Learning the Cell-Graphs: Macroscopic Modeling of Brain Tumors

Çiğdem Gündüz[*] Bülent Yener[†] S. Humayun Gultekin[‡]

December 8, 2003

Abstract

Diffuse gliomas are brain tumors that invade the surrounding normal tissue by an aggressive diffusion process. This diffuse invasive behavior affects the prognosis adversely, and renders radical treatment impossible. Current mathematical models to quantify and analyze a cancer tumor are not scalable due to their enormous complexity. We developed a scalable, graph theoretical model, based on the spatial relationship between the cells, to quantify the properties of the invasion. The graph theoretical model is used by a machine learning algorithm. The learning algorithm uses graph metrics to distinguish (1) gliomas from surrounding normal tissue, and (ii) gliomas from inflammation. We tested the algorithms on real data to validate the proposed approach.

1 Introduction

Cancer is an uncontrolled proliferation of cells that express varying degrees of fidelity to their precursors. Neoplastic process entails not only cellular proliferation but also a modification of the differentiation of the involved cell types. Thus, in a sense cancer may be viewed as a burlesque of normal development [1].

Diffuse (malignant) gliomas are brain tumors that possess the capability to infiltrate the surrounding healthy brain tissues by an initially non-destructive migrational manner. The biological basis for glioma invasion constitutes a complex process involving cell-to-cell interaction, adhesion to the exctracellular matrix, tumor cell motility, and enzymatic remodeling of the extracellular space [12]. Although the state of art medical imaging improved the detection of gliomas; quantification of the extent of invasion, prediction of biological behavior, and radical surgical removal in individual cases remains a challenge.

Mathematical modeling of cancer and quantifications of its properties has been a focus of intensive research [2, 3, 6, 4, 7]. However, current computational and mathematical models at the cellular level are not scalable. Some of these approaches are based on Monte-Carlo algorithm [6, 7]. Others are based on formulating continuous differential equations and finding probability generating functions to model the cell behavior. Clearly, solving large number of equations or simulating millions of cells with Monte-Carlo algorithms have prohibitive computational complexity. Thus, addressing the scalability problem requires new algorithmic approaches and new models.

This work offers novel mathematical technique to model a cancer tumor. It is based on examination of the coordinates of individual cells in a sample tissue to construct a cell-graph. The mathematical properties of the cell-graph are computed to identify subgraphs that represent different biomedical phenomena in the sample tissue. The identification is done by a machine learning algorithm that is trained over numerous samples under human (expert) supervision. The learning algorithm can successfully distinguish (1) gliomas from surrounding normal tissue, and (ii) gliomas from other invasions such as inflammation.

[*] Department of Computer Science, Rensselaer Polytechnic Institute, NY 12180, gunduz@cs.rpi.edu

[†] Department of Computer Science, Rensselaer Polytechnic Institute, NY 12180, yener@cs.rpi.edu

[‡] Department of Pathology, Mount Sinai School of Medicine, NY 10029, Humayun.Gultekin@msnyuhealth.org

The graph theoretical approach is motivated by the fact that many real-world, self-organizing, complex dynamic systems can be represented by graphs. Furthermore, precise metrics are available to quantify the properties of these graphs and identify their characteristics. One example is the Hollywood movie star network, obtained by drawing a line between two actors if they played in the same movie. This network is derived from 150,000 movies and has 300,000 nodes. Another one is the WWW graph in which each page is a node and each *url* is a directed link. This graph has billions of nodes and several billions of links (it was based on 1999 data). Similarly, the Internet router graph has hundreds of thousands nodes and links. Another example is the USA power grid network which has approximately 5,000 nodes. Collaboration network among the mathematicians with 70,000 nodes and 200,000 links (1991-1998 data) is another example. Finally, the tiny neural network of C-elegance worm with 300 nodes (neurons) shares common properties with the earlier mentioned, much large networks. Although the size and domains of these graphs are very different, it is possible to distinguish them from random graphs [5] using some of the metrics that are adapted in this work as well.

In this work we report our initial results that we can construct a cell-graph from sample tissue images, and deploy a learning algorithm that distinguishes between different regions in the tissue based on the graph metrics. The proposed approach is scalable since graphs with order of millions nodes can be tackled to compute the metrics of interest.

The remaining of this paper is organized as follows. In Section 2, we define our methodology to construct a graph from given tissue information. The topological properties defined on this graph are explained in Section 3. We give our experimental results and their interpretations in Section 4. Section 5 concludes and explains the possible future research directions.

2 Methodology

The methodology used in this work is summarized in Figure 1. The first step is to obtain tissue images from the surgically removed clinical data. Staining process of the data enables us to see and take their images under a microscope. Using these images of tissue samples, we develop a tool to distinguish and recognize different type of cells, e.g., healthy, cancer or inflamed cells. These steps are illustrated as the next components of this figure.

