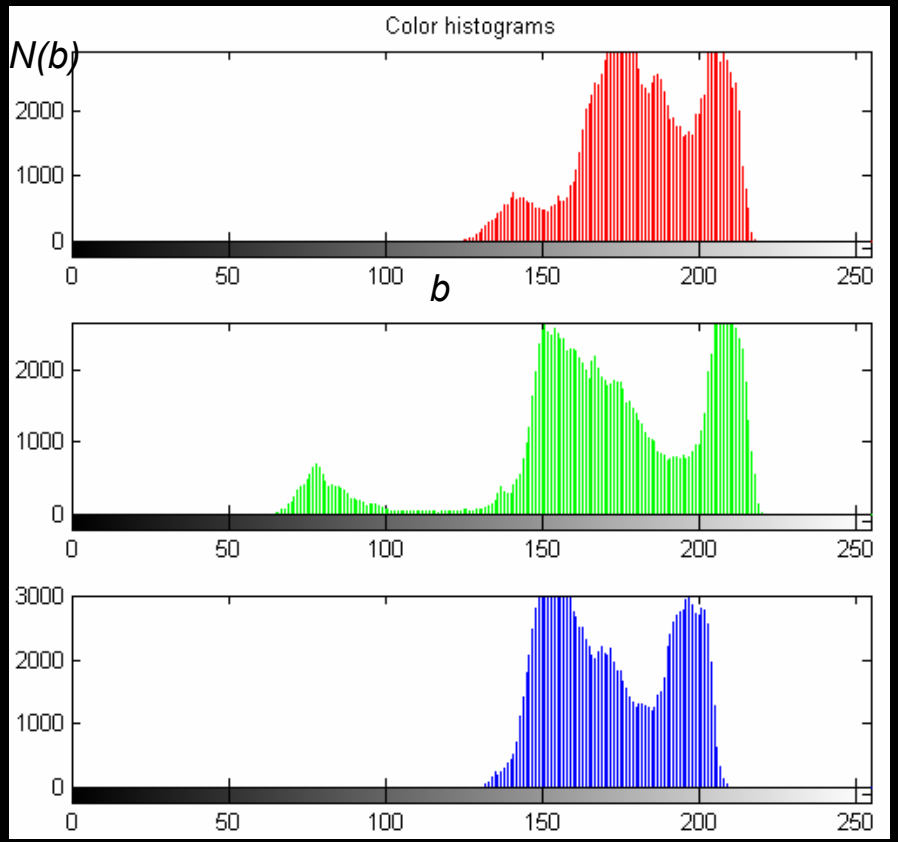
$$Energy = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)^2$$

$$Correlation = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left(i.j.p(i,j) \right) - \mu_x\mu_y}{\sigma_x\sigma_y}$$

$$Homogeneity = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{1}{1+(i-j)^2} p(i,j)$$

Haralick (1976)

2.3.3. Color



$$P(b) = \frac{N(b)}{M}$$

$$S_M = \overline{b} = \sum_{b=0}^{255} b * P(b)$$

$$S_D = [\sum_{b=0}^{255} (b - \overline{b})^2]^1/2$$

$$S_S = \frac{1}{\sigma^3} \sum_{b=0}^{255} (b - \overline{b})^3 P(b)$$

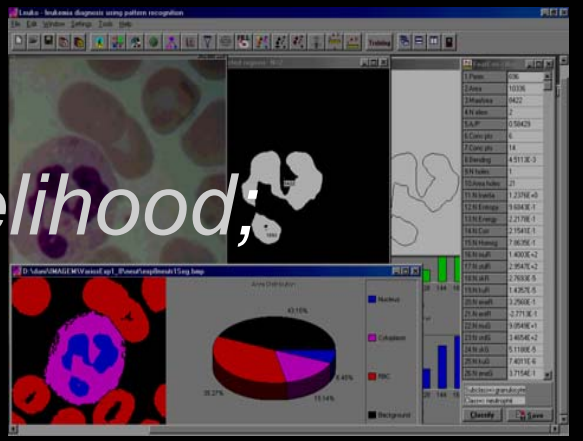
$$S_k = \frac{1}{\sigma^4} \sum_{b=0}^{255} (b - \overline{b})^4 P(b) - 3$$

