Novel Performance Metrics in the Classification of Microscopic Colonic Images

L. A. Gan Lim[1], R.N.G. Naguib[2,3], E.P. Dadios[4], J.M.C. Avila[5]

***Abstract*—Two classification metrics are proposed in this paper. The first one, called the mean relative difference confusion matrix, or MRDCM, is aimed at better quantifying the performance of a classifier that outputs a spectrum of real numbers. Usually, the tendency is to use threshold values to distinctly put classified values into rigid categories. While this approach has proven to be effective in many studies, sometimes it is best to leave the "classification" or the interpretation of actual output values of classifiers to a human expert. The MRDCM has been conceptualized with this idea in mind. The other classification metric that is proposed in this paper is aimed at improving the aggregation of values in a conventional confusion matrix to calculate the accuracy of classification. This other proposed novel metric, called the Classification Performance Index or CPI, includes in its calculation the consideration of both the correct and wrong classifications. These proposed metrics were applied to a classification of microscopic colonic images into three categories, namely: normal, adenomatous polyp, and cancerous. The results show agreement with the conventional confusion matrix and accuracy metric plus more information.**

***Index Terms*—Classifier metrics, colonic image classification, confusion matrix.**

[1] L. A. Gan Lim, Mechanical Engineering Dept., De La Salle University–Manila, Philippines (e-mail:)

[2,3] R.N.G. Naguib, Liverpool Hope University, U.K., BIOCORE Research and Consultancy International, U.K. (email:)

[4] E.P. Dadios, Manufacturing Engineering and Management Dept., De La Salle University–Manila, Philippines (e-mail:)

[5] J.M.C. Avila, Dept. of Pathology, University of the Philippines – Manila, Philippines (e-mail:)

I. Introduction

It has been reported in the Philippines that cancer ranks third among the leading causes of morbidity and mortality [1]. Worldwide, colorectal cancer is considered as the third most common neoplasm [2]. Similar to other types of cancers, early detection of cancer of the colon is the key to a successful treatment. It is, therefore, crucial to be able to have a reliable system to measure the performances of various classification or diagnostic tools being proposed to make a good selection and comparison in the developmental stages of these classification systems. Research in the classification of microscopic images of colonic mucosa has shown that textural features derived from grey-level co-occurrence matrices (GLCMs) are very useful. Similarly, with our previous work in [5] and [6], Esgiar et al. [7], Atlamazoglou et al. [8], Shuttleworth et al. [9], among others, GLCM were used as image property in classification. Several works have also been done to address the issue regarding the measurement of the performance of classifiers. In [9], for instance, Zachariah et al. compared 23 classifier metrics and concluded that the lift metric had the highest coefficient of determination. Hernandez-Orallo et al. [10] used a threshold choice method to link performance metrics and expected loss. Ferri et al. analyzed the performance of 18 different metrics [11] while Seliya et al. analyzed 22 metrics [12]. In 2009, Hand [13] proposed an alternative to area under the ROC curve. There are many more studies about classifier performance metrics that can be cited, but none have addressed the "right" compromise between simplicity and effectiveness.

Two classification metrics are proposed in this paper. The first one, called the mean relative difference confusion matrix (MRDCM), is aimed at better quantifying the performance of a classifier that outputs a spectrum of real numbers. Usually, the tendency is to use threshold values to put classified values into rigid categories distinctly. While this approach has proven to be effective in many studies, sometimes it is best to leave the classification or the interpretation of actual output values of classifiers to a human expert. The MRDCM

has been conceptualized with this idea in mind. The other classification metric that is proposed in this paper is aimed at improving the aggregation of values in a conventional confusion matrix to calculate the accuracy of classification. This other proposed novel metric, called the Classification Performance Index (CPI), includes in its calculation the consideration of both the correct and wrong classifications. These proposed metrics were applied to a classification of microscopic colonic images into three categories, namely: normal, adenomatous polyp, and cancerous. The classifier that was used in this study was adaptive network-based fuzzy inference system (ANFIS). Two textural properties, the sum average and difference entropy, derived from grey level co-occurrence matrix, were used. The results show agreement with the conventional confusion matrix and accuracy metric plus more information.

