Learning the topological properties of brain tumors

Cigdem Demir[1], S. Humayun Gultekin[2,*] and Bülent Yener[1]

[1]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180,

[2]Department of Pathology, Mount Sinai Medical School, New York, NY 10021

Abstract:

Different types of feature representation have been investigated to represent the histopathological images for the purpose of cancer diagnosis. In this work, we demonstrate that cell-graphs provide effective representations as they encode the pairwise relation between every cell by statistically assigning a link between them. Working with photomicrographs of 646 archival brain biopsy samples from 60 patients, we show that without this pairwise relation, neither the spatial distribution of the cells nor the texture analysis of the images yields as accurate results as in the case of the cell graphs to distinguish cancerous tissues from non-cancerous tissues with similar cellular density levels. We use the global graph metrics that are defined on the entire cell-graph as a feature set of a multilayer perceptron for the tissue level diagnosis of a brain cancer called malignant *glioma*. In our experiments, we correctly classify the cancerous and healthy brain tissue samples that have significantly different cellular density levels with accuracy greater than 99 %. Furthermore, we accomplish distinguishing the cancerous tissues from non-neoplastic reactive/inflammatory conditions that may reveal an equally high cellular density; with an accuracy of at least 92 %.

* Present address: Oregon Health and Science University, Department of Pathology, Portland, OR

Index terms: Image representation, machine learning, model development, graph theory, medical information systems.

1. Introduction

Automated classification of the histopathological images has been extensively studied for cancer diagnosis. These studies make use of different classifiers that employ a subset of different types of features. For example, a large subset of these studies uses feature sets that typically consist of the morphological features such as the area, perimeter, and roundness of a nucleus [5, 9, 10, 12, 16, 17, 18, 20, 22, 23] and/or the textural features such as the angular second moment, inverse difference moment, dissimilarity, and entropy derived from the co-occurrence matrix [5, 6, 10, 13, 19, 20, 22]. These studies train their systems to distinguish the healthy and cancerous tissues using artificial neural networks [19, 20, 23], k-nearest neighborhood algorithm [6, 9], support vector machines [10], linear programming [17], logistic regression [22], fuzzy [16], and genetic [18] algorithms. Complimentary to the morphological and textural features, a few of these studies use colorimetric features such as the intensity, saturation, red, green, and blue components of pixels [9, 23] and densitometric features such as the number of low optical density pixels in an image [6, 13, 19].

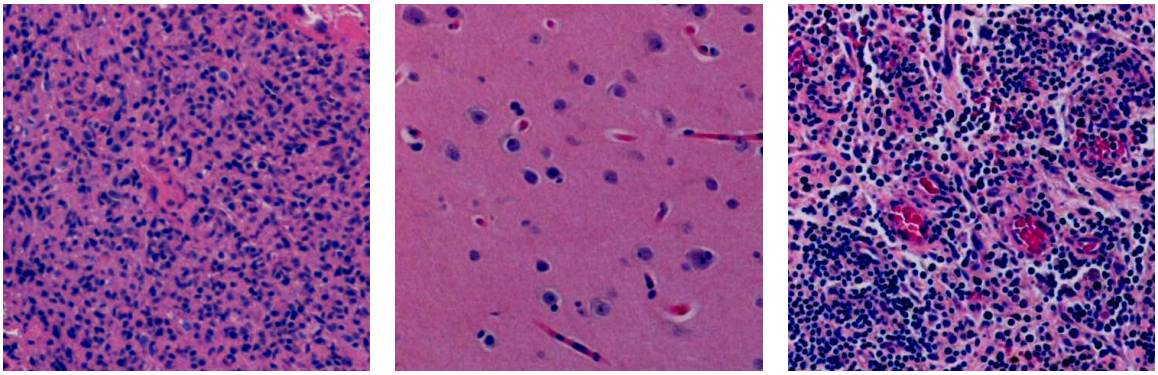
Another subset of these studies uses fractals that describe the similarity levels of different structures found in a tissue image over a range of scales [4, 7]. These studies use the fractal dimensions as their features and use k-nearest neighborhood algorithm [7], neural networks and logistic regression [4] as their classifiers. Finally, the orientational features are extracted by

making use of Gabor filters that respond to contrast edges and line-like features of a specific orientation [21].

Recently, we have demonstrated that the use of cell-graphs generated from the tissue images according to the spatial distribution of the cells leads to successful tissue diagnosis of cancer [11]. In the generation of such graphs, the nodes correspond to the cells and the probability of a link between a pair of nodes is calculated as a decaying exponential function of the Euclidean distance between this node pair. We have showed that the topological features defined on each node of this cell-graph, i.e., the **local graph metrics**, carry characteristic properties to distinguish the images of cancerous brain tissues from those of healthy or non-neoplastic primary inflammatory processes (herein referred to as "inflamed tissues"). The work in [11] introduces a novel method of feature extraction for histopathological images and it proposes to use a machine learning algorithm that employs this representation for the diagnosis of cancer.