$$S_N = \frac{1}{\sigma^3} \sum_{b=0}^{255} [P(b)]^2$$

$$S_E = -\frac{1}{\sigma^3} \sum_{b=0}^{255} P(b) * log_2[P(b)]$$

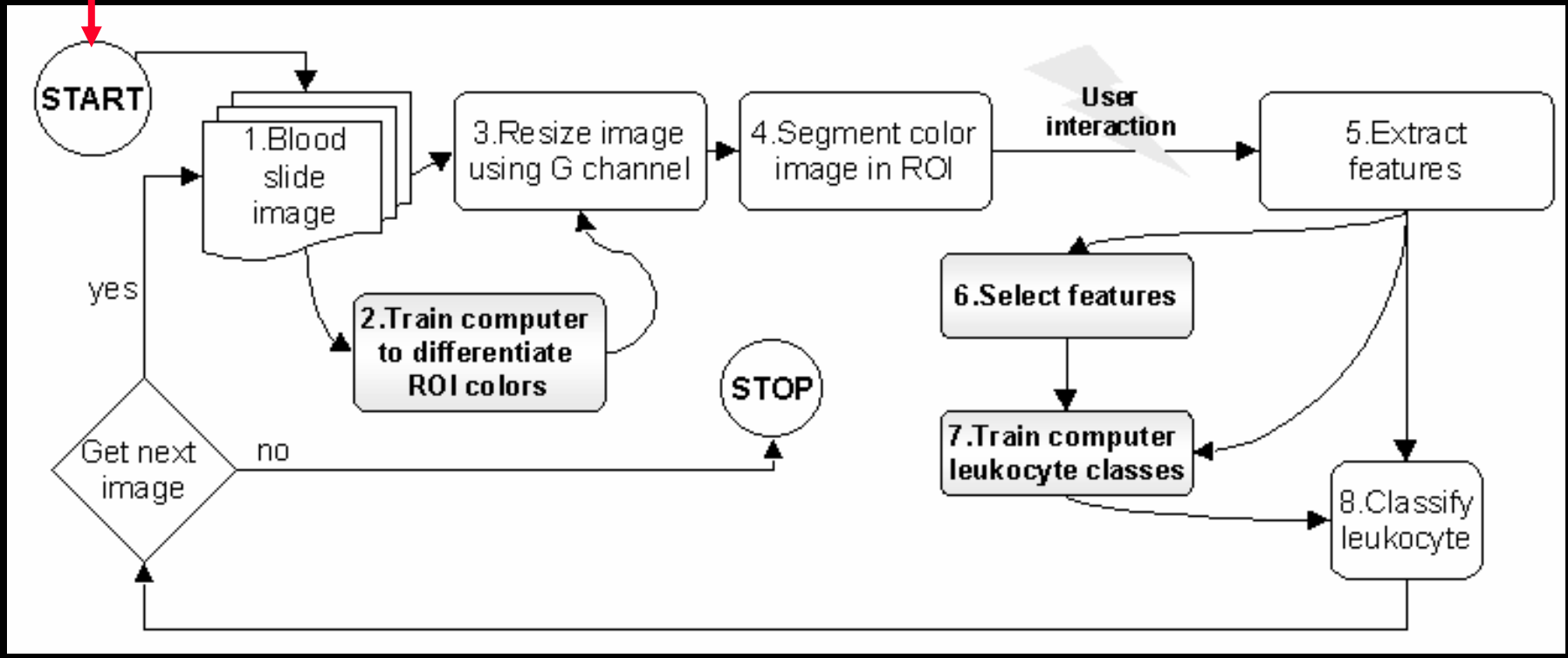
Pratt (1991), Song et al (1997)