In this work, our approach is based on construction of cell-graphs from the tissue images. A cell-graph is denoted by $G = (V, E)$ where the vertex (node) set represents the nucleus of cells and the edge set E defines a locality relationship between them. The cell-graph is obtained after the following steps (these steps are visualized in Figures 4 and 5 on a sample image):

1. First we determine the cell locations in a tissue image. This problem requires to distinguish the cells from their background. We use *k-means* clustering algorithm which is based on the color information of the pixels [8]. After setting the cluster vectors on our training samples, our pathology expert analyzes the cluster information and assign classes to the cluster vectors, i.e., he labels these clusters as either cell or non-cell regions. We label pixels as 1 if they are cells and 0 otherwise. This information is to be used in all of our tissue samples (during testing).

2. Next we transform the cell information to identify the nodes (vertices) of the graph. The main difficulty here is the *noise*: in glioma samples, there are too many cells with different sizes as well as coinciding cells. The noise prevents a one-to-one mapping between a cell and a node. Moreover, if a one to one mapping were possible, then the number of nodes in our graph would be dependent on the number of cells, which makes the computation hard for very large tissue cells.

   Our approach to this problem is to embed a grid over the sample image, and calculate the probability of a grid entry being a cell. For each entry, we compute the probability value as the average of the label of pixels located in this entry. We apply a threshold (node-threshold) to the computed

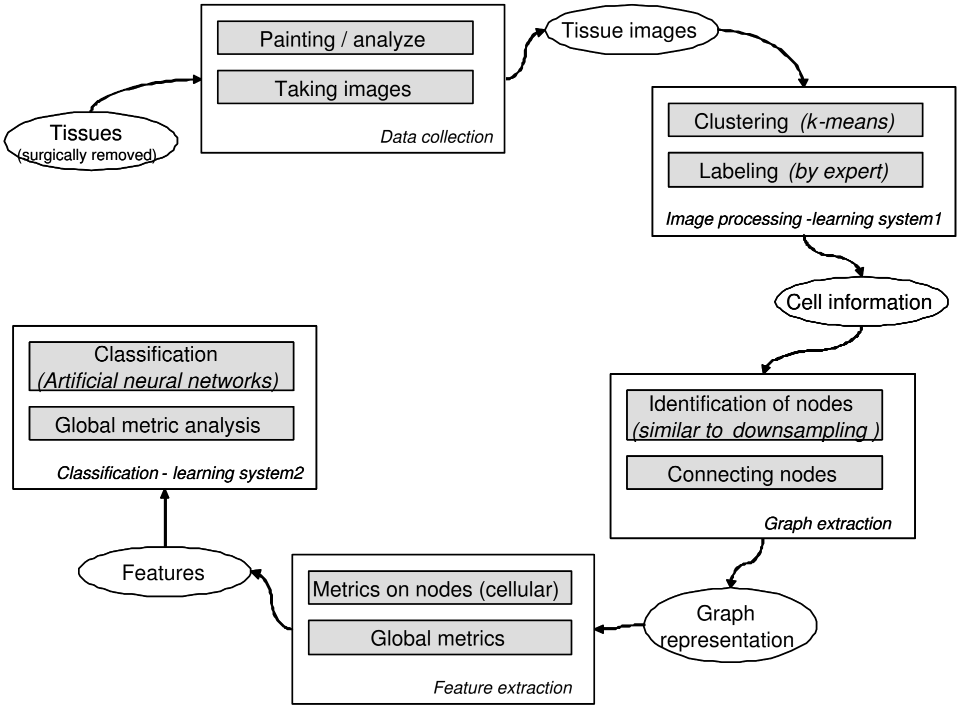

Figure 1: A general view of our system. In the first component the tissue image is extracted from the clinical data surgically removed from the patients. The second component (image processing) distinguishes cells from their background. Third component extracts the *cell-graphs*, and the metrics are computed on them in the fourth component. The last component is the classification tool that classifies the different type of cells, e.g., healthy inflamed.

probability values and the values greater than the node-threshold are labeled as cells, whereas the others are labeled as background. This step can also be considered as downsampling of the image obtained in Step 1. Note that the resolution of a tissue image determines the complexity of whole process.

We have two control parameters in this step: (i) the size of the grid, and (ii) the node-threshold value. Increasing the node-threshold value produces sparser graphs and the grid size determines the downsampling rate. At the end of this step the spatial information of the nodes is translated to their locations on a 2D grid.