In this work, as our first contribution, we compare the cell-graph approach with two different approaches; the first approach uses only the spatial distribution of the cells without defining links and the other approach uses textural features. We demonstrate that the cell-graphs provide an effective tool to represent tissue images not only because they encode the spatial distribution of the cells, but also they encode the **pairwise** relation between the cells by assigning a **link** between them. Our experiments show that this pairwise relation is crucial in obtaining a high classification accuracy to distinguish different types of tissue images, even when they have similar levels of cellular density. For example, although the spatial distribution of cells alone provides sufficient information to distinguish the *cancerous tissues* with higher cellular density

(as shown in Figure 1a) from the *healthy tissues* with lower cellular density (as shown in Figure 1b), it is not sufficient to distinguish the *cancerous tissues* from *inflamed tissues* (as shown in Figure 1c) whose cellular density is equally high. However, the cell-graphs defining links based on the pairwise relation between every cell successfully distinguish the cancerous tissues from both healthy and inflamed tissues regardless of their cellular density. This demonstrates that the cell-graph approach provide further information in the classification of different types of tissues with different cellular density levels. Moreover, in the distinction of cancerous – noncancerous tissues, we compare the accuracy of the classifier that uses the cell-graph representation with the accuracy of the classifier that employs textural features. While the cell-graph representation encodes the pairwise relation between the cells, the textural features reflect the spatial interrelationships of pixel gray values. Although the classifier employing the textural features is as accurate as in the case of the cell-graph approach to distinguish the cancerous and healthy tissues, it yields lower accuracy values than the cell-graph approach to distinguish the cancerous and inflamed tissues.



(a) (b) (c)

Figure 1: Microscopic images of brain biopsies stained with hematoxylin and eosin technique: (a) a brain tumor sample (i.e. glioma), (b) a healthy tissue sample, and (c) an inflamed tissue sample.

As our second contribution, we extend the work in [11] to study the topological properties defined on the entire graph, i.e., the **global graph metrics**. While the local metrics provide information at the cellular level, the global graph metrics provide information at the tissue level. We use the global metrics as our feature set and the artificial neural networks as our classifier in the diagnosis of a brain tumor called malignant *glioma*. These global graph metrics include the average degree, the clustering coefficient, the average eccentricity, the ratio of the giant connected component, the percentage of the end nodes, the percentage of the isolated nodes, the spectral radius, and the eigen exponent.

In our experiments, we use a total of 646 images of tissue samples surgically removed from 60 different patients. For the training set, we use 211 images of tissues taken from 22 different patients. For the testing set, we use 435 images of tissues taken from the remaining 38 patients; these images are not used in the training set at all. In the classification of cancerous and healthy tissues only, we achieve 99.75 % accuracy on the testing set by using the global graph metrics. Similarly, we achieve 99.75 % and 98.46 % classification accuracies on the testing set by using information extracted from the spatial distribution of the cells (cell-distribution) and by using information extracted from texture analysis (texture-based), respectively. On the other hand, in the classification of cancerous, healthy, and inflamed tissues of the testing set, the cell-graph approach leads to accuracy levels of 96.30 % for healthy tissues, 94.00 % for cancerous tissues, and 92.19 % for inflamed tissues, whereas the cell-distribution approach yields only accuracy levels of 98.34 % for healthy tissues, 71.79 % for cancerous tissues, and 42.97 % for inflamed tissues at most and the texture-based approach yields accuracy levels of 96.10 % for healthy tissues, 89.60 % for cancerous tissues, and 79.38 % for inflamed tissues. Therefore, the cell-

distribution and texture-based approaches increase the false negatives and false positives. The significant difference in the classification results of cancerous and inflamed tissues using the cell-graph and cell-distribution approaches suggests the significance of the pairwise relation between the cells (i.e., the links between them), and hence, indicating the effectiveness of the edges in the cell-graph. The difference in the classification results of cancerous and inflamed tissues using the cell-graph and texture-based approaches suggests the importance of the existence of this pairwise relation between the cells, not for example between the gray values of pixels, and hence, indicating the effectiveness of the nodes in the cell-graph.

The remaining of this paper is organized as follows. In Section 2, we briefly explain the methodology to generate a graph from a tissue image and the definitions of the global graph metrics that quantify the topological properties of the generated graphs. In Section 3, we present experimental results. Finally, we provide a summary of our work in Section 4.