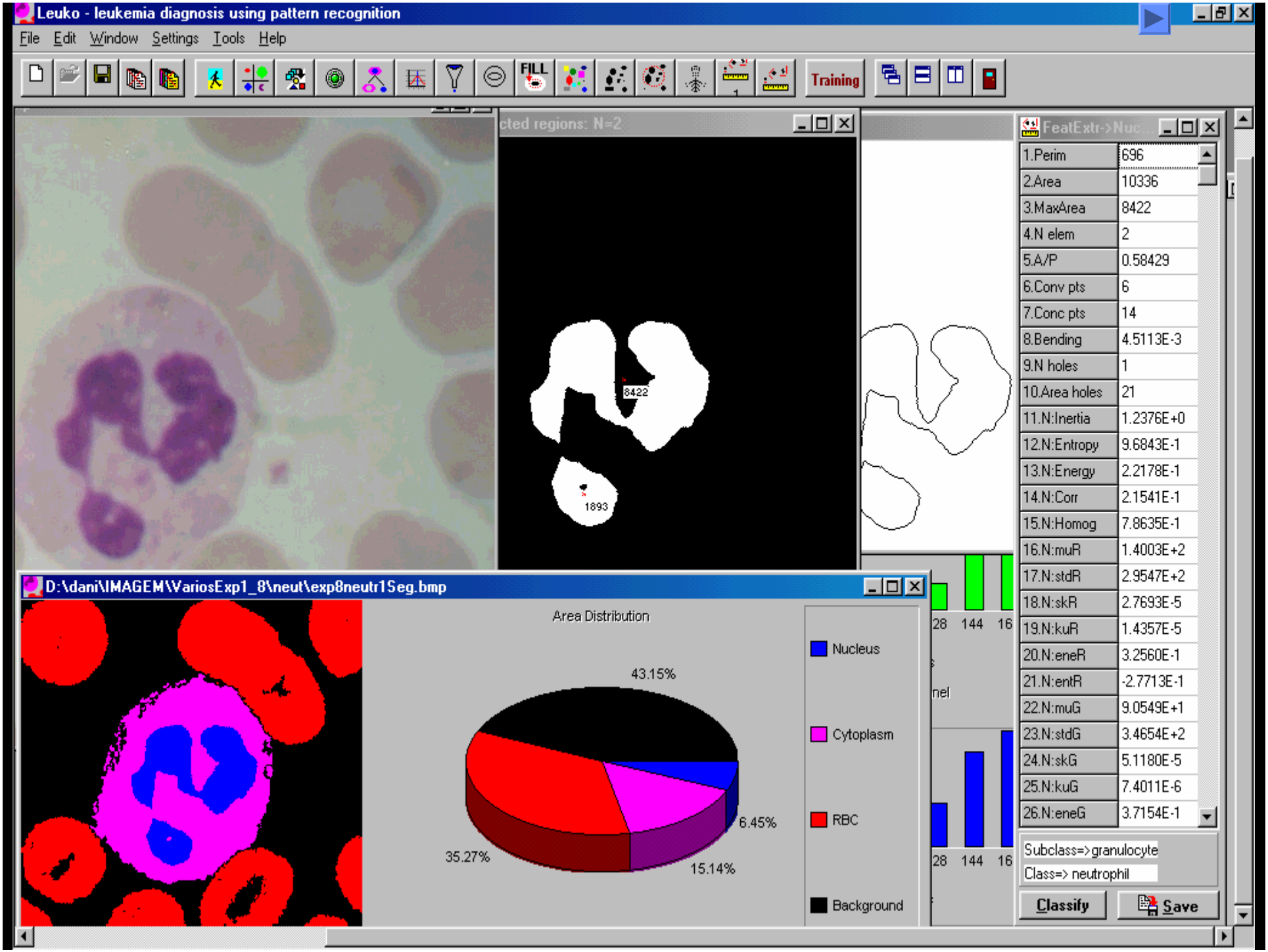
2.4. *Leuko* classification



- Learning: *supervised*;
- Decision rule: *maximum likelihood*;
- Pdf estimation: *parametric;*
- Pdf: *Gaussian*;
- Error estimation methods regarded:
  - *Resubstituion, holdout, cross-validation, leave-one-out.*

Leuko


START
1.Blood slide image
3.Resize image using G channel
4.Segment color image in ROI
User interaction
5.Extract features
6.Select features
2.Train computer to differentiate ROI colors
STOP
yes
Get next image
no
7.Train computer leukocyte classes
8.Classify leukocyte

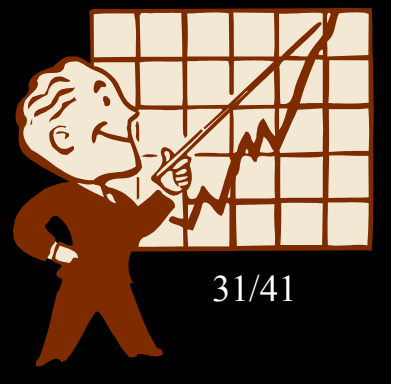

Leuko - leukemia diagnosis using pattern recognition
File Edit Window Settings Tools Help
Training
cted regions: N=2
8422
1893
FeatExtr->Nuc...
1.Perim 696
2.Area 10336
3.MaxArea 8422
4.N elem 2
5.A/P 0.58429
6.Conv pts 6
7.Conc pts 14
8.Bending 4.5113E-3
9.N holes 1
10.Area holes 21
11.N:Inertia 1.2376E+0
12.N:Entropy 9.6843E-1
13.N:Energy 2.2178E-1
14.N:Corr 2.1541E-1
15.N:Homog 7.8635E-1
16.N:muR 1.4003E+2
17.N:stdR 2.9547E+2
18.N:skR 2.7693E-5
19.N:kuR 1.4357E-5
20.N:eneR 3.2560E-1
21.N:entR -2.7713E-1
22.N:muG 9.0549E+1
23.N:stdG 3.4654E+2
24.N:skG 5.1180E-5
25.N:kuG 7.4011E-6
26.N:eneG 3.7154E-1
28 144 16
nel
28 144 16
Subclass=>granulocyte
Class=> neutrophil
Classify
Save
D:\dani\IMAGEM\VariosExp1_8\neut\exp8neutr1Seg.bmp
Area Distribution
43.15%
6.45%
15.14%
35.27%
Nucleus
Cytoplasm
RBC
Background

