*Contributed Sessions*

# Augmented cell-graphs for automated cancer diagnosis

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

[1]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA and [2]Department of Pathology, Mount Sinai School of Medicine, New York, NY 10029, USA

**ABSTRACT**

**Summary:** This work reports a novel computational method based on augmented cell-graphs (ACG), which are constructed from low-magnification tissue images for the mathematical diagnosis of brain cancer (malignant glioma). An ACG is a simple, undirected, weighted and complete graph in which a node represents a cell cluster and an edge between a pair of nodes defines a binary relationship between them. Both the nodes and the edges of an ACG are assigned weights to capture more information about the topology of the tissue. In this work, the experiments are conducted on a dataset that is comprised of 646 human brain biopsy samples from 60 different patients. It is shown that the ACG approach yields sensitivity of 97.53% and specificities of 93.33 and 98.15% (for the inflamed and healthy, respectively) at the tissue level in glioma diagnosis.

**Contact:** demir@cs.rpi.edu

## 1 INTRODUCTION

Histology considers formation and structure of a tissue at the microscopic level and provides computational tools for the diagnosis of diseases including cancer.

There are a number of bio-computational methods proposed to diagnose cancer by automatically capturing the histological changes in a tissue. The automated diagnosis techniques not only decrease the inter-observer variability, but also facilitate large-scale cancer screening by assisting pathologists.

In automated cancer diagnosis, one of the most important steps is the feature (i.e. metric) extraction which aims at identifying and quantifying the properties of a single cell or an entire tissue. In literature, there are four different approaches to feature extraction: morphological (Blekas *et al.*, 1998; Spyridonos *et al.*, 2001; Street *et al.*, 1993; Thiran and Macq, 1996; Wolberg *et al.*, 1995), textural (Albregtsen *et al.*, 2000; Esgiar *et al.*, 1998, 2002; Hamilton *et al.*, 1997; Smolle, 2000), intensity-based (Schnorrenberg *et al.*, 1996; Weyn *et al.*, 1998; Wiltgen *et al.*, 2003; Zhou *et al.*, 2002), and topological (Choi *et al.*, 1997; Keenan *et al.*, 2000; Weyn *et al.*, 1999). Although these approaches yield promising results, they all need to address (1) the difficulty of determining the exact locations of cells/nuclei in the biopsy image, and/or (2) the sensitivity to the noise that arises from the stain artifacts in the image.

In this work, we report a novel computational method based on the augmented cell-graphs (ACG) of cancer which enables us to define a new set of features. An ACG is a simple (i.e. no self loops),

undirected, weighted, and complete (i.e. with an edge between every pair of nodes) graph obtained from the histopathological image of a tissue. In an ACG, a node represents a cell cluster and an edge between a pair of nodes defines a binary relationship between them. Both the nodes and the edges of an ACG are assigned weights to capture more information about the topology of the tissue. In this work, we define the node weight as the size of the cluster corresponding to a node, and the edge weight as the Euclidean distance between the end points of an edge.

The ACG approach is based on the distinctive topological properties of self-organizing malignant cells rather than the exact location of each cell. In this approach, the nodes correspond to cell clusters rather than individual cells. Thus, the ACG approach eliminates the need for the exact loci of the cells, and the coarse loci of the cells would suffice. Furthermore, the ACG approach neither uses the intensity values of the pixels directly in the feature extraction nor it is sensitive to the gray-scale dependencies between the pixels. Thus, it is likely to be immune to the noise inherit in a histopathological image.

There are other topological approaches for the analysis of histopathological images. In Choi *et al.* (1997), Keenan *et al.* (2000) and Weyn *et al.* (1999), Voronoi diagrams and their Delaunay triangulations are suggested to capture the topological information between the adjacent cells. Thus, these approaches focus on local-edges only. In the ACG approach, the definition of an edge enables capture of the spatial information beyond adjacent cells. In Gunduz *et al.* (2004), cell-graphs are constructed also by defining edges between a pair of non-adjacent cells with a certain probability. While this approach remedies the local-edges-only issue, it may miss some edges that can be important for automated diagnosis owing to the probabilistic assignment of edges in the cell-graph. Furthermore, in Gunduz *et al.* (2004), no distinction is made (e.g. no weights are assigned) between nodes and edges once they are established. We distinguish the ACGs presented in this work from the simple cell-graphs used in Gunduz *et al.* (2004) and Demir *et al.* (2005) since an ACG defines edges between all pairs of nodes, and assigns weights to each node and edge to reflect their contribution to the topology of the tissue.