2. Methods

2.1. The cell-graph generation

This technique relies on the clustering information of cells in the tissue. We first generate a mesh of cells based on the locations of the cells in their two-dimensional tissue image. Then we generate a graph by establishing links according to the Euclidean distance between every pair of these cells. This resulting cell-graph represents the tissue image and the topological properties of

the cell-graph are exploited in the classification of different tissue images. This technique is summarized below; the details can be found elsewhere [11].

The first step is **color quantization**. In this step, we learn how to distinguish the cells from their background based on the color information of the pixels. For that, we use k-means algorithm [14] to cluster the pixels of training samples and learn the clustering vectors. After that, each of these clustering vectors is assigned to be either “cell” or “background” class by a pathologist. We use these clustering vectors and their class assignments in the classification of the pixels of testing images as either “cell” or “background”.

The next step is **node identification** where we translate the class information of the pixels to the node information of a cell-graph. For that, we put a grid over the tissue image and for each grid entry, we compute a probability of being a cell as follows: First, we assign a value of 1 to the pixels of “cell” class and a value of 0 to the pixels of “background” class and we then compute the average of the values of the pixels located in this grid entry as its probability value. Grid entries with probability values greater than a threshold are considered as the nodes of a cell-graph. In this step, a node can represent a single cell, a part of a cell, or bunch of cells depending on the grid size. Because of that, the topological features extracted using this method do not require high magnification images to resolve the details of a cell in contrast with the morphological features.

The last step is **link establishing**. In this step, we set the links between the nodes identified in the previous step to generate a cell-graph. Therefore, in this step, we translate the pairwise spatial

relation between every two nodes to the possible existence of links in a cell-graph. The probability of an existence of a link between the nodes u and v is given by $P(u,v) = d(u,v)^{-\alpha}$, where $d(u,v)$ is the Euclidean distance between the nodes u and v, and α is the exponent that controls the density of a graph. This probability quantifies the possibility for one of these nodes to be grown from the other. Thus, the links of the cell-graph model the prevalence of cancer. More formally, suppose $G = (V, E)$ be a generated cell-graph with V and E being the set of nodes and links of the graph, respectively. After determining V in the node identification step, we define a binary relation E on V such that $E = \{(u,v) : r < d(u,v)^{-\alpha}, \forall u, v \in V\}$, where r is a real number between 0 and 1 that is generated by a random number generator.

We use the topological properties extracted from the resulting cell-graph as the feature set of the corresponding tissue image. In the classification of a tissue, we use artificial neural networks [1, 15], where the inputs are these topological properties extracted from the cell-graphs and the output is whether the tissue is cancerous, healthy, or inflamed.

2.2. The global graph metrics

In this work, we use eight different topological properties defined on the entire graph (i.e., the **global graph metrics**), namely the average degree, the clustering coefficient, the average eccentricity, the ratio of the giant connected component, the percentage of the end nodes, the percentage of the isolated nodes, the spectral radius, and the eigen exponent.

(1) The degree of a node is defined as the number of its links. Using the distribution of the node degrees, we compute the **average degree** as a global metric.

(2) The clustering coefficient C_i of a node *i* is defined as $C_i = (2 \cdot E_i)/(k \cdot (k+1))$, where *k* is the number of neighbors of the node *i* and E_i is the number of existing links between its neighbors [3]. This metric quantifies the connectivity information in the neighborhood of a node. We use the **average clustering coefficient** as a global metric.

(3) The eccentricity of a node *i* is the length of the maximum of the shortest paths between the node *i* and every other nodes reachable from *i*. We use the **average eccentricity** as a global metric.

(4) The giant connected component of a graph is the largest set of the nodes where all of the nodes in this set are reachable from each other. We use **the ratio of the size of the giant connected component** over the size of the entire graph as a global metric.

(5) A node in a graph is an "isolated node" if it does not have any neighbors, i.e., if it has a degree of 0. A node in a graph is an "end node" if it is connected to a single node, i.e., if it has a degree of 1. We use the **percentages of the isolated and the end nodes** in the entire graph as global metrics.

(6) The last two metrics are related to the spectrum of a graph, which is the set of graph eigenvalues (i.e., eigenvalues of the adjacency matrix of a graph). The spectrum of a graph is closely related to the topological properties of a graph such as the diameter, the number of the connected components and the number of spanning trees [2]. In this work, we use the **spectral radius**, which is defined as a maximum absolute value of eigenvalues in the spectrum, as a global metric. The **eigen exponent** is defined as the slope of the sorted eigenvalues as a function of their orders in log-log scale [8]. As our

last global metric, we use the eigen exponent computed on the first largest 50 eigenvalues of each graph.