3
Feature
selection

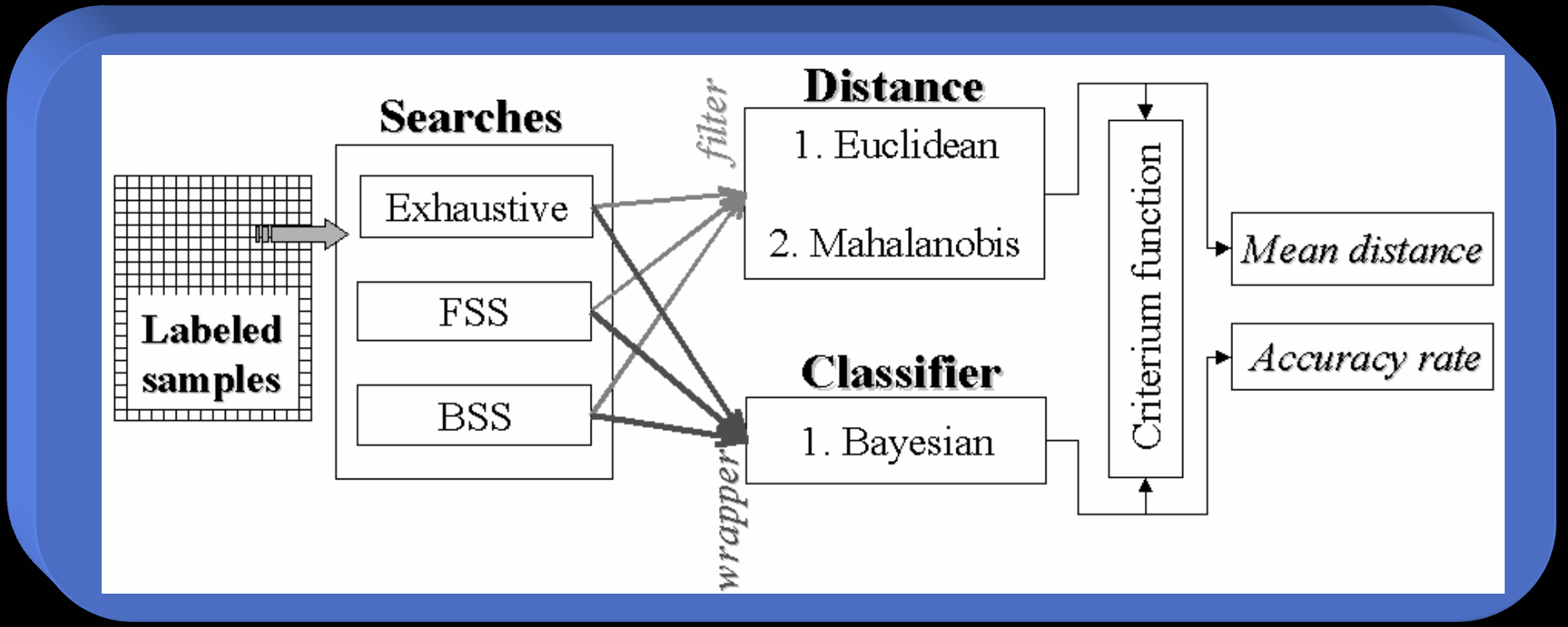
3. Feature selection

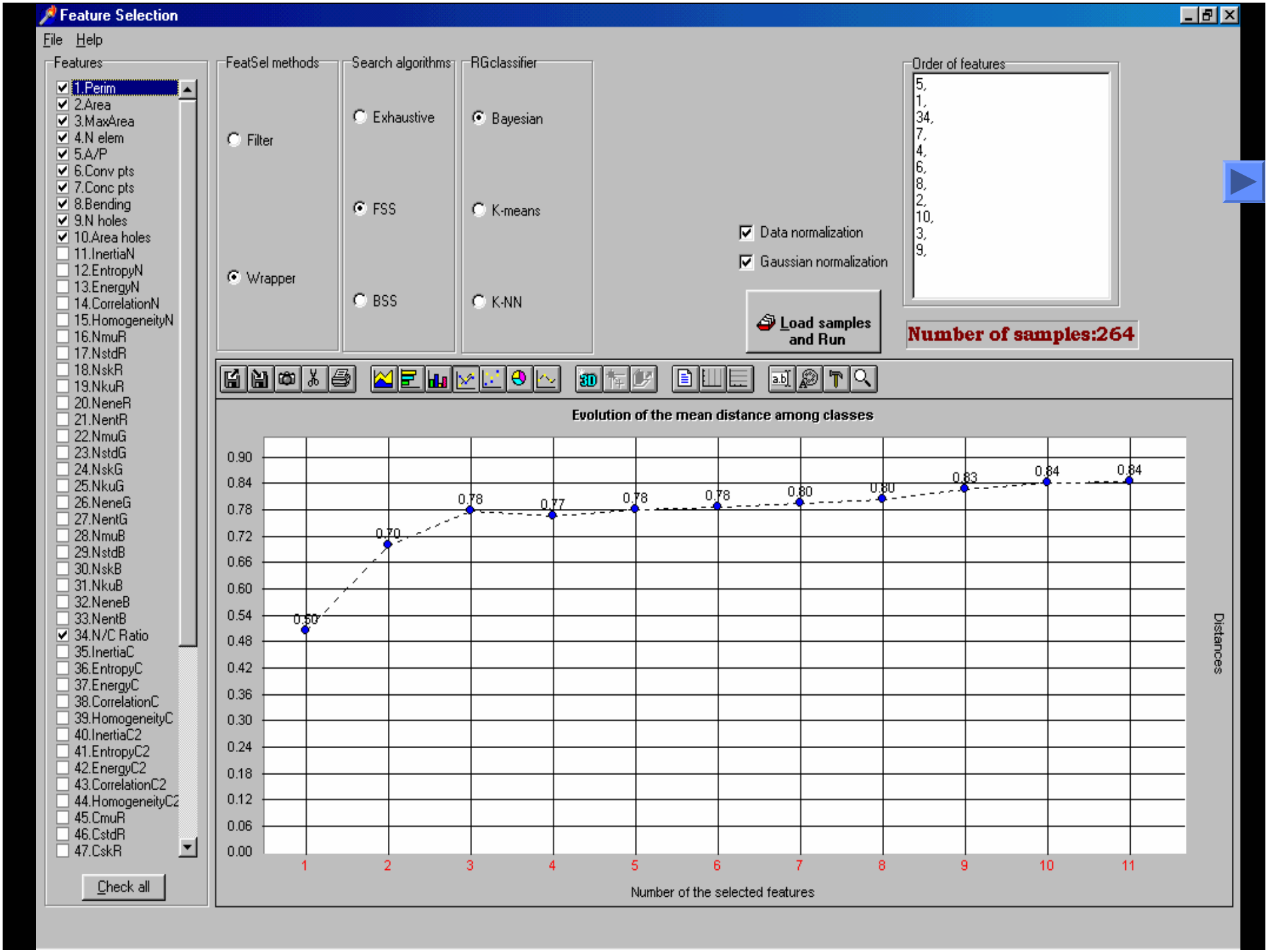
- Problems of dealing with many features:
  - Overfitting *(decision surface = many details)*
  - Overtraining *(decision surface = few details)*
  - Curse of dimensionality *(lacking samples)*;
  - Peaking phenomena.