On an ACG, we compute global metrics defined over the entire graph (quantifying the properties of the entire tissue sample) and show that the ACG approach improves the accuracy of automated cancer diagnosis. The dataset used in this work consists of 646 brain biopsy samples of 60 different patients. It contains three different types of tissue samples as shown in Figure 1: (1) a brain tumor (i.e. malignant glioma), (2) a healthy tissue, and (3) a tissue sample of a benign inflammatory process. Remark that the cancerous samples are easily distinguished from the healthy ones even with untrained eyes (as shown in Fig. 1a and b, respectively). However, it

---

*To whom correspondence should be addressed.

†Present address: Department of Pathology, Oregon Health and Science University, Portland, OR 97205, USA.
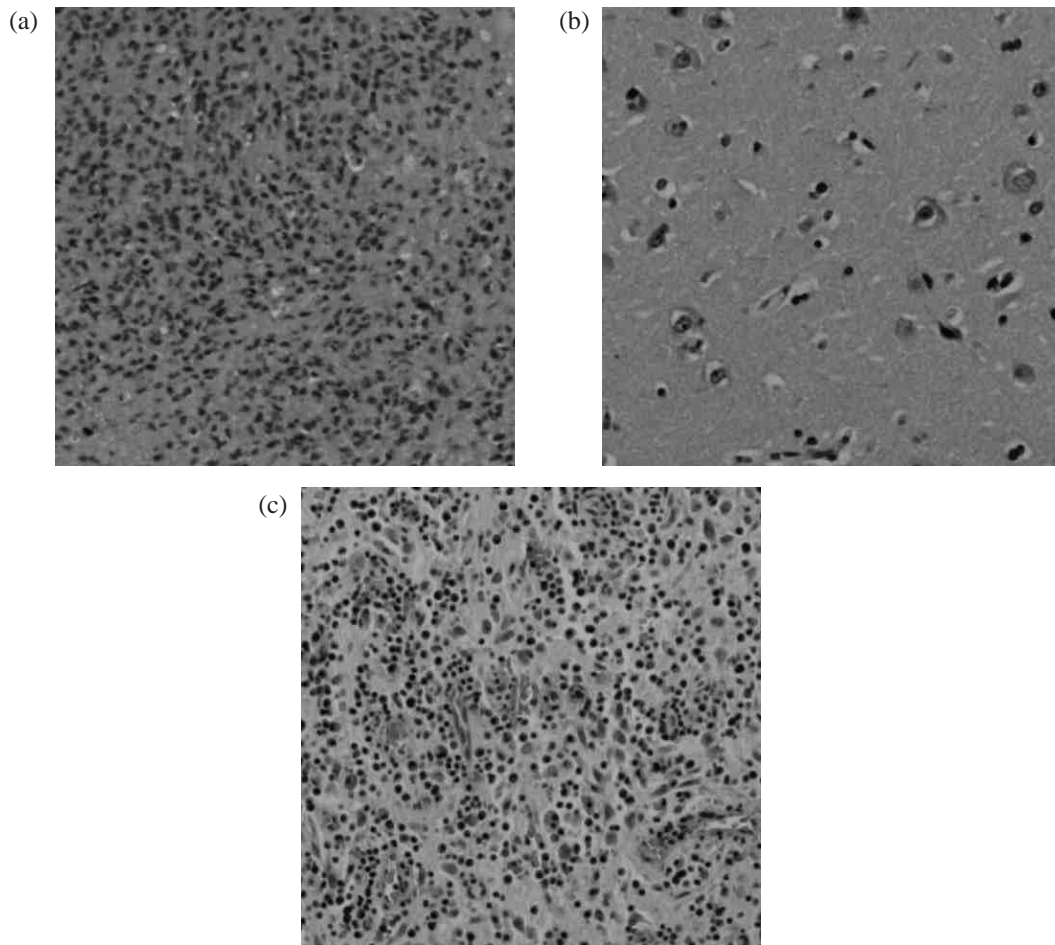
(a)