3. The last step is to define the edge set to connect the nodes found in Step 2 to construct the graph. In this step, we use spatial information, which are the locations of the nodes in a 2D grid. We define an edge-threshold such that any two nodes are to be connected if the distance between them is smaller than the edge-threshold. This threshold affects the connectivity of the graph. Increasing the edge-threshold results in denser graphs.

We compute six different metrics on the resultant graphs, reflecting their different topological properties. The metrics are used by a machine learning algorithm to classify different cell concentrations as cancerous, normal or inflammation. The learning algorithms used in this paper are briefly explained next.

2.1 K-means Algorithm

The k-means algorithm is an unsupervised learning algorithm that clusters the data based on their features [8]. As its name implies we have k cluster vectors and each sample belongs to one of the clusters whose center is the closest to that sample. After assigning the sample to one of the clusters, the sample is represented by this cluster vector.

K-means algorithm is trained as to minimize the distances between the samples and their corresponding cluster vectors. We begin with random cluster vectors, and after assigning each sample to its closest vector, cluster vectors are recomputed as the mean of all samples that belong to them. This continues iteratively until reaching a convergence point.

In this work, we use k-means algorithm to cluster the color information of the tissue images, where the color information is represented by red-green-blue (RGB) values. Each cluster vector, which is also composed of RGB values, represents the group of colors.

Obviously k-means algorithm is unsupervised learning and after learning one must assign classes to the determined clusters. In this work we have classes of *cell* and *background*, and our expert assigns them to each cluster vector.

2.2 Artificial Neural Networks

A neural network is composed of nodes (*perceptrons*) that are tied with weighted connections. Each perceptron takes a vector of input values and computes a single output value as the weighted sum of its input values. The output value is activated only if exceeds the threshold defined by an activation function [9, 10]. Figure 2 shows a single perceptron with inputs x_i and output o. In this figure, weights w_i are associated with each input, where w_0 is a bias term.

In this work, we use multilayer perceptrons; the outputs of each layer are connected to the inputs of another layer. In this work, inputs, x_i are the topological metrics, and the output, o, is the class label, indicating whether a cell is cancerous, healthy or generated as synthetically. The input layer is connected to a hidden layer with weights w_{ij} and the hidden layer connects to a output layer with weights v_{ij}. A multilayer network used in this work is illustrated in Figure 3. Note that in cell classification we use six different local metrics which are explained in detail in the next section.

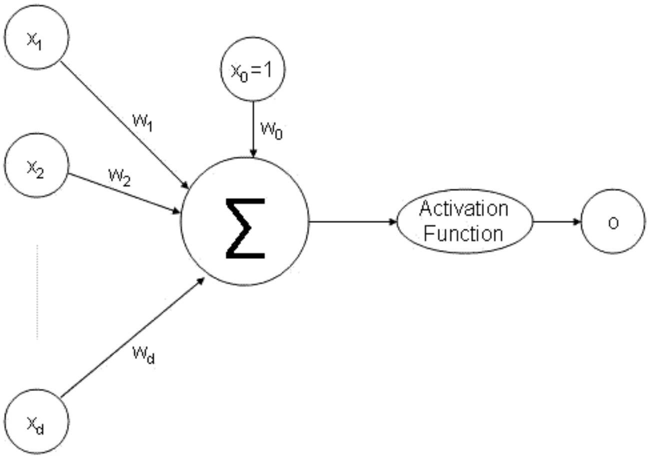

Figure 2: A single perceptron.

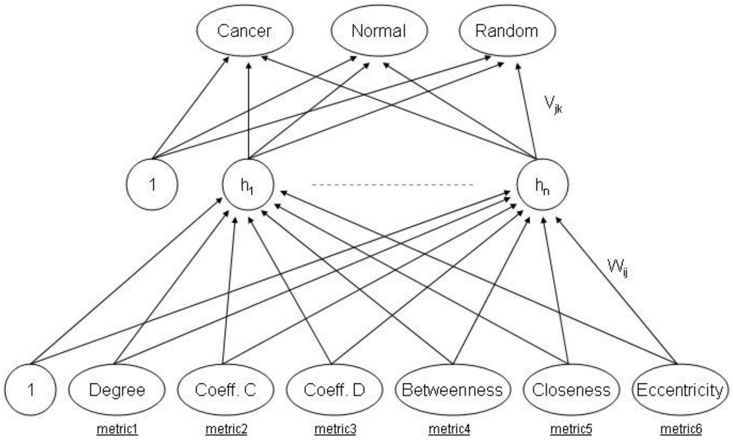

Figure 3: One of the multilayer perceptron used in this work. The inputs are the local metrics defined for the nodes of the extracted graphs. The output indicates whether a cell is cancerous, healthy or generated synthetically.