II. Preparation of Images

The images used in this study were derived from slides and cases randomly chosen from the 2007 and 2008 surgical pathology files of Medical Center Manila Hospital, previously diagnosed as colonic adenocarcinoma, adenomatous polyps from the colon, as well as resection planes of the colonic resections without tumor to serve as controls. These slides were routinely processed using a Sakura tissue processor and cut at 8 micra using a standard microtome. All were stained with hematoxylin and eosin. All images were taken at 400x magnification using an Olympus DP20 digital photomicrography apparatus mounted on an Olympus microscope (trinocular) at 1200x1800 dpi resolution. There were a total of 75 1600x1200-pixel-images (25 for each diagnosed case) used in this study. Shown in Figs. 1, 2, and 3 are sample images representative of each classification.

III. Textural Properties

A digital image can be represented as a matrix or set of matrices wherein each element contains numerical information about each pixel of the image. Texture can be defined as the mutual relationship among intensity values of neighboring pixels repeated over an area larger than the size of the relationship [3]. Haralick et al. [4] proposed textural features based on grey-level co-occurrence matrices or GLCMs. These features are effective in discriminating microscopic images of colon cancer tissues and cells.

For an Nx x Ny image I with each pixel quantized to Ng levels, let Lx be the horizontal spatial domain, Ly the vertical spatial domain, and G the set of quantized grey levels, such that Lx = {1,2,…Nx}, Ly = {1,2,…Ny}, and G = {1,2,…Ng}. The elements of a GLCM are then the relative frequencies Pij with which two neighboring pixels separated by distance d and angle Θ occur on the image, one with grey level i and the other with grey level j. With angles quantized to intervals of 45°, the un-normalized frequencies were defined in [4] as:

$$\mathbf{P}(i, j, d, 0^{\circ}) = \#\{[(k, l), (m, n)] \mid k - m = 0, |l - n| = d, I(k, l) = i, I(m, n) = j\} \tag{1a}$$

$$\mathbf{P}(i, j, d, 45^{\circ}) = \#\{[(k, l), (m, n)] \mid (k - m = d, l - n = -d), (k - m = -d, l - n = d), I(k, l) = i, I(m, n) = j\} \tag{1b}$$

$$\mathbf{P}(i, j, d, 90^{\circ}) = \#\{[(k, l), (m, n)] \mid |k - m| = d, l - n = 0, I(k, l) = i, I(m, n) = j\} \tag{1c}$$

$$\mathbf{P}(i, j, d, 135^{\circ}) = \#\{[(k, l), (m, n)] \mid (k - m = d, l - n = d), (k - m = -d, l - n = -d), I(k, l) = i, I(m, n) = j\} \tag{1d}$$

where # denotes the number of elements in the set. The co-occurrence matrix can be normalized by dividing each entry by the total number of pairs. In [4], 14 textural properties were introduced that are derivable from the GLCM. Below is a summary of the properties calculated from the normalized GLCM that were used in this study.

1) Sum Average:

$$f_7 = \sum_{i=2}^{2N_g} i p_{x+y}(i) \tag{2}$$

2) Difference Entropy:

$$f_{12} = \sum_{i=0}^{N_{g-1}} p_{x-y}(i) \log\left\{p_{x-y}(i)\right\} \tag{3}$$

Notation:

- $p(i, j)$ (i, j)-th entry in a normalized gray-tone spatial-dependence matrix, = P(i,j)/R.
- $p_x(i)$ i-th entry in the marginal-probability matrix obtained by summing the rows of $p(i, j) = \sum_{j=1}^{N_{g-1}} p(i, j)$.
- N_g Number of distinct gray levels in the quantized image.