- ∴ Define feature subsets using strategies:
  - Exaustive;
  - Heuristics;
  - Random;



3. Feature selection




Feature Selection
File Help
Features
1.Perim
2.Area
3.MaxArea
4.N elem
5.A/P
6.Conv pts
7.Conc pts
8.Bending
9.N holes
10.Area holes
11.InertiaN
12.EntropyN
13.EnergyN
14.CorrelationN
15.HomogeneityN
16.NmuR
17.NstdR
18.NskR
19.NkuR
20.NeneR
21.NentR
22.NmuG
23.NstdG
24.NskG
25.NkuG
26.NeneG
27.NentG
28.NmuB
29.NstdB
30.NskB
31.NkuB
32.NeneB
33.NentB
34.N/C Ratio
35.InertiaC
36.EntropyC
37.EnergyC
38.CorrelationC
39.HomogeneityC
40.InertiaC2
41.EntropyC2
42.EnergyC2
43.CorrelationC2
44.HomogeneityC2
45.CmuR
46.CstdR
47.CskR
Check all
FeatSel methods
Filter
Wrapper
Search algorithms
Exhaustive
FSS
BSS
RGclassifier
Bayesian
K-means
K-NN
Data normalization
Gaussian normalization
Load samples and Run
Order of features
5,
1,
34,
7,
4,
6,
8,
2,
10,
3,
9,
Number of samples:264
Evolution of the mean distance among classes
0.90
0.84
0.78
0.72
0.66
0.60
0.54
0.48
0.42
0.36
0.30
0.24
0.18
0.12
0.06
0.00
0.50
0.70
0.78
0.77
0.78
0.78
0.80
0.80
0.83
0.84
0.84
1
2
3
4
5
6
7
8
9
10
11
Number of the selected features
Distances

$\frac{\partial \varphi}{\partial t} = k\nabla^2\varphi$
$\frac{\partial \varphi}{\partial t} = k\nabla^2\varphi$

4

Results

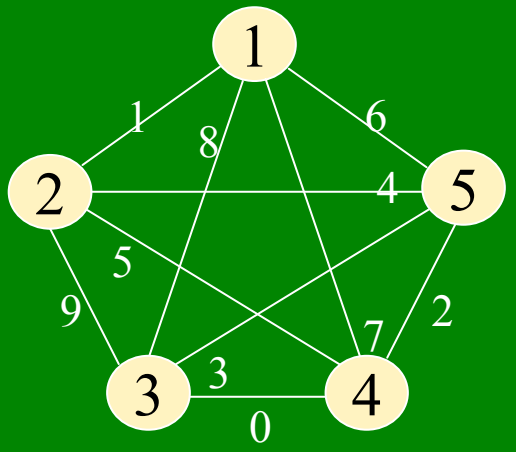
4. Results

- Differentiation among 6 types of leukocytes, including the CLL (the most common leukemia in adults from Western);

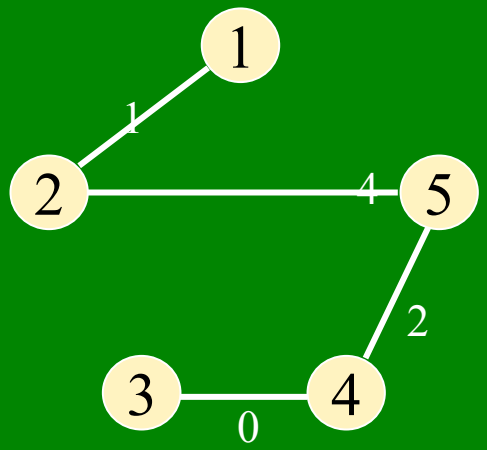
%	baso	eosi	linf	mono	neut	LLC
baso	100					
eosi		86			14	
linf			92	1	1	6
mono		3	13	81	3	
neut	1	1			98	
LLC		2	8	2	2	86

- Project and development of pattern recognition and feature selection tools.