3 Metrics

Metrics on a graph reflects its topological properties, and provides information of its characteristics. The metrics defined in this section are most commonly used in analyzing the other types of graphs, e.g., Internet, actor or C-elegance worm graphs. These metrics quantify the information about the degree distribution of a node, the connectivity information of its neighbors, and the connectedness information of itself as well as the whole graph. Metrics defined on the nodes are local, but by using statistics, they also provide the global information for the graph. A precise mapping from these metrics to properties of glioma cells is outside the scope of this work and left for future study. We simply use these metrics to identify and distinguish mathematical properties of gliomas from other cell structures.

1. **Degree** is the most trivial metric and it is defined as the number of the connections of a single node for an undirected graph. Its value on a tumor graph is higher, but the higher degree values are not always an indicator of a cancer.

2. **Clustering coefficient** reflects the connectivity information in the neighborhood environment of a node [11]. They provide the transitivity information [13], since it controls whether two different nodes are connected or not, if they are connected to the same node.

We use clustering coefficient C_i which is defined as the percentage of the connections between the neighbors of node i, and it is given as:

$$C_i = \frac{2 \cdot E_i}{k \cdot (k-1)} \tag{1}$$

where k is the number of neighbors of node i and E_i is the existing connections between its neighbors.

Random and scale-free graphs can be distinguished by using the clustering coefficient C. Random graphs have small values of clustering coefficients C, whereas scale-free graphs have larger values than those of the random graphs. We also observe larger values for our tissue images. This indicates the scale-free-ness of our graphs. This also means that our cell-graphs are not *random.*

We also use modified version of the clustering coefficient defined in [11]. Clustering coefficient D_i is defined similar to C_i with an exception. It also considers node i and its connections in the computation of the clustering coefficient [11]. The formula of D_i is given as:

$$D_i = \frac{2 \cdot (E_i + k)}{k \cdot (k+1)} \tag{2}$$

3. **Closeness** and **betweenness** are local metrics that measure the connectedness of a graph [13]. The closeness of a node is the average of the distances between the node and every other nodes except itself. It reflects the centrality property of a single node and smaller values indicate that this node places close to the center of a graph. Betweenness of a node is the total number of the shortest paths that pass through this node. These metrics may indicate the location of a cell within the tumor. For example, having a smaller closeness value or higher betweenness value may suggest that the cell is close to the center of the tumor.

4. **Eccentricity** of a node is a local metric defined as the minimum number of hops required to reach at least 90 per cent of its reachable nodes. The higher values of this metric may indicate the density of the diffuse invasion.

4 Experiments

The experiments are conducted on clinical data for brain tumors. We use the digital images of surgically removed tissues to construct a graph representing the data as explained in Section 2. Each pixel of these images is represented by its RGB values.

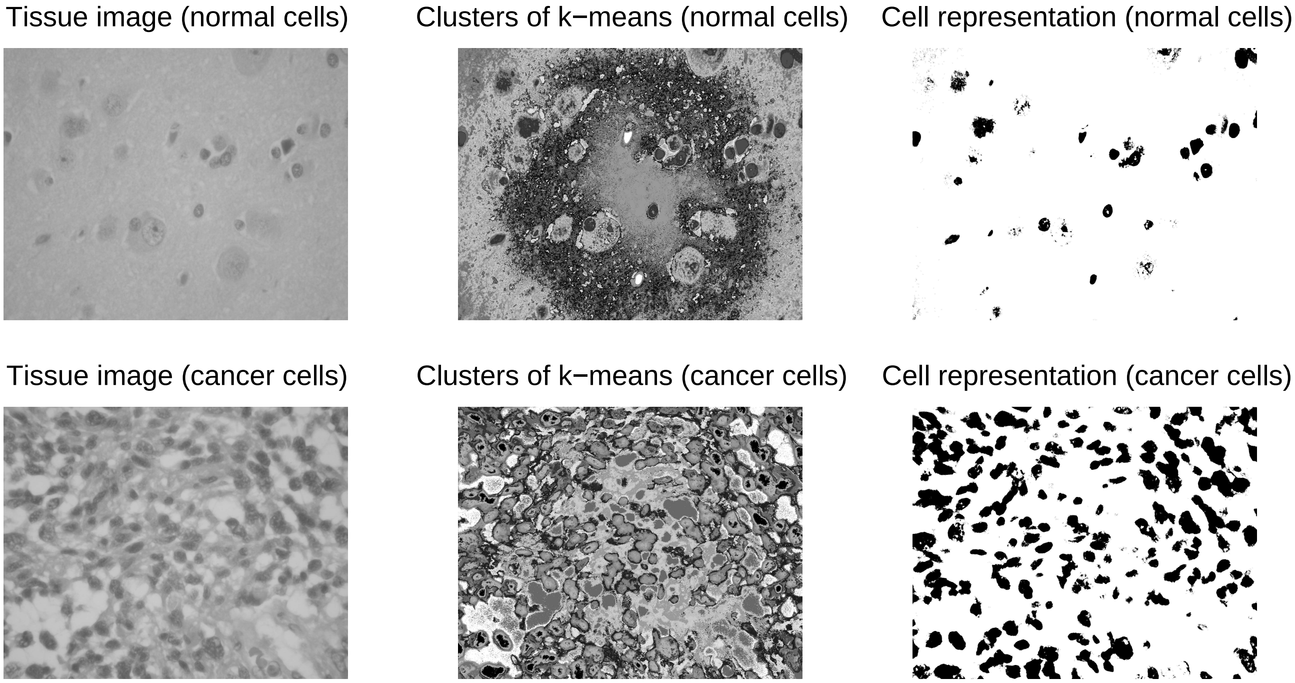
Tissue image (normal cells)
Clusters of k−means (normal cells)
Cell representation (normal cells)
Tissue image (cancer cells)
Clusters of k−means (cancer cells)
Cell representation (cancer cells)