(b)

(c)

**Fig. 1.** Exemplary microscopic images of tissue samples surgically removed from human brain: (**a**) a brain tumor sample (i.e. malignant glioma), (**b**) a healthy tissue sample, and (**c**) a tissue sample of a benign inflammatory process.

is not a straightforward differentiation between the cancerous tissues and benign inflammatory processes[1] (as shown in Fig. 1a and c, respectively).

Despite such visual similarity of the cancerous and inflamed biopsy samples, we demonstrate that the ACG approach yields sensitivity of 97.53% and specificities of 93.33 and 98.15% (for the inflamed and the healthy, respectively) in the cancer diagnosis at the tissue level. In comparison to results obtained on the same dataset using similar global metrics computed on simple cell-graphs (Demir *et al*., 2005), the ACG approach leads to statistically significant improvement in the accuracy and sensitivity.

## 2 METHODOLOGY

The ACG approach contains three steps: segmentation, feature extraction, and diagnosis, as illustrated in Figure 2.

### 2.1 Segmentation

In the segmentation step, we construct the nodes of the cell-graph by identifying the cell clusters in the biopsy image. Since we utilize the cell clusters

---

[1]Herein referred to as 'inflamed tissues'.

instead of individual cells, only the coarse locations of the cell clusters are necessary.

In the first part of the segmentation, we classify the pixels of the image as 'cell' or 'background'. For this, we first learn how to classify them using the $k$-means algorithm (Haralick, 1979). The $k$-means algorithm clusters the pixels of the training samples into $k$ clustering vectors. Subsequently, a human expert assigns each of these clustering vectors either 'cell' or 'background' class. Once learned, the clustering vectors together with their class assignments are used in the classification of the pixels in the other images. In this part, we have the control parameter of $k$. The value of this parameter should be selected large enough to represent the different parts of a biopsy sample. On the other hand, unlike other parameters used in the cell-graph approach, the selection of this parameter is limited to the human perception; the human expert should be able to reproducibly distinguish the different clusters and successfully assign the corresponding classes to these clusters. Considering this trade-off, we set the value of this parameter to be 16 in our work.

In the second part of the segmentation, we identify the nodes of the cell-graph making use of the pixel classification. After classifying the pixels, we convert the values of the pixels of 'cell' and 'background' classes to 1 and 0, respectively. Then, we place a grid on the resulting image. For each grid entry, we compute the ratio of the 'cell' pixels in the grid entry. The grid entries with ratios greater than a node threshold are identified to be the nodes of the graph and the ratio of each identified node will be used as the weight of its corresponding node.

STEP 1: Segmentation

   Clustering
   Expert labeling
   Node identification

STEP 2: Feature extraction

   Edge establishing
   Global metric extraction

STEP 3: Diagnosis

   Training
   (using neural networks)
   Parameter selection
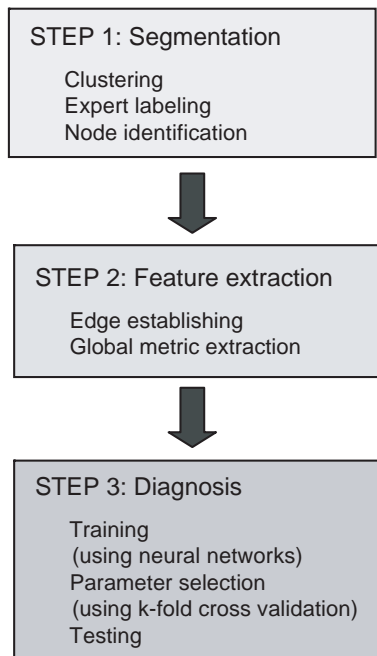   (using k-fold cross validation)
   Testing

