

# Computer-Aided Cytogenetical Method of Breast Cancer Diagnosis. Part I - Decision Rule

**R.I.Andrushkiw**

*Department Mathematical  
Sciences and Center for  
Applied Mathematics and  
Statistics, New Jersey  
Institute of Technology,  
Newark, NJ, USA*

**Yu.I.Petunin,  
L.I.Ostapchenko**

*Kiev National  
Taras Shevchenko  
University, Kyiv,  
Ukraine*

**N.V.Boroday,  
Y.V.Lofovskaya,  
V.F.Chekhun**

*R.E.Kavetsky Institute of  
Oncology and Radiobiology  
of National Academy of  
Sciences of Ukraine*

**I.V.Dosenko**

*Institute of  
Oncology of  
Academy of  
Medical  
Sciences of  
Ukraine*

**Abstract.** *A Computer-aided cytogenetic method for the diagnosis of breast cancer is proposed. The approach is based on statistical analysis of indexes of interphase nuclei of buccal epitheliocytes, calculated with respect to their RGB-image after Feulgen staining.*

*Key words: breast cancer, fibroadenomatosis, buccal epithelium, discriminant analysis.*

## 1 Introduction

The computer-aided methods of cancer diagnosis, proposed by the authors in papers [1-3], were based on the investigation of morphological and densitometric indexes of interphase nuclei of buccal epitheliocytes. In this paper we describe a new approach, based on the analysis of RGB-image of these nuclei. Our purpose is to construct a filter that can be used to distinguish patients with breast cancer from those with fibroadenomatosis.

## 2 Material and methods

For our investigations we considered 68 patients suffering from breast cancer (BC), 33 patients suffering from fibroadenomatosis (FAM) and 30 healthy women (control). Each diagnosis was verified by histological

investigation of the removed tumor. The health of women in control group was verified by clinical examination. After gargling and removing the superficial cell layer of buccal mucous, we obtained smears from the median depth of the spinous layer from the patients' oral cavity. The smears were dried out under room temperature and fixed for 30 min in Nikiforov's mixture, followed by Feulgen staining with cold hydrolysis in 5 n. HCl for 15 min under temperature  $t=21-22^{\circ}\text{C}$ . Then, RGB-images (R-red, G-green, B-blue) were made of 30 to 100 typical nuclei, consisting of  $160 \times 160$  pixels. Finally, for every RGB-image we calculated 112 indexes: 25 vector and 87 scalar quantities. These indexes were calculated on the basis of RGB-images that were created using yellow and violet filters, and also without any filter. The first 25 vector indexes characterize the entropy distribution of the nuclei, the entire image of a cell, and the exterior of nuclei in 3, 4, 5, 6, 7 and 8-dimensional spaces, using confidence ellipsoids. In addition, some of these 3D-parameters are combinations of area, perimeter and form-factor. The other 87 indexes are scalar parameters that characterize the average entropy, curvature of spanning surfaces, and the

distribution of frequencies of some threshold levels of colors.

To identify the above indexes, let us introduce the following notation: *Ent* – entropy, *Nucleus* – parameter of *RGB*-image of nucleus, *Backg* – parameters of *RGB*-image of space outside of nucleus, *Total* – parameters of whole *RGB*-image (R red component, G green component, B blue component), *SC* – parameter of scanogram, *Area* – area of nucleus, *Perimeter* – perimeter of nucleus, *Fform* – form-factor, *CV* – curvature, *S* – standard deviation, *N* – without filter, *Y* – orange filter, *V* – violet filter, *MC* – modal classes, i.e. levels of the color (1, 2, ..., 255) for which the frequencies  $p_1$  and  $p_2$  of the pixels of the whole scanogram (of the nucleus only) having such color are calculated,

$$CI_1 = \frac{1}{(n-1)^2} \left( \sum_{i=1}^n \sum_{j=1}^{n-1} |s_{ij+1} - s_{ij}| + \sum_{i=1}^{n-1} \sum_{j=1}^n |s_{i+1j} - s_{ij}| \right),$$

$$CI_2 = \frac{1}{N_C} \left( \sum_{i \in Pr_x C} \sum_{j: (i,j) \in C} |s_{ij+1} - s_{ij}| + \sum_{j \in Pr_y C} \sum_{i: (i,j) \in C} |s_{ij+1} - s_{ij}| \right),$$