Figure 4: Extracting cell information: k-means algorithm is used to find the clusters and each cluster is assigned as cell or background. k is taken as 17.

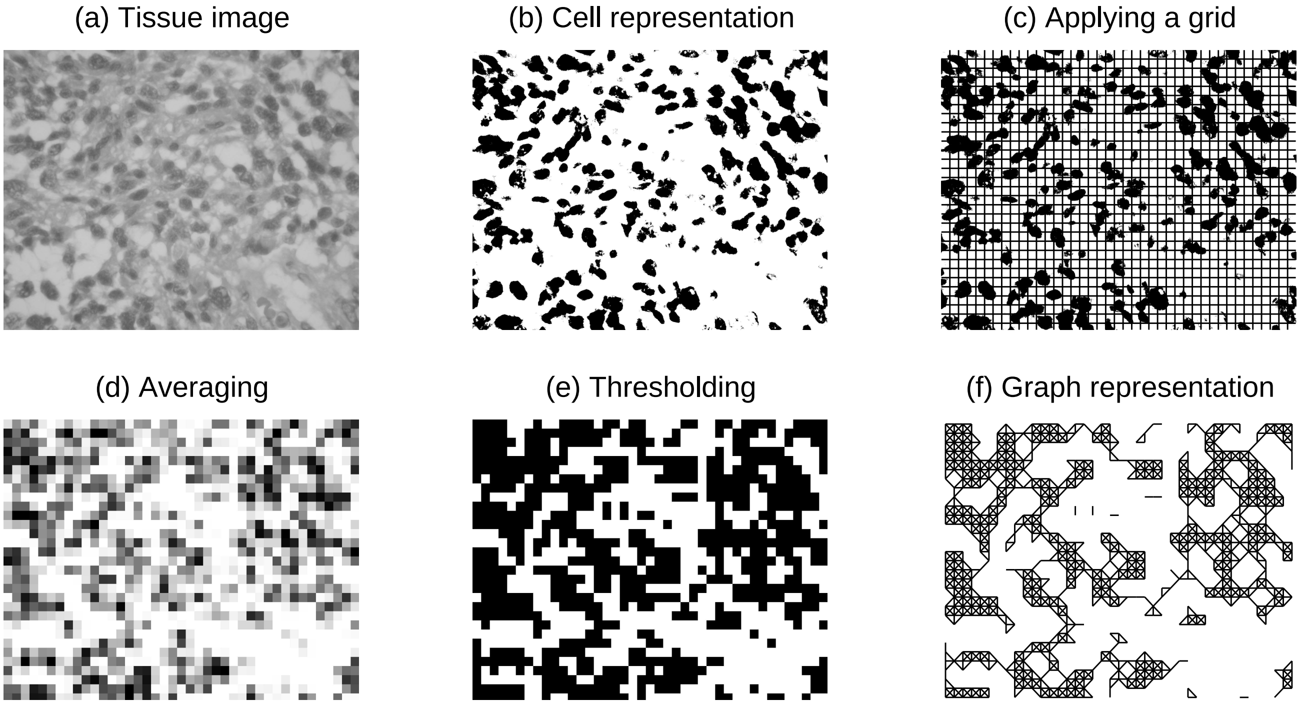

Figure 5: The steps to extract a graph from cell information (grid size = 50, node-threshold = 0.1, edge-threshold = 1). These images are obtained on a cancer cell-graph.

We run the k-means algorithm on the data to learn cluster vectors on training samples and we use these cluster values for our test samples. We have tried various k values and based on the clusters, and based on human expertise have labeled these clusters as either cell or background. We illustrate these steps in Figure 4 for both cancer and normal tissues. Remark that the images in this graph are from the test set and are not used in training. The value of k is selected as 17 in this graph.