$\sum_i$ and $\sum_j$ $\sum_{i=1}^{N_g}$ and $\sum_{j=1}^{N_g}$, respectively.

$$p_y(j)=\sum_{i=1}^{N_g} p(i,j).$$

$$p_{x+y}(k)=\sum_{\substack{i=1\\ i+j=k}}^{N_g}\sum_{j=1}^{N_g} p(i,j)$$

where k = 2, 3,…, $2N_g$

$$p_{x-y}(k)=\sum_{\substack{i=1\\ |i-j|=k}}^{N_g}\sum_{j=1}^{N_g} p(i,j)$$

where k = 0, 1, ….., N_g-1.

IV. The Mean Relative Difference Confusion Matrix

A commonly used tool to examine the performance of a classifier is the confusion matrix, which is a table of numbers of correct classifications and misclassifications. If one wants to produce a single number out of the confusion matrix as a measure of classification performance, the sum of the diagonals of the matrix is usually chosen and normalized to produce what is called the percent accuracy. The problem with this performance parameter is that it does not show the gravity of mistakes committed by the classifier in problems with more than two classes. For example, in this study where there are three classes of images—normal, adenomatous polyp, and cancerous cases—the percent accuracy will not yield information as to whether a cancerous case was misclassified as normal or as an adenomatous polyp. In "human" logic, it is less of a mistake to classify a cancerous case as adenomatous polyp than to classify it as normal. Erroneous downgrading from cancerous to normal can lead to a serious case not given enough scrutiny and is, therefore, the worst mistake that can be made by a classifier. As for the confusion matrix, although it is in itself an excellent tool to analyze the performance of a classifier, it is not directly compatible with the output of a classifier, such as ANFIS, that outputs a real number. The confusion matrix tabulates the counts (whole numbers) of classifications and misclassifications while ANFIS, since it is a Sugeno-type fuzzy inference system or FIS, generally gives out real numbers. Two alternatives can be adopted to fix this. One is to introduce threshold values for the output of ANFIS, and the other is to devise another classification performance matrix which can "handle" the ANFIS output values. The latter choice is more preferred in this study because it has the advantage of maintaining the spectral nature of histopathologic image classification and characterization. It is believed in this study that this approach is closer to how human pathologists view this kind of problem. Therefore, in this research, a new classification performance matrix, called the MRDCM, is proposed. The MRDCM tabulates the average differences of classification output values of the images and three constants defined by the following:

0.0 – for normal case
0.5 – for adenomatous polyp case, and
1.0 – for cancerous case.

Table I shows the general format of an MRDCM. Unlike the usual confusion matrix, the main diagonal elements of an MRDCM are ideally zero or close to zero since it is desired that the classification of the images should be correct and, therefore, have very small, if not zero, average differences with the ideal ANFIS output value for each case. For the off-diagonal elements, it is desirable to have non-zero values close to 0.5 or 1.0.

TABLE I
GENERAL FORMAT OF AN MRDCM. THE ELEMENTS A, E, AND I ARE THE MAIN DIAGONAL ELEMENTS. THE REST OF THE ELEMENTS ARE THE OFF-DIAGONAL ELEMENTS.

	Expected Normal	**Expected Aden. Polyp**	**Expected Cancerous**
Predicted Normal	a	b	c
Predicted Aden. Polyp	d	e	f
Predicted Cancerous	g	h	i

Each element in the matrix can be expressed as:

$$x_{ij}=\frac{\sum_{k=1}^{n_j}\left|o_j(k)-c_i(k)\right|}{n_j}$$

where

x_{ij} = element in the matrix at row i and column j
$o_j(k)$= ANFIS output value for image k at class j
n_j = total number of images in class j