where  $N_C$  is the number of pixels in the scanogram,  $s_{ij}$  is an element of the scanogram,  $CI_1$  is the first curvature index characterizing surface curvature along  $x$  and  $y$  axis when whole scanogram is considered (both nucleus and background),  $CI_2$  is the curvature index of nucleus where  $Pr_x C$  is the projection of  $C$  on  $x$ -axis and  $Pr_y C$  is the projection of  $C$  on  $y$ -axis,  $C$  is a set of all pairs  $(i,j)$ , where  $i,j$ -th pixel belongs to the nucleus,  $MCVF1$  – the first modal class volume factor =  $p_1/p_2$ ,  $MCVF2$  – the second modal class volume factor =  $p_1/p_2$  (for pixels from nucleus), *R correct %* and *B correct %* are the percentages of scanograms with correctly built boundary for red and green components, respectively.

- 1 Ent N/G Nucleus + Ent N/G Backg + Ent N/G Total
- 2 Ent Y/G Nucleus + Ent Y/G Backg + Ent Y/G Total
- 3 Ent V/G Nucleus + Ent V/G Backg + Ent V/G Total
- 4 SC N/G Area + SC N/G Perimeter

- + SC N/G FForm
- 5 SC Y/G Area + SC Y/G Perimeter + SC Y/G FForm
- 6 SC V/G Area + SC V/G Perimeter + SC V/G Fform
- 7 CV Y/R  $CI_1$  + CV Y/R  $CI_2$  + CV Y/G  $CI_1$  + CV Y/R  $CI_2$
- 8 MC Y/R MCVF1 + MC Y/R MCVF2 + MC Y/G MCVF1 + MC Y/G MCVF2
- 9 Ent N/G Nucleus + Ent N/G Backg
- 10 Ent N/G Nucleus + Ent N/G Backg + Ent N/G Total
- 11 Ent N/R Nucleus + Ent N/R Backg + Ent N/R Total + Ent N/G Nucleus + Ent N/G Backg + Ent N/G Total
- 12 Ent Y/G Nucleus + Ent Y/G Backg
- 13 Ent Y/G Nucleus + Ent Y/G Backg + Ent Y/G Total
- 14 Ent Y/R Nucleus + Ent Y/R Backg + Ent Y/R Total + Ent Y/G Nucleus + Ent Y/G Backg + Ent Y/G Total
- 15 Ent V/G Nucleus + Ent V/G Backg
- 16 Ent V/G Nucleus + Ent V/G Backg + Ent V/G Total
- 17 Ent V/R Nucleus + Ent V/R Backg + Ent V/R Total + Ent V/G Nucleus + Ent V/G Backg + Ent V/G Total
- 18 SC N/R Area + SC N/R Perimeter + SC N/R FForm
- 19 SC N/G Area + SC N/G Perimeter + SC N/G FForm
- 20 SC N/B Area + SC N/B Perimeter + SC N/B FForm
- 21 SC Y/R Area + SC Y/R Perimeter + SC Y/R FForm
- 22 SC Y/G Area + SC Y/G Perimeter + SC Y/G FForm
- 23 SC V/R Area + SC V/R Perimeter + SC V/R FForm
- 24 SC V/G Area + SC V/G Perimeter + SC V/G FForm
- 25 SC V/B Area + SC V/B Perimeter + SC V/B FForm
- 26 Ent N/R correct %
- 27 Ent N/R Nucleus
- 28 Ent N/R Backg
- 29 Ent N/R Total
- 30 Ent N/G Nucleus
- 31 Ent N/G Backg
- 32 Ent N/G Total
- 33 Ent N/B correct %
- 34 Ent N/B Nucleus
- 35 Ent N/B Backg