**Fig. 2.** The summary of three-step methodology: segmentation, feature extraction and diagnosis.

In the second part, we have two free parameters: grid size and node threshold. The grid size determines the size of the cell clusters and, depending on its value, a cell cluster can typically represent a single cell, a part of a cell or a bunch of cells. Together with the grid size, the node threshold determines the cellular density of the graph. Larger values of the threshold result in sparser graphs. Smaller values of this parameter make the segmentation more sensitive to the noise that arises from the stain artifacts and the misassignment of pixels in the first part of the segmentation. In our experiments, we optimize the values of these parameters using $k$-fold cross-validation. We will discuss the details of the parameter selection in Section 3.2.

## 2.2 Feature extraction

In the feature extraction step, we first generate an ACG by defining an edge between each pair of nodes and assigning the Euclidean distance between its end points as the weight of that edge.[2] Since an ACG is a complete graph, there is an edge between every pair of nodes, thus we can avoid several control parameters used in the previous work (Gunduz *et al.*, 2004; Demir *et al.*, 2005) for establishing edges.

Next, we define the graph metrics and compute them on the ACGs. In this work, we use 7 such global metrics:

(1) The average degree: For weighted graphs, the degree of a node is defined as the sum of the weights of the edges that belong to this node. In this work, we normalize the calculated degree dividing it by the sum of degrees of all nodes.

(2) The average eccentricity: The eccentricity of a node is the length of the maximum of the shortest paths between the node and every other node reachable from it. In this work, we compute the shortest path lengths considering the edge weights.

[2]The importance of the edge increases with decreasing Euclidean distance. We construct our weighted graphs according to this principle, since some graph algorithms (e.g. shortest path algorithms) require associating smaller weights to shorter edges which correspond to closer nodes.

(3) The average node weight: We compute the node weights as the ratio of the 'cell' pixels in the cell clusters in the segmentation step.

(4) The most frequent edge weight: We group the edges according to the integral part of their weights and use the most frequent integral part observed in the ACG as a global metric.

(5) The other three global metrics are related to the spectral decomposition of the graph, i.e. the set of the eigenvalues of the graph. In graph theory, the graph spectrum is closely related to the topological properties of the graph (Cvetkovic *et al.*, 1978). In this work, we use three metrics that characterize the ∼30 eigenvalues with the largest absolute values in the spectrum. The first metric is the spectral radius; it is defined as the largest absolute value of the eigenvalues in the spectrum. Similarly, we use the second largest absolute value of the eigenvalues in the spectrum as the second metric. The last metric is the eigen exponent which is defined as the slope of the sorted eigenvalues as a function of their orders in log-log scale (Faloutsos *et al.*, 1999); in this work, we use the slope between the third largest and its next largest 30 eigenvalues.

## 2.3 Diagnosis

The last step is the classification of the tissues according to their histological properties. For this, we employ the global metrics explained in Section 2.2 as the feature set and an artificial neural network as the classifier. Neural networks are non-linear models that capture complex interactions among the input data and they tolerate the noisy and irrelevant information (Bishop, 1995). In this work, we use a back-propagation multilayer perceptron (MLP) with a single hidden layer. In the MLP we use, the input layer consists of the global metrics derived from ACGs and the output layer consists of three nodes each of which corresponds to a tissue class (i.e. cancerous, inflamed or healthy). The number of hidden units in the hidden layer is a free parameter and we optimize this parameter again by using $k$-fold cross-validation.

# 3 EXPERIMENTS

## 3.1 Dataset preparation

We conduct our experiments on a dataset consisting of 646 microscopic images of brain biopsy samples of 60 different patients each of which has a malignant glioma (a type of brain cancer), a benign inflammatory process or a healthy tissue. These 60 patients are randomly chosen from Pathology Department archives in the Mount Sinai School of Medicine.[3] The number of patients with the cancerous, inflamed, and healthy tissue samples is 41, 9 and 14, respectively; for 4 patients, we have both the cancerous and healthy tissue samples. These tissue samples consist of 5–6 μm thick tissue section stained with the hematoxylin and eosin technique.