3. Experiments

3.1. Methodology

In our experiments, we use a data set that consists of 646 microscopic images of brain biopsy samples of 60 randomly chosen patients from the pathology archives. All patients were adults with both sexes included. This data set includes samples of 41 cancerous (glioma), 14 healthy and 9 reactive/inflammatory processes; for 4 of these patients, we have both cancerous and healthy tissue samples. The training data set consists of 211 images taken from 22 different patients and the testing data set consists of 435 images taken from the remaining 38 patients. Each sample consists of a 5-6 micron-thick tissue section stained with hematoxylin and eosin technique and mounted on a glass slide[1]. The images are taken in the RGB color space with a magnification of 100X and each image consists of 480x480 pixels. After taking the images, we convert the RGB values of the pixels into their corresponding values into the La*b* color space. Unlike the RGB color space, the La*b* color space is a uniform color space and the color and detail information are completely separate entities. Therefore, using the La*b* color space yields better quantization results in our experiments. We cluster the La*b* values of the pixels using k-

[1] The identifiers were removed, and slides were numerically recoded corresponding to diagnostic categories by the pathologist, prior to obtaining digital images of the tissues. Therefore, the remaining two investigators had access to images and diagnoses only, without retraceable personal identifiers.

means algorithm, where the value of k is 16. We observe that the values of k greater than 16 do not introduce a significant change in the resulting processed images.

In identifying the nodes of the cell-graph, we have two control parameters: the grid size and the probability threshold. We select a grid size of 6 that matches the size of a typical cell in the magnification of 100X. The probability threshold determines the density of the nodes in a cell-graph. A larger threshold produces sparser graphs, whereas a smaller threshold makes the assignment of the nodes more sensitive to the noise arising from misassignment of "cell" classes in the color quantization step. Therefore, we choose a reasonable threshold value of 0.25 that yields dense enough graphs eliminating the noise. In establishing the links of the cell-graph, we use an exponentially decaying probability function with an exponent of $-\alpha$ with $0 \leq \alpha$. The value of α determines the density of the links in a cell-graph; larger values of α produce sparser graphs. α values close to 0 produce the graphs that are almost connected and from such graphs, it is not possible to extract the distinguishing topological properties. On the other hand, as the value of α increases, the resulting graphs can have only a few links and from such graphs, it is also not possible to compute the distinguishing properties. Considering these, we choose α to be 3.6 that produces dense enough graphs to capture the distinguishing properties of these graphs.

To compare with the cell-graph approach and investigate the significance of encoding the pairwise relation between the nodes, we separately use (i) features extracted from the spatial distribution of the cells that do not include any link information and (ii) textural features derived from the gray-level co-occurrence matrix in the classification of different tissues.

We extract the features of the cell-distribution approach that only uses the spatial distribution of the cells as follows: After the node identification step, we embed a grid on the resulting mesh of nodes instead of establishing links and extracting a cell-graph. For each grid entry, we average the values of the mesh entries located in a particular grid entry of interest, assigning a value of 1 to each mesh entry consisting of a node and a value of 0 otherwise. Then we use these average values of grid entries as the feature set of the cell-distribution approach.

The co-occurrence matrix C computed on a gray-level image P is defined by a distance d and an angle θ. $C(i,j)$ indicates how many times the gray value i co-occurs with the gray value j in a particular spatial relationship defined by d and θ. Mathematically, it is given as $C(i,j) = \left|\{m,n\} : P(m,n) = i \text{ and } P(m + d\cos\theta, n + d\sin\theta) = j\right|$. We compute 12 different normalized gray-level co-occurrence matrices at four different angles (0, 45, 90, and 135^{o}) and three different distances (1, 5, and 9). On each normalized co-occurrence matrix, we compute six different features, including the angular second moment, the contrast, the correlation, the inverse difference moment, the dissimilarity, and the entropy. More on these features and their derivations can be found in [6].

In the classification of the images of tissue samples, we use multilayer perceptrons. For each classifier, which uses features extracted from using the cell-graph, cell-distribution or texture-based approaches, we choose the number of hidden units as to optimize the classification accuracies. We train each classifier on 20 different runs and the results presented in this section are obtained averaging the accuracies over these runs.

3.2. Results

3.2.1. Classification of cancerous and healthy tissues

In this subsection, we provide the accuracy of each approach in the classification of cancerous and healthy tissues that have dense and sparse cellular density levels and compare these accuracy values. In Table 1, we report the average accuracy values and their standard deviations obtained in the classification of the cancerous and healthy tissue images by using cell-graphs. In addition to the overall accuracy obtained on the entire data set (including both cancerous and healthy tissues), we report the accuracy levels for each class type. This table shows that using topological properties of the cell-graphs extracted from the tissue images, all samples in the training set are correctly classified. It also indicates that the topological properties of the cell-graphs distinguish the cancerous and healthy tissues in the testing set with a high level of accuracy >99 %.