5
Latest developments



Graph *G*



MST

T

5.1. *Leuko* + SVM

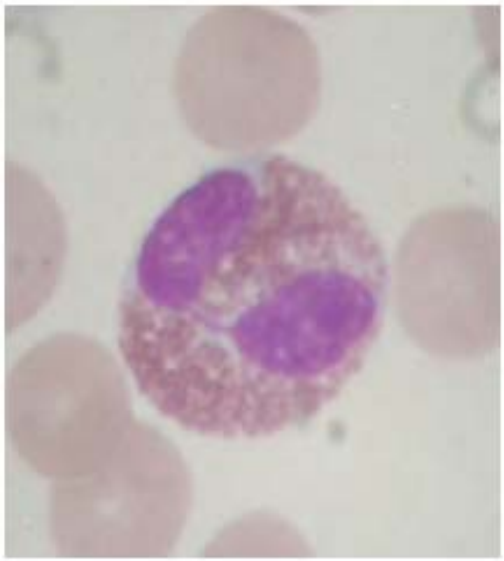
- SVM extension to multiclass task, applying tree-based strategy;
- SVM and minimum spanning tree:
  - Adapted version of Kruskal algorithm
    - Generates a tree in a bottom-up iterative way
    - Polynomial complexity
      - Including weight assignment
    - Allows the **automatic** determination of the multiclass tree.

5.1. *Leuko* accuracy

	CLASSES								
Technique	Baso	Eosi	Lymp	Mono	Neut	CLL	PL	HCL	Total
1AA	92.5	**96.4**	90.8	90.6	97.6	86.8	79.5	**82.9**	91.5 (3.0)
AAA	91.0	93.6	91.2	89.6	97.1	**89.4**	78.0	*72.9*	90.6 (2.5)
ECOC	**95.0**	94.6	**94.0**	**91.7**	**98.8**	84.1	79.0	81.4	**92.1** (1.9)
DDAG	**95.0**	95.5	92.2	91.6	97.2	89.3	73.5	75.7	91.5 (3.0)
ADAG	**95.0**	95.5	91.9	91.6	97.2	89.3	73.5	77.1	91.5 (3.0)
Tree	*88.5*	94.6	90.8	87.9	97.6	84.8	**80.0**	77.1	90.1 (2.3)
WW	90.0	**96.4**	89.4	86.0	96.3	84.8	*71.5*	**82.9**	89.6 (2.2)
CS	92.5	**96.4**	91.2	88.0	95.5	85.4	73.0	78.6	90.1 (3.0)
NB	90.5	*92.5*	*63.7*	*73.1*	*70.2*	*72.8*	74.0	75.7	*73.0* (*4.6*)

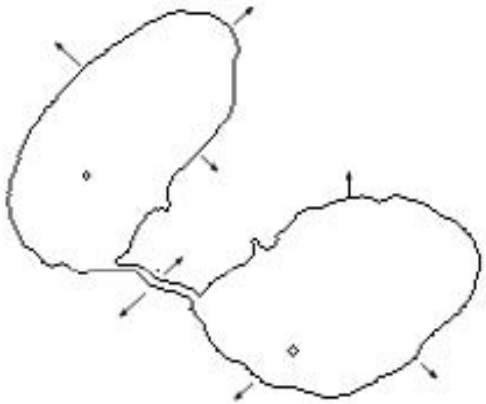
5.2. Level set

- $\gamma(t)$ = boundary of the nucleus in $\Phi(x,t=0)$
- $\Phi t + F|\nabla\Phi| = 0$



(a)