After determining the cell and background regions, the nodes are to be extracted on these data. The grid is embedded on them (Figure 5-c), and for each entry of a grid, a probability value of having a cell is computed by averaging the labeled data in the grid entry. Note that we label cell regions as 1, and the background as 0 (Figure 5-d uses gray scale levels to represent the average values). A pair of cells are connected if the distance between them is smaller than a edge-threshold (Figure 5-e). We set these three parameters as follows: the grid size= 50 (i.e., 50 pixels are grouped to represent a cell or not), the node-threshold = 0.1 (i.e., at least 10 per cent of a grid entry should consist of cell regions to being a cell), and the edge-threshold = 1 (i.e., two nodes are to be connected if they are adjacent in the grid.

In the next three subsections, we compare the cell-graphs extracted from the cancerous tissues to three different types of structures. Our first aim is to show that the cell-graphs of cancerous tissues are different than those of healthy tissues. Since the degree distributions of cell-graphs for cancerous and healthy tissues are obviously different, we also examine other dense structures explained below.

The first dense structure we discuss is cell-graphs extracted from inflamed tissues. Our aim is to show that the graph structure of glioma is different than the structure of other biological phenomenon; using cell-graphs they are also distinguished from inflammation.

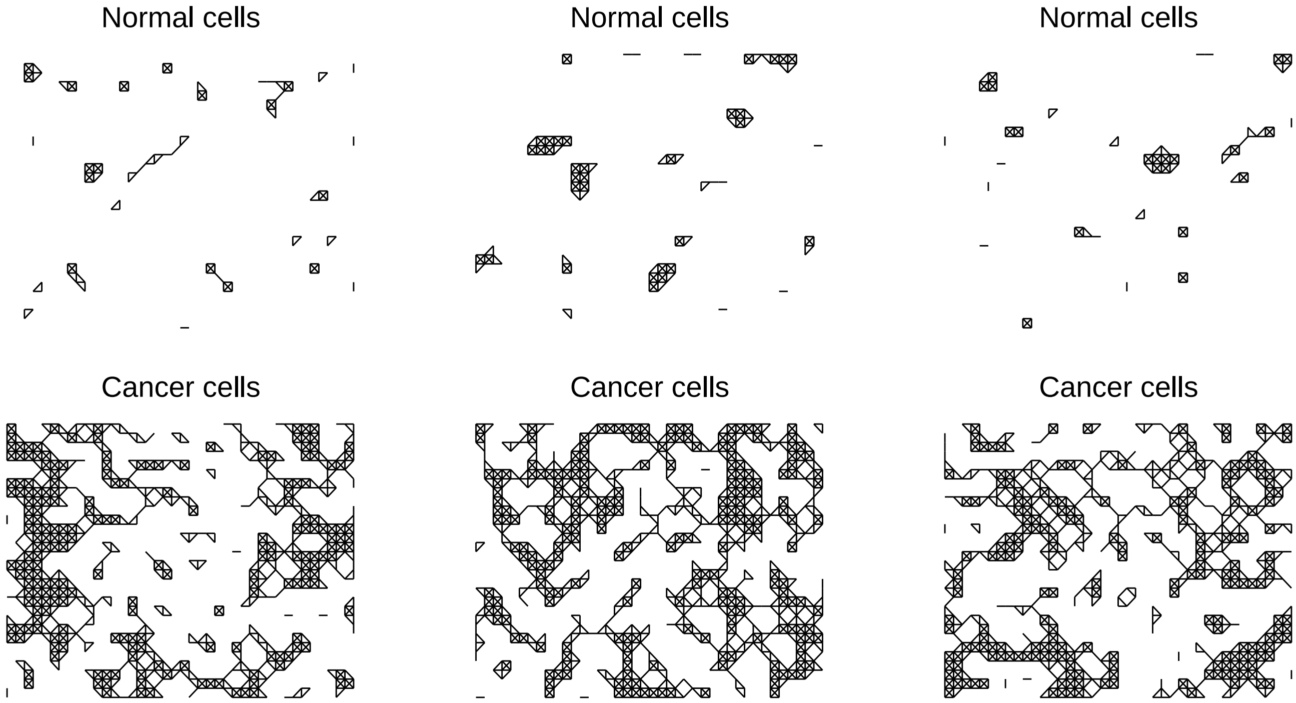

Figure 6: Different cell-graphs representing cancer and normal cells.

Last, we show that cancerous cell-graphs are not random, and show the properties of the scale-free graphs. We compare them with synthetically generated random graphs of the same size. They are completely separable and the clustering coefficient C of the cancerous cell-graphs are much larger than those of the generated random graphs as expected.