Table 1: Cancerous and healthy tissue classification using the **cell-graph** approach.

	Training set accuracy	Testing set accuracy
Overall	100.00 ∓ 0.00	99.75 ∓ 0.00
Cancerous	100.00 ∓ 0.00	100.00 ∓ 0.00
Healthy	100.00 ∓ 0.00	99.35 ∓ 0.00

In the cell-distribution approach, the grid size determines the dimension of the extracted feature set. Since the dimension of the mesh for the images used in this work is 80x80, we choose the grid size ranging from 1 to 20. For different exemplary values of the grid size, the average

classification accuracy levels and their standard deviations obtained using the cell-distribution approach are presented in Table 2 and Table 3 for the training and testing data sets, respectively. For the texture-based approach, the average classification accuracies and their standard deviations are reported in Table 4.

Table 2: Cancerous and healthy tissue classification using the **cell-distribution** approach on the training set.

Grid size	Training set accuracy		
	Overall	Cancerous	Healthy
1	98.90 ∓ 0.38	98.13 ∓ 0.64	100.00 ∓ 0.00
2	100.00 ∓ 0.00	100.00 ∓ 0.00	100.00 ∓ 0.00
4	100.00 ∓ 0.00	100.00 ∓ 0.00	100.00 ∓ 0.00
8	100.00 ∓ 0.00	100.00 ∓ 0.00	100.00 ∓ 0.00
10	100.00 ∓ 0.00	100.00 ∓ 0.00	100.00 ∓ 0.00
16	100.00 ∓ 0.00	100.00 ∓ 0.00	100.00 ∓ 0.00
20	100.00 ∓ 0.00	100.00 ∓ 0.00	100.00 ∓ 0.00

Table 3: Cancerous and healthy tissue classification using the **cell-distribution** approach on the testing set.

Grid size	Testing set accuracy		
	Overall	Cancerous	Healthy
1	99.47 ∓ 0.24	99.40 ∓ 0.21	99.48 ∓ 0.58
2	98.48 ∓ 0.37	97.75 ∓ 0.64	99.35 ∓ 0.00
4	99.08 ∓ 0.00	98.80 ∓ 0.00	99.35 ∓ 0.00
8	99.75 ∓ 0.00	100.00 ∓ 0.00	99.35 ∓ 0.00
10	99.75 ∓ 0.00	100.00 ∓ 0.00	99.35 ∓ 0.00
16	99.75 ∓ 0.00	100.00 ∓ 0.00	99.35 ∓ 0.00
20	99.75 ∓ 0.00	100.00 ∓ 0.00	99.35 ∓ 0.00

Table 4: Cancerous and healthy tissue classification using the **texture-based** approach.

	Training set accuracy	Testing set accuracy
Overall	100.00 ∓ 0.00	98.46 ∓ 0.31
Cancerous	100.00 ∓ 0.00	99.76 ∓ 0.21
Healthy	100.00 ∓ 0.00	96.36 ∓ 0.63

Similar to Table 1 (cell-graph approach), Table 2 (cell-distribution approach) indicates that the training samples are almost classified correctly, except a few of the cancerous tissue samples when the grid size is 1 and Table 4 (texture-based approach) indicates that all the training samples are correctly classified. The tissue samples in the testing sets are also classified with accuracy greater than 95 % for all three approaches (Tables 1, 3, and 4).

3.2.2. Classification of cancerous, healthy, and inflamed tissues

Tables 1–3 demonstrate that the spatial distribution of the cells provides sufficient information to distinguish different class types when their cellular density is significantly different. To show that the cell-graph approach does not solely rely on the difference in the cellular density of different classes, we also use the images of the inflamed tissues that are as dense as the cancerous tissues. In Figure 2, the histogram of the amount of nodes is given for each class type. This figure exhibits that there is a significant difference between the number of nodes, i.e., cellular density, in the graphs of healthy and cancerous tissues. However, the numbers of nodes in the graphs of inflamed and cancerous tissue fall in the same range.

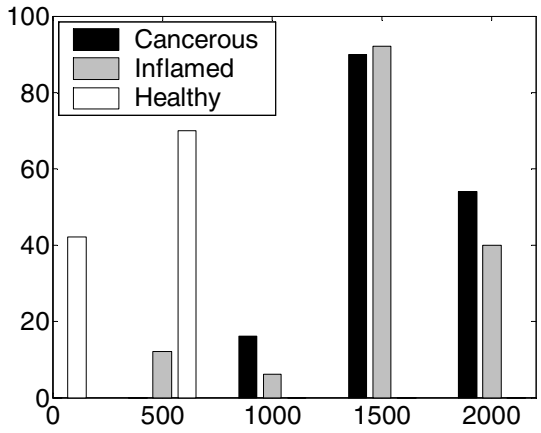


Figure 2: The histograms of the number of nodes in the graphs extracted from the different types of tissue images.