We take the images of these tissue samples by using a Nikon Coolscope Digital Camera. The images are taken in the RGB color space with a magnification of 100×. Prior to segmentation, the RGB values of the pixels are converted to their corresponding values in the La*b* color space since this space is a uniform color space that provides separate color and detail information (Wyszecki and Stiles, 2000). The images used in the dataset consist of 480 × 480 pixels.

We divide this dataset into training and test sets. To reflect the real-life situation in the patient distribution of the test set, we use half of the patients of each type in the test set and the rest in the training set. For the test set, the number of the biopsy images of each

[3]All patients were adults with both sexes included. Prior to obtaining digital images of the tissues, the identifiers were removed and slides were numerically recoded corresponding to diagnostic categories by the pathologist; therefore, the personal identifiers cannot be retraced by the other members of our team.

patient is ∼8 (varying between 6 and 10). For the training set, we still use ∼8 biopsy images for each cancerous patient. We use larger amounts of biopsy samples for the healthy and the inflamed since it might be harder for a neural network to learn the rarer classes, if the number of training samples of each class varies significantly between the different classes. Additionally, since the number of available inflamed tissues is less than those of healthy and cancerous samples, we replicate the inflamed samples in the training set. In summary, we use 163 cancerous tissues of 20 patients, 150 inflamed tissues of 5 patients (the dataset included 75 inflamed tissues prior to the replication), and 156 healthy tissues of 7 patients in the training set. In the test set, we use 166 cancerous tissues of 21 patients, 32 inflamed tissues of 4 patients, and 54 healthy tissues of 7 patients.

This dataset consists of some dependent biopsy samples; the samples of the same patient are not independent. It would result in over-optimistic accuracies results for the test set, if different biopsy samples of the same patient were both used in training and testing. To avoid such over-optimistic results, we use the biopsy samples of entirely different patients in training and test sets. Furthermore, we optimize the free parameters on the cross-validation sets (within the training set) without considering the accuracy of the test set.

## 3.2 Parameter selection

We have selected the free parameters (the grid size, node threshold, and number of hidden units) by using 30-fold cross-validation. In $k$-fold cross-validation, the training set is randomly partitioned into $k$ non-overlapping subsets; the $k - 1$ of them are used to train the classifier, and the remaining one is used to estimate the performance of the classifier. This is repeated $k$ times as all distinct subsets are used in estimating the performance. The classifier performance is estimated as the average of the performances obtained in separate $k$ trials.

In Figures 3 and 4, we illustrate the dependency of the classification accuracy (obtained using 30-fold cross-validation) on the value of the grid size for two different exemplary node thresholds (0.25 and 0.50, respectively); in these figures, we illustrate the accuracies for different numbers of hidden units, namely 4, 8, 12, and 16. These figures demonstrate that the better classification results are obtained for the smaller grid sizes. For the grid sizes below a threshold (e.g. grid values of 4 and 6 in our experiments), the results are very close to each other. Especially for larger node thresholds (e.g. for the value of 0.50), the accuracy decrease with the increasing grid size is obvious (in Fig. 4).

In Table 1, we report the classification results of smaller grid sizes that are obtained when 16 hidden units are used; we report the average accuracy obtained on the cross-validation sets and its standard deviation. For these results, we also perform the $t$-test on the difference between the classification accuracy obtained for different parameter sets (we use the $t$-test with a significance level of 0.05). The $t$-test demonstrates that there is no significant difference between the accuracy of the following parameter sets: {4, 0.25}, {4, 0.50}, {6, 0.25}, {6, 0.50}, and {8, 0.50}, where the first element in each set is the grid size and the second is the node threshold.