$$F = 1/(1+\nabla(G*I))$$



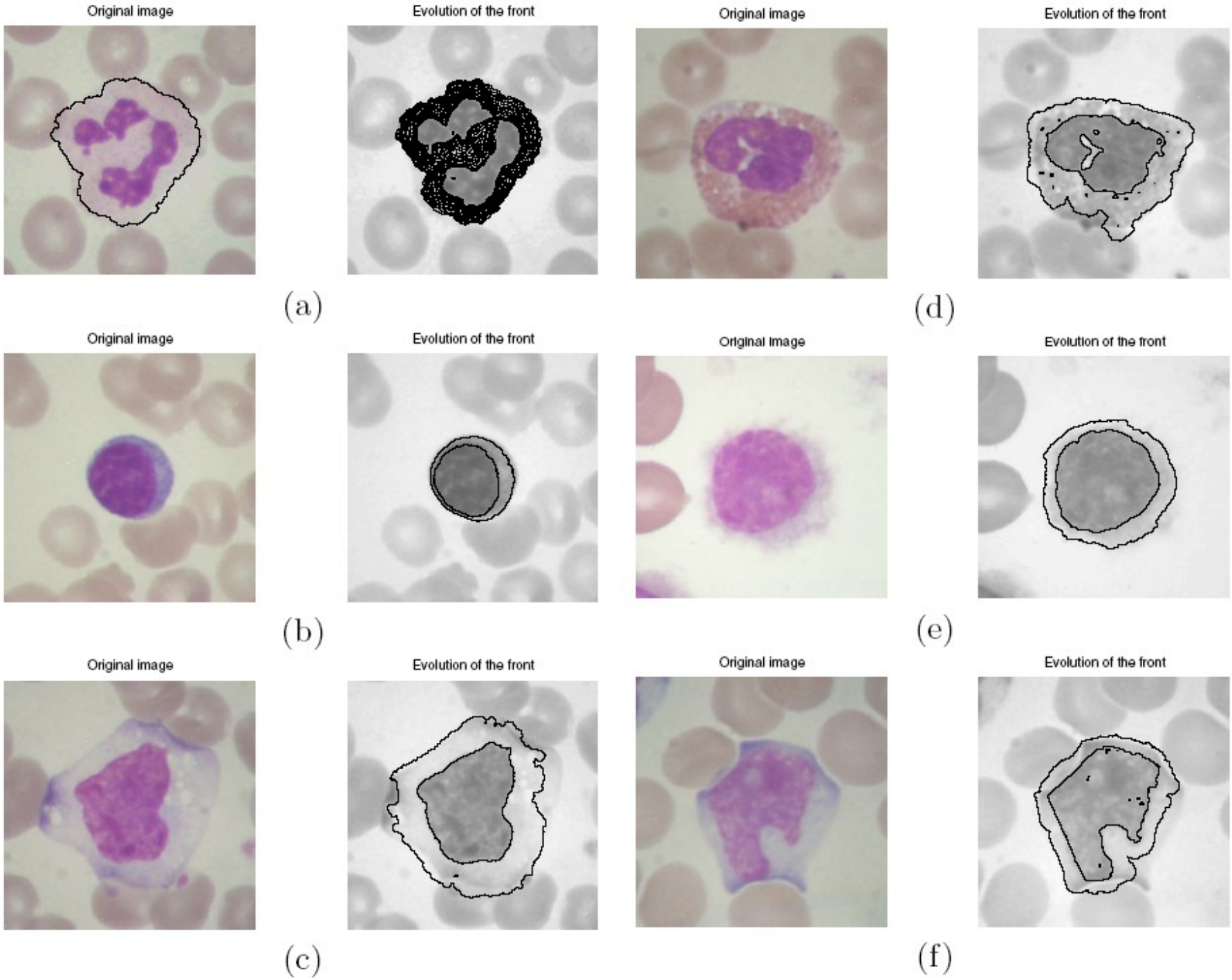
(b)

5.2.1. Colourful speed function

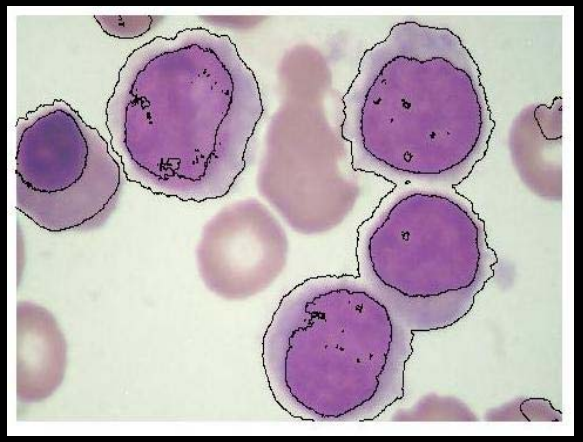
- Given C, the set of points of the nucleus contour and $p \in C$: $D_C(p) = min(d(p,q))$
- If $0<d(p,q)<\delta$,

  ColorSamplesSet ← RGB2Lab($p(x,y)$);
- Lightness and undersampling?
- Maximum likelihood:

$$ML \equiv argmax_{h \in \mathbf{H}} P(S|h)$$

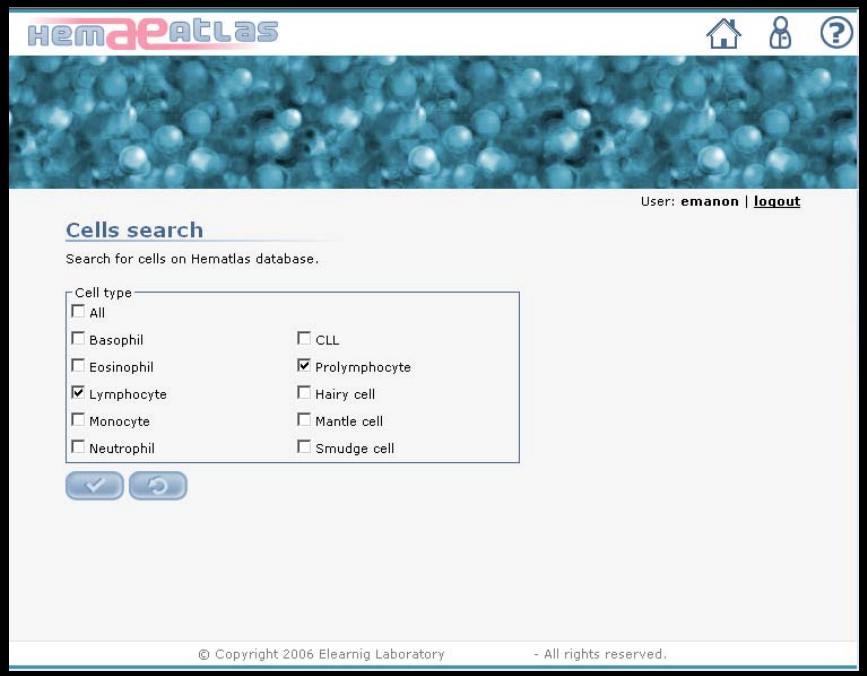

Original image
Evolution of the front
Original image
Evolution of the front
(a)
(d)
Original image
Evolution of the front
Original image
Evolution of the front
(b)
(e)
Original image
Evolution of the front
Original image
Evolution of the front
(c)
(f)