The average accuracy levels obtained in the classification of the cancerous, healthy, and inflamed tissues and their standard deviations using the cell-graph approach are presented in Table 5. This table demonstrates that the samples in both of the training and testing sets are

classified with accuracy > 94 %. In addition to the high accuracy in classification of the healthy tissues, the cancerous and inflamed tissues are distinguished from each other as well as from the healthy tissues with accuracy > 92 %.

Table 5: Cancerous, healthy, and inflamed tissue classification using the **cell-graph** approach.

	Training set accuracy	Testing set accuracy
Overall	95.93 ∓ 1.14	94.68 ∓ 0.71
Cancerous	93.95 ∓ 1.46	94.00 ∓ 0.79
Healthy	100.00 ∓ 0.00	96.30 ∓ 1.16
Inflamed	95.02 ∓ 2.03	92.19 ∓ 1.90

In Table 6 and Table 7, we report the average accuracy levels and their standard deviations obtained using the cell-distribution approach on the samples of the training and testing sets, respectively. We observe that the healthy tissue samples of the training set are almost classified correctly and the healthy tissue samples of the testing set are classified with accuracy greater than 93 %, regardless of the grid size. For all the grid sizes ranging between 1 and 20, the training samples of the cancerous tissues are also classified with a high accuracy level. On the other hand, approximately 30 % of the testing samples of cancerous tissues are misclassified for all of the exemplary values of the grid size. Note that, although the cell-distribution approach leads to higher accuracy on the training samples of cancerous tissues, the cell-graph approach yields significantly better testing accuracy, which is the real criterion to assess a classification in machine learning, than the cell-distribution approach. The classification accuracy of the training samples of inflamed tissues increases with the grid size. The maximum of these values is 98 %

when the grid size is 20; for this grid size, the overall accuracy is better than 99 %. On the other hand, for the grid size of 20, the classification accuracy of the testing samples of inflamed tissues is only 43 %. This accuracy for other grid sizes smaller than 20 is even lower indicating that the system cannot distinguish the inflamed tissues. We conclude that the pairwise relation encoded in the link establishing step of graph extraction provides critical information to distinguish different types of tissue samples regardless of their cellular density levels.

Table 6: Cancerous, healthy, and inflamed tissue classification using the **cell-distribution** approach on the training set.

Grid size	Training set accuracy			
	Overall	Cancerous	Healthy	Inflamed
1	91.04 ∓ 2.70	93.87 ∓ 95.50	100.00 ∓ 0.00	81.33 ∓ 9.15
2	93.77 ∓ 2.30	97.13 ∓ 2.63	100.00 ∓ 0.00	85.53 ∓ 6.15
4	96.09 ∓ 2.83	98.87 ∓ 2.50	99.82 ∓ 0.80	90.33 ∓ 8.51
8	98.44 ∓ 1.02	99.56 ∓ 0.93	100.00 ∓ 0.00	96.07 ∓ 2.98
10	97.99 ∓ 1.36	99.69 ∓ 0.56	99.73 ∓ 0.65	94.87 ∓ 4.20
16	98.46 ∓ 1.10	99.38 ∓ 1.18	100.00 ∓ 0.00	96.33 ∓ 3.40
20	99.12 ∓ 0.69	99.38 ∓ 0.86	100.00 ∓ 0.00	98.20 ∓ 1.90

Table 7: Cancerous, healthy, and inflamed tissue classification using the **cell-distribution** approach on the testing set.

Grid size	Testing set accuracy			
	Overall	Cancerous	Healthy	Inflamed
1	77.86 ∓ 5.29	71.79 ∓ 9.52	93.38 ∓ 1.64	26.41 ∓ 9.04
2	76.61 ∓ 4.41	69.66 ∓ 7.51	98.34 ∓ 2.22	26.09 ∓ 9.95
4	74.48 ∓ 6.37	65.68 ∓ 12.39	98.02 ∓ 1.78	29.69 ∓ 12.31
8	71.85 ∓ 3.23	62.85 ∓ 5.65	94.12 ∓ 1.89	34.69 ∓ 6.40
10	76.24 ∓ 2.81	70.74 ∓ 4.55	93.80 ∓ 3.67	34.53 ∓ 9.75
16	77.37 ∓ 3.24	70.66 ∓ 4.62	97.24 ∓ 1.81	33.91 ∓ 11.35
20	75.29 ∓ 2.39	66.47 ∓ 4.43	96.27 ∓ 2.30	42.97 ∓ 8.41

Table 8 demonstrates the classification accuracy and their standard deviations obtained using texture analysis. This table indicates that the classification accuracies of the training set obtained using the texture-based approach (> 99 %) are higher than those obtained using the cell-graph approach (> 93 %). On the other hand, for the testing set, the texture-based approach yields lower classification accuracies in the classification of the cancerous and inflamed tissues (89.60 % and 79.38 %) than the cell-graph approach (94.00 % and 92.19 %). For both of the approaches, the healthy tissues are classified with a similar level of accuracy of ~ 96 %. Although the results presented in Table 8 are not as low as in the case of the cell-distribution approach (Table 7), the cell-graph approach improves the accuracy of the texture-based approach in the classification of cancerous and inflamed tissues, indicating the effectiveness of the nodes in a cell-graph.