36	Ent N/B Total	87	SC V/R Correct %
37	Ent Y/R correct %	88	SC V/R Area
38	Ent Y/R Nucleus	89	SC V/R Perimeter
39	Ent Y/R Backg	90	SC V/R FForm
40	Ent Y/R Total	91	SC V/G Correct %
41	Ent Y/G correct %	92	SC V/G Area
42	Ent Y/G Nucleus	93	SC V/G Perimeter
43	Ent Y/G Backg	94	SC V/B Correct %
44	Ent Y/G Total	95	SC V/B Area
45	Ent V/R correct %	96	SC V/B Perimeter
46	Ent V/R Nucleus	97	SC V/B FForm
47	Ent V/R Backg	98	CI <sub>1</sub> N/R Correct %
48	Ent V/R Total	99	CI <sub>1</sub> N/R X-bar
49	Ent V/G correct %	100	CI <sub>1</sub> N/R S
50	Ent V/G Nucleus	101	CI <sub>1</sub> N/G S
51	Ent V/G Backg	102	CI <sub>1</sub> V/B Correct %
52	Ent V/G Total	103	CI <sub>1</sub> V/R Correct %
53	Ent V/B correct %	104	CI <sub>1</sub> V/R X-bar
54	Ent V/B Nucleus	105	CI <sub>1</sub> V/G S
55	Ent V/B Backg	106	CI <sub>1</sub> V/B Correct %
56	Ent V/B Total	107	CI <sub>1</sub> Y/R Correct %
57	CV Y/R Correct %	108	CI <sub>1</sub> Y/R X-bar
58	CV Y/G Correct %	109	CI <sub>1</sub> Y/R S
59	CV Y/R CI <sub>1</sub>	110	CI <sub>1</sub> Y/G Correct %
60	CV Y/R CI <sub>2</sub>	111	CI <sub>1</sub> Y/G X-bar
61	CV Y/G CI <sub>1</sub>	112	CI <sub>1</sub> Y/G S
62	CV Y/G CI <sub>2</sub>		
63	MC Y/R Correct %		
64	MC Y/G Correct %		
65	MC Y/R MCVF1		
66	MC Y/R MCVF2		
67	MC Y/G MCVF1		
68	MC Y/G MCVF2		
69	SC N/R Correct %		
70	SC N/R Area		
71	SC N/R Perimeter		
72	SC N/R Fform		
73	SC N/G Area		
74	SC N/G Perimeter		
75	SC N/G FForm		
76	SC N/B Correct %		
77	SC N/B Area		
78	SC N/B Perimeter		
79	SC Y/R Correct %		
80	SC Y/R Area		
81	SC Y/R Perimeter		
82	SC Y/R FForm		
83	SC Y/G Correct %		
84	SC Y/G Area		
85	SC Y/G Perimeter		
86	SC Y/G FForm		

### 3 First stage of differential diagnosis

Let us denote the confidence ellipses for BC-patients by  $E_{BC}^{(k)}$ ,  $k = 1, \dots, 25$ , and the confidence ellipses for FAM-patients by  $E_{FAM}^{(k)}$ ,  $k = 1, \dots, 25$ . Let us denote the confidence intervals for healthy patients constructed by minimal and maximal order statistics by  $I_i = (\alpha_{\min}^{(i)}, \alpha_{\max}^{(i)})$ ,  $i = 1, \dots, 112$ , and the confidence intervals for healthy patients constructed by means of the 3s-rule by  $J_i = (\bar{x}_i - 3s_i, \bar{x}_i + 3s_i)$ ,  $i = 1, \dots, 112$  (for details see Part II of our paper).

At the first stage our objective was to identify only those patients who had BC. To accomplish this, we investigated FAM patients using the “leave-one-out” scheme, which showed that the number of indexes that fall outside the confidence ellipses  $E_{FAM}^{(k)}$ ,  $k = 1, \dots, 25$  varied from 0 to 3 for almost all FAM patients (for one patients this number was 5). Moreover, the number of patients’ indexes that

fell outside the remaining 87 confidence intervals was equal to 0 or 1. Thus, we can propose the following rule: if the number of patient's indexes falling outside the confidence ellipses  $E_{FAM}^{(k)}$   $k = 1, \dots, 25$  and  $(\alpha_{\min}^{(i)}, \alpha_{\max}^{(i)})$   $i=1, \dots, 112$  exceeds 3 and 1, respectively, then this patient suffers from BC. In the sample of 68 BC patients this rule was satisfied by 26 patients.

The remaining 42 patients did not satisfy these conditions. To identify BC patients in this group, we considered the confidence interval for indexes of healthy women. It turned out that the number of indexes that fell outside the FAM-patient's control confidence interval varied from 4 to 33, and for BC patients this number varied from 2 to 43. Therefore, we identified a patient as BC, if the number of the patient's indexes falling outside the above confidence interval exceeded 33. We identified 8 such patients. But among these patients only 4 were new, since the remaining 4 were included in the group of 26 patients mentioned above.

Further filtration was based on the confidence intervals for the control group constructed by 3s-rule. In this case the number of indexes of FAM-patient's falling outside the control confidence interval varied from 5 to 26, and for BC patients this number varied from 4 to 35. Therefore, we identified a patient as BC if the number of the patient's falling-out indexes exceeded 26. Following this procedure, we were able to identify 6 new patients that were not identified at previous stages.