$$c_i(k)=\begin{cases}0.0 & if\ i=1\\ 0.5 & if\ i=2\\ 1.0 & if\ i=3\end{cases}$$

V. The Classification Performance Index

In optimizing classifiers, it would be very advantageous to be able to express the performance of a classifier into a single number or scalar just like the percent accuracy of a confusion matrix. As pointed out earlier, the percent accuracy parameter does not take into account the gravity of the misclassifications of a classifier for problems with more than two cases. The new idea that is being proposed in this study is to introduce a parameter called the CPI that precisely brings with it the information conveyed by percent accuracy plus additional measures of classification failures. The CPI metric is arrived at by first calculating the corresponding confusion matrix using threshold values for the adenomatous polyp and cancerous cases and normalizing the elements by using the sum of elements per column or class as a divisor. Next, the confusion matrix with normalized elements is then multiplied element-wise by a new matrix referred to here as factor matrix, which is essentially a weight matrix. The product, which is sometimes referred to as Hadamard or Schur product in matrix multiplication, is another matrix similar in size to the confusion matrix and the factor matrix. The factor matrix contains elements that act as multipliers similar to connection weights in a feed-forward neural network. Finally, the CPI parameter is calculated as the sum of all the elements of the element-wise product of the normalized confusion matrix and the factor matrix. The idea behind the factor matrix is to select specific real numbers as elements that will seek proportional contributions of the specific elements of the confusion matrix to the CPI parameter. To make the CPI reflect the failure-to-success spectrum of a classifier, the entries in the factor matrix must be selected to get more contribution from the successes and less from the failures in the numbers tabulated in the confusion matrix. This was accomplished in this study by suggesting a ranking of the elements of the confusion matrix according to the degree of success and gravity of failure of the classifier expressed as a set of multiplying factors. Tables 2, 3, and 4 show the format of the confusion matrix used in this study, the format of the factor matrix, and the suggested ranking of the corresponding elements according to a set of multiplying factors, respectively.

TABLE II
Format of the Confusion Matrix Used in this Study

	Expected Normal	**Expected Aden. Polyp**	**Expected Cancerous**
Predicted Normal	A	B	C
Predicted Aden. Polyp	D	E	F
Predicted Cancerous	G	H	I

TABLE III
Format of the Factor Matrix

A	B	C
D	E	F
G	H	I

The letters assigned to each element of the matrix correspond to the left column of Table IV and the entries in Table II as multipliers.

TABLE IV
The Suggested Ranking of the Elements of the Factor Matrix with the Multiplying Factors

Location in the factor matrix	**Multiplying factor**
i	+1/3
e	+1/3
a	+1/3
d	-0.05
h	-0.1
g	-0.2
b	-0.3
f	-0.4
c	-0.5

Match the letters on the left column to the entries in Table III.

It can be observed that the multiplying factors in Table III together produce an effect on the CPI wherein the positive and negative factors counteract each other when multiplied by the confusion matrix. The entries i, e, and a get +1/3 each since the numbers in these locations in the confusion matrix represent the correct classifications. Their factors have been purposely chosen to sum-up to 1.0 or 100% because they represent the perfect score. The rest of the entries are all assigned negative factors, representing a penalty against the CPI because they are the multipliers of the off-diagonal elements of the confusion matrix. It can be observed that the factors in entries c, g, and b, all sum-up to -1.0 or -100% which is considered to be the exact opposite of a perfect score in classification in this study. Entry c is assigned the greatest penalty effect because it corresponds to the worst possible mistake that can be committed by a classifier, which is a misclassification of cancer into normal. Since entry f is considered as between entries c and b, therefore, c = –0.5, b = –0.3, g = –0.2 and f = –0.4. Entry d is considered here as the element in the factor matrix that corresponds to the least serious misclassification wherein a truly normal case is classified as adenomatous polyp by mistake while entry

h had to be just worse than entry d. With g = -0.2 and a = 1/3, entries h and d had to assume -0.1 and -0.05 values, respectively. Therefore, Table 4 suggests that the factor matrix should be expressed as in equation 4.

$$FM = \begin{bmatrix} +\frac{1}{3} & 0.3 & -0.5 \\ -0.05 & +\frac{1}{3} & -0.4 \\ -0.2 & -0.1 & +\frac{1}{3} \end{bmatrix} \quad (4)$$

where: FM = factor matrix.

Putting it all together now, the CPI can be calculated by first getting the entry-wise product of the confusion matrix and the factor matrix, and then obtaining the sum of all the elements of the resulting matrix. Mathematically, for three classes can be expressed as:

$$CPI = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{CM_{ij} FM_{ij}}{N_j} \quad (5)$$

where:

CPI = classification performance index
CM_{ij} = entry in the confusion matrix at row i and column j
FM_{ij} = entry in the factor matrix at row i and column j
N_j = total number of elements in class or column j.