Table 8: Cancerous, healthy, and inflamed tissue classification using the **texture-based** approach.

	Training set accuracy	Testing set accuracy
Overall	99.62 ∓ 0.20	91.15 ∓ 1.03
Cancerous	99.00 ∓ 0.53	89.60 ∓ 1.97
Healthy	100.00 ∓ 0.00	96.10 ∓ 0.00
Inflamed	100.00 ∓ 0.00	79.38 ∓ 2.18

3.3. Comparison of the local and global graph metrics

The local graph metrics provide the information for each individual node of a graph and enable the cancer diagnosis at the cellular level. However, the global metrics provide the information for the entire graph, and thus, enable the detection of cancer at the tissue level. Since the global metrics are typically computed on the distributions of the local metrics, they are expected to be more reliable. This is also observed in the comparison of the accuracies obtained on the local graph metrics in [11] and on the global graph metrics in this study. For the classification of cancerous and inflamed tissue samples, we have obtained accuracy levels of 83–88 % and 92–95 % on the local and global metrics, respectively. For the classification of healthy tissue samples, we have obtained similar levels of accuracies.

In [11], we also used the classification results of the nodes (at the cellular level) to determine whether the tissue is correctly classified (at the tissue level) by examining the percentages of the nodes with correct classes. If this percentage is larger than an assumed *N* percent, we consider

that the tissue is classified correctly; otherwise we consider that it is misclassified. That is an indirect way of tissue classification necessitating to set the appropriate value for *N*. In this work, using the global metrics as the feature set in the classification introduces a direct way of tissue classification and eliminates the need of setting a value of *N*.

4. Conclusion

This work investigates the strength of the cell-graph representation in the diagnosis of cancer. In addition to encoding the spatial distribution of the cells, the cell-graphs encode the pairwise relations between the cells by assigning links between them. We demonstrate that this pairwise relation is crucial in classifying different types of tissues with similar cellular density levels.

In our experiments, we use 646 images of brain tissue samples surgically removed from 60 patients. We demonstrate that the cell-graph representation successfully distinguishes the images of cancerous tissues from the images of both healthy and inflamed tissues by using the global graph metrics. We obtain 94.68 % accuracy on the overall testing samples; the percentages of correct classification of the testing samples of healthy, cancerous, and inflamed tissues are 96.30 %, 94.00 %, and 92.19 %, respectively. On the other hand, the cell-distribution approach successfully classifies only the healthy tissues, but fails to distinguish the cancerous and inflamed tissues from each other. The maximum accuracy on the overall testing samples is 77.86 %; the percentages of correct classification of the testing samples of healthy, cancerous, and inflamed tissues are 98.34 %, 71.79 %, and 42.97 % at most, respectively. The texture-based approach successfully classifies the healthy tissues (96.10 %), as well. Although the testing set

accuracy in the classification of cancerous and inflamed tissues is not as low as in the case of the cell-distribution approach, it yields lower accuracy levels; 89.60 % and 79.38 % for cancerous and inflamed tissues, respectively. The summary of the classification accuracies of the cell-graph, cell-distribution, and texture-based approaches for the testing sets is given in Table 9. In this table, we report the value belonging to the grid size that leads to the maximum classification accuracy for the cell-distribution approach.

Table 9: Comparison of the classification accuracy obtained using the cell-graph, cell-distribution, and texture-based approaches on the testing sets.

	Cell-graph	Cell-distribution	Texture-based
Overall	94.68 ∓ 0.71	77.86 ∓ 5.29	91.15 ∓ 1.03
Cancerous	94.00 ∓ 0.79	71.79 ∓ 9.52	89.60 ∓ 1.97
Healthy	96.30 ∓ 1.16	98.34 ∓ 2.22	96.10 ∓ 0.00
Inflamed	92.19 ∓ 1.90	42.97 ∓ 8.41	79.38 ∓ 2.18

This work also compares the global graph metrics defined in this paper and the local graph metrics previously defined in [11]. We see that the global graph metrics yield better accuracy results than the local metrics to distinguish the cancerous, healthy, and inflamed classes.