We also investigate the effects of the node threshold selection. For this, we fix the grid size as 4 [one of the grid sizes that yields best accuracy results on cross-validation sets (Table 1)]. For different values of the node threshold (ranging between 0.25 and 0.99), the classification accuracies obtained using 30-fold cross-validation are shown in Figure 5. This figure exhibits that, for the smaller values of
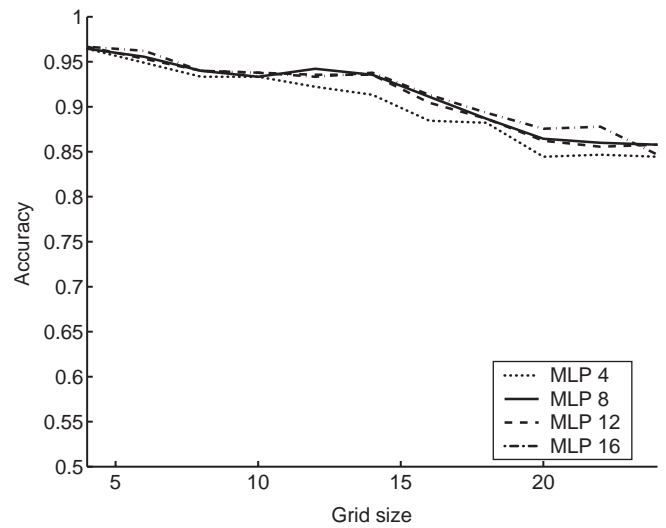


**Fig. 3.** The dependency of the classification accuracy of the cross-validation sets on the selection of the grid size for different numbers of hidden units in a MLP. (Here the node threshold selected is 0.25.)
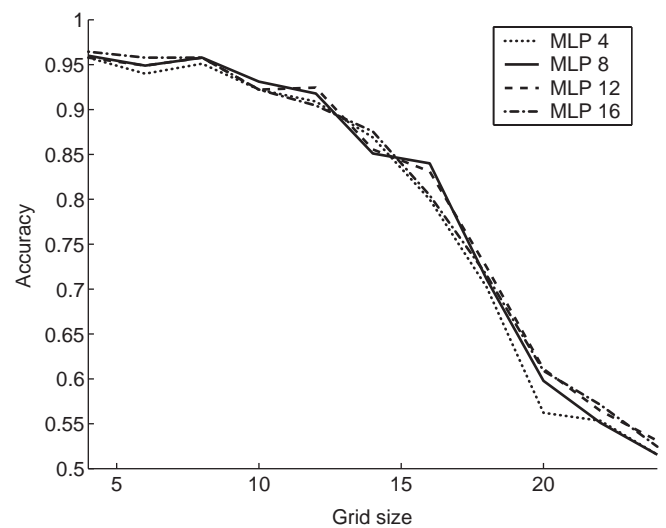


**Fig. 4.** The dependency of the classification accuracy of the cross-validation sets on the selection of the grid size for different numbers of hidden units in a MLP. (Here the node threshold selected is 0.50.)

**Table 1.** Average classification accuracies obtained on the cross-validation sets and their standard deviations, for smaller grid sizes

| Grid size | Accuracy on CV | |
| | Node threshold = 0.25 | Node threshold = 0.50 |
| --- | --- | --- |
| 4 | 96.67 ± 4.55 | 96.44 ± 5.17 |
| 6 | 96.22 ± 6.93 | 95.78 ± 6.43 |
| 8 | 94.00 ± 7.50 | 95.78 ± 6.19 |
| 10 | 93.78 ± 6.54 | 92.22 ± 6.80 |

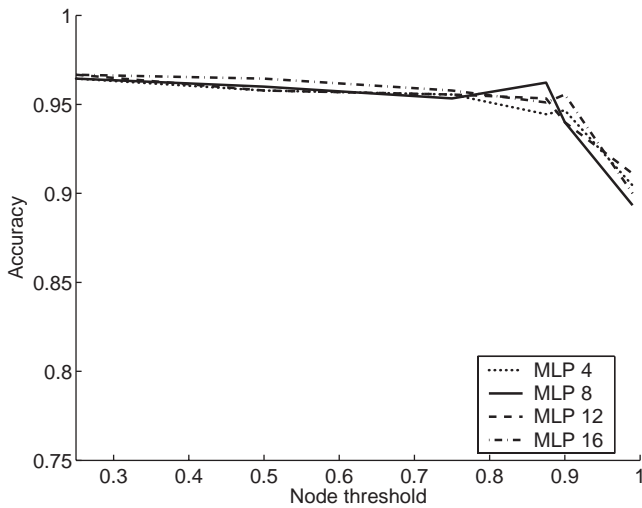In the classification, a MLP with 16 hidden units is used.