I. ANFIS Classification and Results

The classification of images using ANFIS was evaluated using three tools: the conventional confusion matrix, the MRDCM, and the CPI. More discussion about the ANFIS implementation can be found in our earlier paper in [5]. Table 6 shows that the classifier was quite successful. There were no misclassifications in the normal-cancerous conditions. However, there were some mistakes in the middle region—the adenomatous polyp cases. This was to be expected since the middle region is the most difficult part of the classification task. This is also reflected in the MRDCM in Table 5. Ideally, the number in the main diagonal of an MRDCM should all be zero. The numbers in the main diagonal of Table 5, therefore, means that the classification performance was very good because the numbers are close to 0.0. At first glance, one would be more comfortable using the conventional confusion matrix like the one shown in Table 6. However, the convenience that comes with analyzing Table 6 comes at the price of choosing threshold values 0.25 and 0.75. Adjusting these values would surely affect the numbers in the matrix of Table 6. The selection of the threshold values, therefore, puts a limitation or uncertainty in the interpretation of the confusion matrix. This is the price to pay for forcing to choose threshold values. The MRDCM does not share this problem and is, therefore, less biased. The numbers in the MRDCM can easily be mapped into a grey color spectrum to make it easier to be interpreted by a human pathologist or oncologist. In contrast to a conventional confusion matrix, the values of the off-diagonal elements in Table 5 are "far" from 0.0, indicating that misclassifications were minimal. Ideally, the elements closest to the main diagonal should be 0.5 while the elements farthest should be close to 1.0.

The CPI data in Table 6 shows that the calculated accuracy values overestimate the performance of the classifier. When the CPI values are converted into percentages, it is clear that these values are lower than their corresponding accuracy values. This is obviously a result of the conventional accuracy formula only considering the main diagonal elements.

TABLE V
MRDCM FOR TRAINING AND TESTING DATA SETS USING FEATURES SUM AVERAGE AND DIFFERENCE ENTROPY

	Training Data Set			Testing Data Set		
	Expected Normal	**Expected Aden. Polyp**	**Expected Cancerous**	**Expected Normal**	**Expected Aden. Polyp**	**Expected Cancerous**
Predicted Normal	0.0728	0.5434	0.894	0.0641	0.4829	0.882
Predicted Aden. Polyp	0.4374	0.1688	0.394	0.4471	0.1824	0.382
Predicted Cancerous	0.9374	0.4566	0.1136	0.9471	0.5171	0.1245

TABLE VI
Confusion Matrix, Percent Accuracy, and CPI for Training and Testing Data Sets Using Features Sum Average and Difference Entropy with Threshold Values of 0.25 and 0.75

	Training Data Set			Testing Data Set		
	Expected Normal	**Expected Aden. Polyp**	**Expected Cancerous**	**Expected Normal**	**Expected Aden. Polyp**	**Expected Cancerous**
Predicted Normal	64	4	0	28	3	0
Predicted Aden. Polyp	6	56	8	2	24	3
Predicted Cancerous	0	10	62	0	3	27
Percent Accuracy		86.6667%			87.7778%	
Classification Performance Index		0.7852			0.7944	

VII. Conclusion

The purpose of MRDCM is to allow clinicians or pathologists to make use of the real number output of the ANFIS classifier and thereby avoid the use of threshold values to characterize an image. The CPI, on the other hand, is considered here as a better alternative to the percent accuracy parameter when expressing the classification quality reflected by a confusion matrix. The CPI utilizes a set of numbers called factor matrix that collectively imposes a kind of penalty to elements in the confusion matrix that represent bad classification performance. It was pointed out that one of the disadvantages of using the percent accuracy is that it does not distinguish between bad and worse misclassifications. An example of this is misclassification of cancerous into adenomatous polyp compared to misclassification of cancerous into normal. Unlike the percent accuracy parameter, the CPI puts more "penalty" on the latter case of misclassification.