4.1 Distinguishing Cancer Cells from Normal Tissue Cells

The extracted cell-graphs for tumor and normal tissues can be seen in Figure 6. The sparsity (density) of the graphs show that the tumor and normal tissues have completely different graphs. We validated this visual observation by computing the metric values to quantify these differences statistically. The data histograms for each metric is shown in Figure 7. These histograms indicate that normal and cancer cells can be distinguished by using these metrics.

4.2 Cancer Cell-graphs v.s. Inflammation Cell-graphs: “Distinguishing” Dense Graphs

We compared the cancer cell-graphs with the inflammation cell-graphs. The images of inflammation and tumor tissues and their corresponding graphs are given in Figure 8. The above two figures correspond to the inflammation data, whereas the other two are for tumor data.

The histograms of the metrics computed on both inflammation and tumor cell-graphs are given in Figure 9. These histograms show that it is not so easy to distinguish them as in previous histograms. We run a classifier algorithm, a multilayer perceptron with 5 hidden units. Its average accuracy results on training and testing sets are given in Table 1. They are more than 75 per cent which indicates the classification is based on the metric values. If it were random, the accuracy results would be

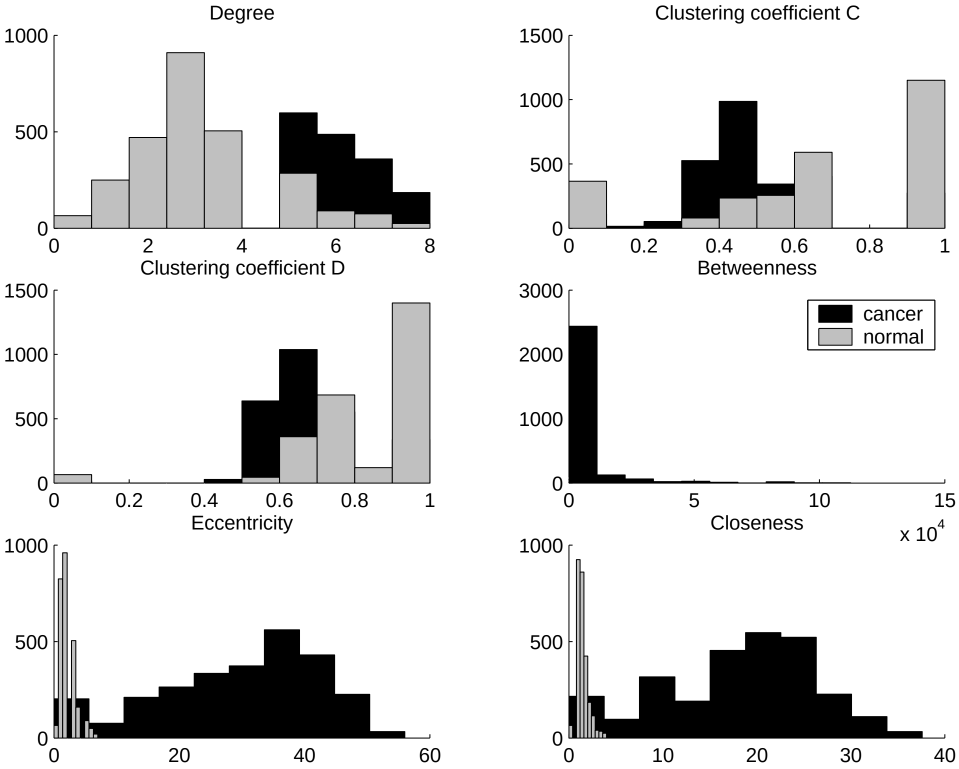

Figure 7: Histograms of the metrics computed for normal and cancer cells.