**Fig. 5.** The dependency of the classification accuracy of the cross-validation sets on the selection of the node threshold for different numbers of hidden units in a MLP. (Here the grid size selected is 4).

**Table 2.** The average accuracy, sensitivity and specificity (obtained over 30 runs) for the ACG in the first two rows and simple cell-graph (Demir *et al.*, 2005) in the third row

| Parameters | Accuracy | Sensitivity | Specificity Inflamed | Healthy |
|---|---|---|---|---|
| {4, 25} | 96.93 (±0.52) | 97.51 (±0.52) | 91.88 (±1.76) | 98.15 (±0.00) |
| {4, 50} | 97.13 (±0.32) | 97.53 (±0.52) | 93.33 (±1.08) | 98.15 (±0.00) |
| {4, 50, −4.4} | 95.45 (±1.33) | 95.14 (±2.03) | 92.50 (±1.76) | 98.15 (±0.00) |

Parameters in the first column are given in the forms of {grid size, node threshold} in the first two rows and {grid size, node threshold, edge exponent} in the third row.

the node threshold, the classification accuracy is similar regardless of the node threshold value. When we increase its value above ∼0.9, the classification accuracy suddenly decreases.

### 3.3 Results

By making use of the 30-fold cross-validation results, we select two sets of parameters: {4, 0.25} and {4, 0.50}, where the first element in each set is the grid size and the second is the node threshold; for both the parameter sets, we set the number of hidden units to be 16. For each parameter set, we train our system by running a MLP 30 times. The accuracy as well as the sensitivity and specificity obtained on the test set are given in the first two rows in Table 2. In Table 2 (in the third row), we also present the accuracy, sensitivity, and specificity obtained using the global metrics extracted for the simple cell-graphs in Demir *et al.* (2005) on the same dataset. In that work, the cell-graph parameters {the grid size, node threshold, edge exponent} are also selected using *k*-fold cross-validation, and the best classification results (on the cross-validation sets) are obtained when these parameters are 4, 0.50, and −4.4, respectively.

The *t*-test conducted on these classification results demonstrates that the accuracy and the sensitivity of the cancer diagnosis are significantly improved using ACGs. For the specificity of the inflamed type, we obtain statistically better results using ACGs with a parameter set of {4, 0.50}. On the other hand, there is no significant difference between the approaches of simple cell-graphs and ACGs with a parameter set of {4, 0.25}. The specificity of the healthy type is the same for both the cell-graph approaches.

## 4 CONCLUSION

In this work, we report a new method based on the construction of ACGs with weighted nodes and weighted edges for the computational diagnosis of malignant glioma. In this article, we describe the methodological steps of the ACG approach (segmentation, graph metric extraction, and diagnosis) together with the experimental results. Utilizing ACGs for the first time, we present that unprecedented accuracy is achieved in the classification of glioma based on the distinctive topological properties of its self-organizing malignant cells and that ACGs outperform their counterparts, the simple cell-graphs. ACGs, which are fully connected and whose edges are deterministically established between every possible pair of cell clusters in a tissue image, encode the complete topological information available in the image (such as the cell cluster size and the Euclidean distance between the clusters); given the low resolution of the image, no spatial information is wasted.

In this work, we use images of 646 human brain biopsy samples taken from a total of 60 patients. These images include 329 malignant glioma samples from 41 patients, 107 benign inflamed samples from 9 patients, and 210 healthy samples from 14 patients. From tissue images of $480 \times 480$ pixels with $100 \times$ magnification, we construct ACGs and compute global graph metrics such as the degree, eccentricity, spectral radius and eigen exponent that quantify different aspects of cell distribution across the tissue sample in its entirety. This is the first work that relies on global graph metrics defined on fully-connected cancer cell-graphs. We select the control parameters using *k*-fold cross-validation; this optimizes the free parameter selection. Our experimental results show that the ACG approach leads to sensitivity of 97.53% and specificities of 93.33 and 98.15% (for the inflamed and healthy, respectively) at the tissue level in glioma diagnosis. To the best of our knowledge, these are the best results ever reported with high reliability on cell-graph based glioma diagnosis.