References

[1] Bishop, C. M. *Neural Networks for Pattern Recognition*, Oxford: Oxford University Press, 1995.

[2] Cvetkovic, D. M., M. Boob, and H. Sachs. *Spectra of Graph*, Academic Press, 1978.

[3] Dorogovtsev, S. N. and J. F. F. Mendes. Evolution of networks. *Adv. Phys*. 51: 1979-1187, 2002.

[4] Einstein, A. J., H. S. Wu, M. Sanchez, and J. Gil. Fractal characterization of chromatin appearance for diagnosis in breast cytology. *J. Pathol*. 185: 366-381, 1998.

[5] Esgiar A. N., R. N. Naguib, M. K. Bennett, and A. Murray. Automated feature extraction and identification of colon carcinoma. *Anal. Quant. Cytol. Histol*. 20(4):297-301, 1998.

[6] Esgiar, A. N., R. N. G. Naguib, B. S. Sharif, M. K. Bennett, A. Murray. Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa. *IEEE T. Inf. Technol. B*. 2(3):197-203, 1998.

[7] Esgiar, A. N., R. N. G. Naguib, B. S. Sharif, M. K. Bennett, A. Murray. Fractal analysis in the detection of colonic cancer images. *IEEE T. Inf. Technol. B*. 6(1):54-58, 2002.

[8] Faloutsos, M., P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. *Proceedings of ACM/SIGCOMM,* pp. 251-262, Cambridge, MA, 1999.

[9] Ganster, H., P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler. Automated melanoma recognition. *IEEE T. Med. Imaging*. 20(3): 233-239, 2001.

[10] Glotsos, D., P. Spyridonos, P. Petalas, G. Nikiforidis, D. Cavouras, P. Ravazoula, P. Dadioti, and I. Lekka. Support vector machines for classification of histopathological images of brain tumour astrocytomas. *Proceedings of the International Conference on Computational Methods in Sciences and Engineering*. pp. 192-195, Kastoria, Greece, 2003.

[11] Gunduz, C., B. Yener, and S. H. Gultekin. The cell graphs of cancer. *Bioinformatics*. 20: i145-i151, 2004.

[12] Hamilton, P. W., D. C. Allen, P. C. Watt, C. C Patterson, and J. D. Biggart. Classification of normal colorectal mucosa and adenocarcinoma by morphometry. *Histopathology*. 11(9):901-911, 1987.

[13] Hamilton, P. W., P. H. Bartels, D. Thompson, N. H. Anderson, and R. Montironi. Automated location of dysplastic fields in colorectal histology using image texture analysis. *J. Pathol*. 182(1):68-75, 1997.

[14] Hartigan, J. A. and M. A. Wong. A k-means clustering algorithm. *Appl. Stat*. 28:100-108, 1979.

[15] Jain, A. K., J. Mao, and K. M. Mohiuddin. Artificial neural networks: a tutorial. *Computer*. 29:31-44, 1996.

[16] Jain, R. and A. Abraham. A comparative study of fuzzy classification methods on breast cancer data. *Australiasian Physical And Engineering Sciences in Medicine*. 2004 (to appear).

[17] Mangasarian, O.L, W.N. Street, and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Oper. Res*., 43(4):570-577, 1995.

[18] Pena-Reyes, C. A. and M. Sipper. A fuzzy genetic approach to breast cancer diagnosis. *Artif. Intell. Med*. 17(2):131-155, 1999.

[19] Schnorrenberg, F., C. S. Pattichis, C. N. Schizas, K. Kyriacou, and M. Vassiliou. Computer-aided classification of breast cancer nuclei. *Technol. Health Care*. 4(2):147-161, 1996.

[20] Tasoulis, D. K., P. Spyridonos, N. G. Pavlidis, D. Cavouras, P. Ravazoula, G. Nikiforidis, and M. N. Vrahatis. Urinary bladder tumor grade diagnosis using on-line trained neural networks. *Proc. Knowl. Based Intell. Inform. Eng. Syst*. pp.199-206, 2003.

[21] Todman, A. G., R. N. G. Naguib, and M. K. Bennett. Orientational coherence metrics: classification of colonic cancer images based on human form perception. *Canadian Conference on Electrical and Computer Engineering*. 2:1379-1384, 2001.

[22] Wolberg, W. H., W. N. Street, D. M. Heisey, and O. L. Mangasarian. Computer-derived nuclear features distinguish malignant from benign breast cytology. *Hum. Pathol*. 26(7):792-796, 1995.

[23] Zhou, Z. H., Y. Jiang, Y. B. Yang, and S. F. Chen. Lung cancer cell identification based on artificial neural network ensembles. *Artif. Intell. Med*. 24(1):25-36, 2002.