MRDCM was a natural extension of the ANFIS classifier since the conventional confusion matrix could not be used with the ANFIS output without resorting to selecting threshold values for the three classes. This novel confusion matrix supported the idea of providing the human pathologist or the user with more information by pointing out the state of the image in question relative to extreme cases in the normal-to-cancerous spectrum. Thinking in terms of numbers in the classification spectrum might promote more objectivity on the part of the pathologist. The classifier algorithm/s developed in this study was never meant to replace human experts but rather be used as effective supporting tools.

Sometimes it is necessary to express the performance of a classifier through a single number such as the percent accuracy computed from the conventional confusion matrix. CPI was devised in this study as a better parameter than percent accuracy and is simple to use. The advantage of using CPI is in its ability to account for the successes and severity of failures of a classifier. The CPI parameter achieved this through the use of a novel matrix known as the factor matrix, also devised in this study.

References

[1] C. A. Ngelangel and E. H. Wang, "Cancer and Philippine cancer control program," *Japanese Journal of Clinical Oncology*, vol. 32, supp. 1, pp. S52-S61, 2002.

[2] J. K. Shuttleworth, A. G. Todman, R. N. G. Naguib, R. M. Newman, and M. K. Bennett, "Enhancing feature extraction from colon microscopy images using colourspace rotation," in P*roc. Medical Image Understanding and Analysis Conf.*, Bristol, UK, 2005, pp. 11–14.

[3] A. D. Kulkarni, *Computer Vision and Fuzzy-Neural Systems*. Upper Saddle River, NJ. Prentice Hall PTR, 2001.

[4] R. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification." *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC- 3, no. 6, pp. 610–621, 1973.

[5] L. A. Gan Lim, R. N. G. Naguib, E. P. Dadios, and J. M. C. Avila, "Implementation of GA-KSOM and ANFIS in the classification of colonic histopathological images," in *Proc. of IEEE TENCON*, Cebu City, Philippines, 2012. doi: 10.1109/TENCON.2012.6412240.

[6] L. A. Gan Lim, R. N. G. Naguib, E. P. Dadios, and J. M. C Avila, "Image classification of microscopic colonic images using textural properties and KSOM," *International Journal of Biomedical Engineering and Technology (IJBET)*, vol. 3, no. ¾, pp. 308–318, 2010.

[7] A. N. Esgiar, R. N. G. Naguib, B. S. Sharif, M. K. Bennett, and A. Murray, "Texture descriptions and classification for pathological analysis of cancerous colonic mucosa," in *IEE Proc., Int. Conf. on Image Processing and its Applications*, University of Manchester, UK, vol. 1, 1999, pp 335–338.

[8] V. Atlamazoglou, D. Yova, N. Kavantzas, and S. Loukas, "Texture analysis of fluorescence microscopic images of colonic tissue sections," *Medical and Biological Engineering & Computing*, vol. 39, no. 2, pp. 145-151, 2001.

[9] N. Zachariah, S. Kothari, S. Ramamurthy, A. O. Osunkoya, and M. D. Wang, "Evaluation of performance metrics for histopathological image classifier optimization," 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6943990/.

[10] J. Hernandez-Orallo, P. Flach, and C. Ferri, "A unified view of performance metrics: Translating threshold choice into expected classification loss," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2813–2869, 2012.

[11] C. Ferri, J. Hernandez-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters*, vol. 30, no. 1 pp. 27–38, 2009.

[12] N. Seliya, T. M. Khoshgoftaar, and J. V. Hulse, "A study on the relationships of classifier performance metrics," in *21st IEEE International Conference on Tools with Artificial Intelligence***,** Newark, New Jersey, 2009. [Online]. Available: http://ieeexplore.ieee.org/document/5364367/.

[13] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," *Machine Learning*, vol. 77, no. 1, pp. 103–123, 2009. [Online]. Available: http://link.springer.com/article/10.1007/s10994-009-5119-5.