### REFERENCES

Albregtsen,F., Nielsen,B. and Danielsen,H.E. (2000) Adaptive gray level run length features from class distance matrices. In *15th International Conference on Pattern Recognition*, Barcelona, **3**, 746–749.

Bishop,C.M. (1995) *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.

Blekas,K. *et al.* (1998) Cytological diagnosis based on fuzzy neural networks, *J. Intell. Syst.*, **8**, 55–79.

Choi,H.-K. *et al.* (1997) Image analysis based grading of bladder carcinoma. Comparison of object, texture and graph based methods and their reproducibility, *Anal. Cell Pathol.*, **15**, 1–18.

Cvetkovic,D.M., Boob,M. and Sachs,H. (1978) *Spectra of Graph*. Academic Press.

Demir,C., Gultekin,S.H. and Yener,B. (2005) Learning the topological properties of brain tumors. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, inpress.

Esgiar,A.N. *et al*. (1998) Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa. *IEEE Trans. Inf. Technol. B*, **2**, 197–203.

Esgiar,A.N. *et al*. (2002) Fractal analysis in the detection of colonic cancer images. *IEEE Trans. Inf. Technol. B*, **6**, 54–58.

Faloutsos,M., Faloutsos,P. and Faloutsos,C. (1999) On power-law relationships of the Internet topology. In *Proceedings of ACM/SIGCOMM*, Cambridge, MA, pp. 251–262.

Gunduz,C. *et al*. (2004) The cell graphs of cancer. *Bioinformatics*, **20**, i145–i151.

Hamilton,P.W. *et al*. (1997) Automated location of dysplastic fields in colorectal histology using image texture analysis. *J. Pathol.*, **182**, 68–75.

Haralick,R.M. (1979) Statistical and structural approaches to texture. *Proc IEEE*, **67**, 786–804.

Keenan,S.J. *et al*. (2000) An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN), *J. Pathol.*, **192**, 351–362.

Schnorrenberg,F. *et al*. (1996) Computer-aided classification of breast cancer nuclei. *Technol. Health Care*, **4**, 147–161.

Smolle,J. (2000) Computer recognition of skin structures using discriminant and cluster analysis. *Skin Res. Technol.*, **6**, 58–63.

Spyridonos,P. *et al*. (2001) Computer-based grading of haematoxylin-eosin stained tissue sections of urinary bladder carcinomas. *Med. Inform. Internet Med.*, **26**, 179–190.

Street,W.N., Wolberg,W.H. and Mangasarian,O.L. (1993) Nuclear feature extraction for breast tumor diagnosis. In *Proceedings of the IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*, San Jose, CA, **1905**, 861–870.

Thiran,J.-P. and Macq,B. (1996) Morphological feature extraction for the classification of digital images of cancerous tissues, *IEEE Trans. Bio-Med. Eng.*, **43**, 1011–1020.

Weyn,B. *et al*. (1998) Automated breast tumor diagnosis and grading based on wavelet chromatin texture description, *Cytometry*, **33**, 32–40.

Weyn,B. *et al*. (1999) Computer-assisted differential diagnosis of malignant mesothelioma based on syntactic structure analysis. *Cytometry*, **35**, 23–29.

Wiltgen,M. *et al*. (2003) Tissue counter analysis of benign common nevi and malignant melanoma. *Int. J. Med. Inform.*, **69**, 17–28.

Wolberg,W.H. *et al*. (1995) Computer-derived nuclear features distinguish malignant from benign breast cytology. *Hum. Pathol.*, **26**, 792–796.

Wyszecki,G. and Stiles,W.S. (2000) *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2nd edition. Wiley and Sons.

Zhou,Z.H. *et al*. (2002) Lung cancer cell identification based on artificial neural network ensembles. *Artif. Intell. Med.*, **24**, 25–36.