Ushizima et al., Leukocyte detection using nucleus contour propagation, LNCS 4091:389-396, 2006.



Hematlas

http://lsin.unisantos.br:8090/hematlas

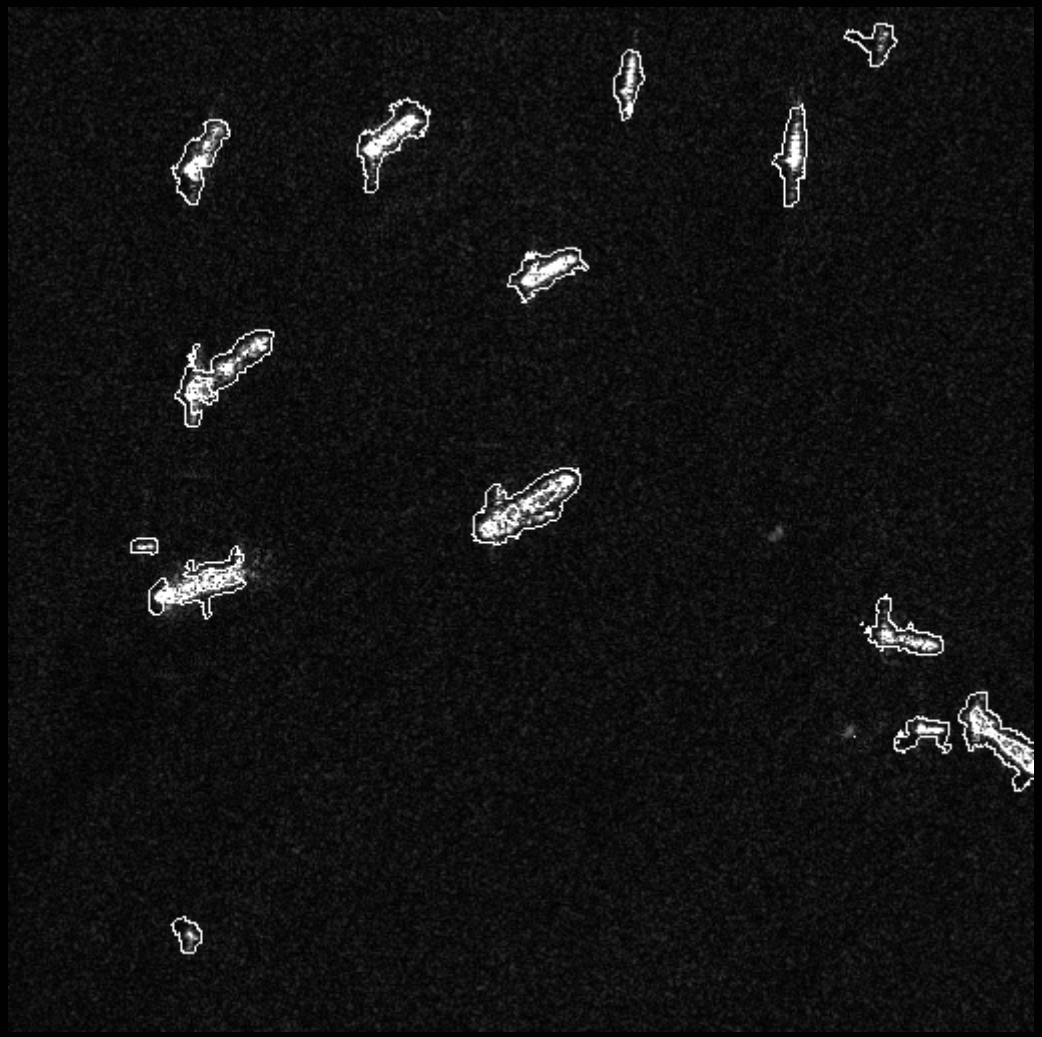


Acknowledgments:

- Unisantos: Profa. Marta Rosatelli, Daniella, Rodrigo, Aline, Rafael
- NIH: Dr. Edgar Rizzatti e Dr. Rodrigo Calado
- Instituto Fleury: Dr. Sérgio Martins
- Universidade de Sao Paulo: Prof. André Carvalho e Ana
- Universidade Federal de Santa Maria: Prof Marcos e Luciana
- Universidade Federal do Ceará: Profa Fátima Medeiros

Contact:

usdani@unisantos.br

http://www.unisantos.br/mestrado/informatica/daniela

Where is the ship?