Thus, applying the above three-stage filtration procedure to 68 patients, the correct diagnosis of BC was made in 36/68 patients (or 52.9%), and incorrect diagnosis was made for 1 patient (FAM was diagnosed as BC). No decision (rejection of decision) was made in the case of the remaining 31 patients.

## 4 Second stage of differential diagnosis

On the second stage we searched only for the FAM patients. At this stage we used the confidence ellipses  $E_{BC}^{(k)}$  and  $E_{FAM}^{(k)}$ ,  $k = 1, \dots, 25$ , and the confidence intervals  $I_{FAM}^{(k)}$  and  $I_{BC}^{(k)}$ ,  $k = 1, \dots, 25$ , constructed by 3s-rule.

To present these results, let us introduce the following notation:

$n_{FAM}$  — the number of patient's indexes that fall outside the confidence ellipses, constructed for vector indexes of FAM patients;

$n_{BC}$  — the number of patient's indexes that fall outside the confidence ellipses, constructed for vector indexes of FAM patients;

$m_{FAM}$  — the number of patient's indexes that fall outside the confidence ellipses, constructed for scalar indexes of FAM patients by 3s-rule;

$m_{BC}$  — the number of patient's indexes that fall outside the confidence ellipses, constructed for scalar indexes of BC patients by 3s-rule.

Consider the indexes  $l_{FAM} = n_{FAM} + m_{FAM}$  and  $l_{BC} = n_{BC} + m_{BC}$ . The rule for diagnosis of FAM has the following form: if  $l_{FAM} < l_{BC}$ , then patient has fibroadenomatosis, in all other cases we reject making a decision. Our investigation showed that for almost all BC patients (excluding one patient) the condition  $l_{FAM} \geq l_{BC}$  was satisfied. Hence, at the second stage we could not make any decision for almost all BC patients, and in one case we made the incorrect diagnosis (with probability 1/68). For the FAM patients we rejected to make a decision in 23 cases, and made 10 correct diagnoses.

Taking into account the number of BC and FAM patients with unconfirmed diagnosis, it is clear that in 56 of 101 cases (i.e. 55.4%) no decision was reached (rejection of decision). To make a decision in these cases, repetition of the analysis would have to be made on new smears from the patients.

## 5 Conclusion

Let us denote by  $H$  the null hypothesis (BC), and by  $H'$  the alternative competitive hypotheses (FAM). Using the formulas for calculating errors of type I and II, and the probability of rejection of decision (RD), we obtain the following table.

**Table.** Estimated probabilities of errors of type I and II corresponding to the number N of repetitive analyses

N	I Type	II Type	RD	Decision
1	1.4%	3.3%	55.4%	44.6%
2	2.2%	5.1%	30.7%	24.7%
3	2.6%	6%	17%	13.7%
4	2.8%	6.5%	9.4%	6%
5	2.8%	6.7%	5.2%	2.6%

Note that after 5 repetitions of the analyses described above, the correct diagnosis was obtained with probabilities of error of type I and II not exceeding 2.8% and 6.7%, respectively, and the probability of rejection of making a decision (RD) was obtained with probability not exceeding 5.2%.

## 6 References

[1] Yu.I. Petunin, D.A. Klyushin, K.P. Ganina, N.V. Boroday, R.I. Andrushkiw, "Computer-aided diagnosis of breast cancer based on analysis of malignancy associated changes in buccal epithelium", *Applied Statistical Sciences IV*, M.Ahnsanullah (ed), Nova Science Publ., pp. 181-204, 1999.

[2] Yu.I. Petunin, D.A. Klyushin, R.I. Andrushkiw, K.P. Ganina, N.V. Boroday, "Computer-aided differential diagnosis of breast cancer and fibroadenomatosis based on malignancy associated changes in buccal epithelium", *Automedica*, Vol.19, No 3-4, pp. 135-164, 2001.

[3] R.I. Andrushkiw, D.A. Klyushin, Yu.I. Petunin, N.V. Boroday, V.Lysyuk "Diagnosis of Breast Cancer by the Modified Nearest Neighbor Recognition Method", Proc. Int. Conf. on Mathematics and Engineering Techniques in Medicine and Biological Sciences, Vol.1, pp. 176-180, 202.