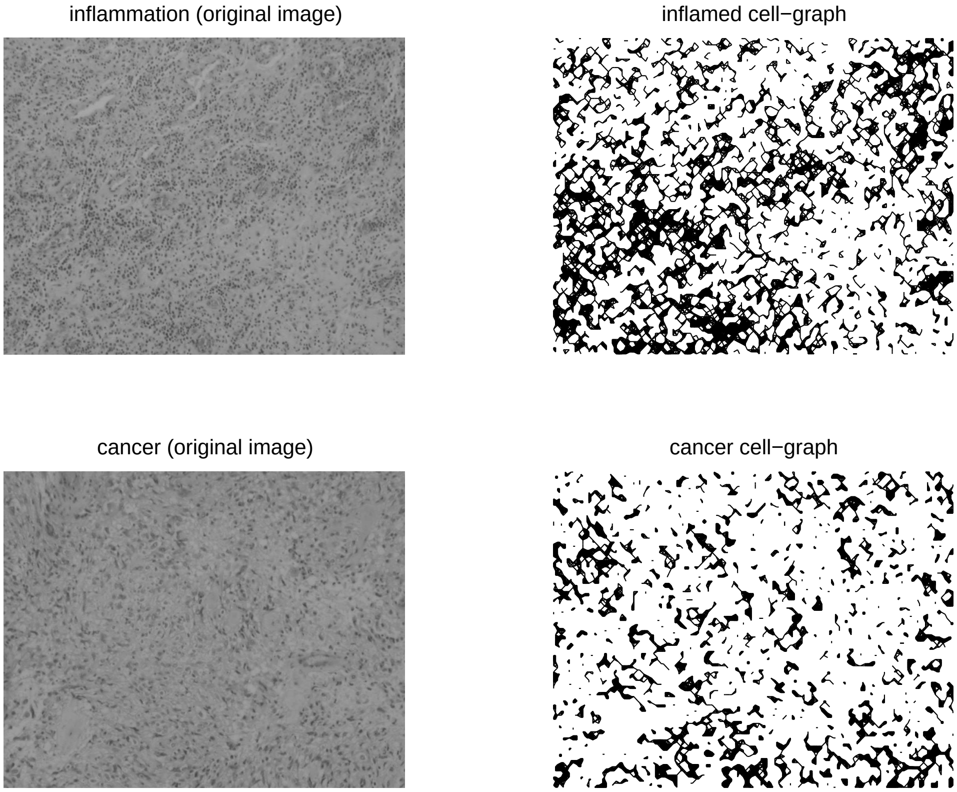

Figure 8: Images of two different inflammatory processes

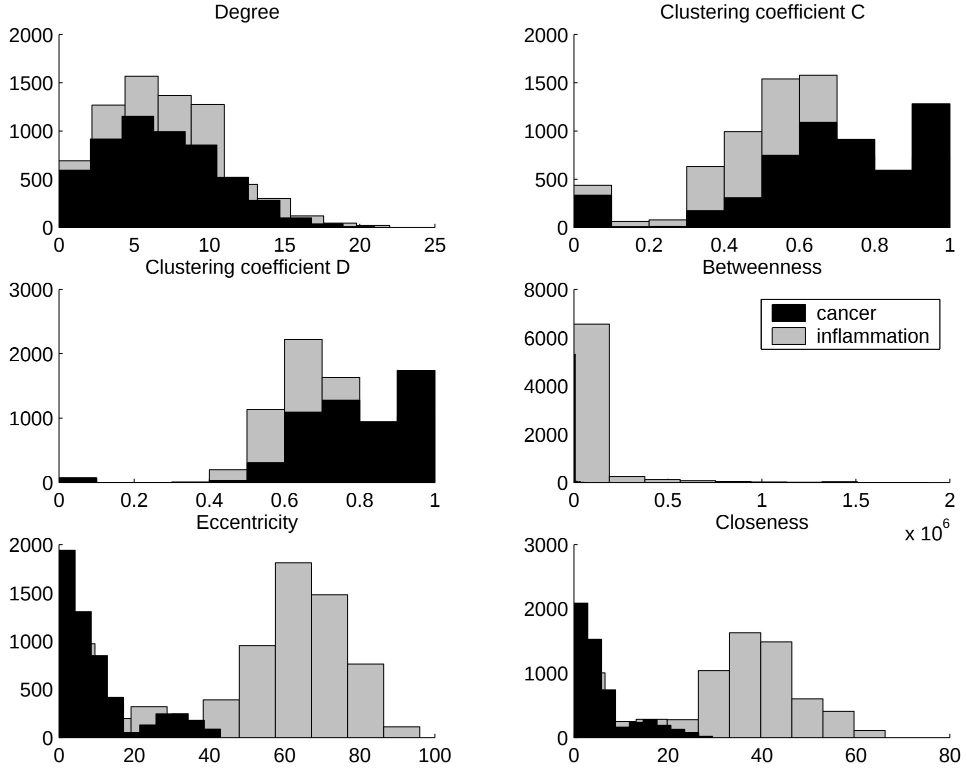

Figure 9: Histograms of the metrics computed for inflammation and cancer cells.

approximately 50 per cent for two classes classification. These results are just the preliminary results and we believe that they will get better as we tune up the system parameters and the resolution of the images.

4.3 Distinguishing Cancer Cell-graph from Random graphs

We also generated random graphs of the same size with the cancer subgraph, and computed the aforementioned metrics on them. The data histograms of their metrics are given in Figure 10. These histograms show that a tumor cell-graph is different than the random graph. We run a classification algorithm to distinguish the cancer and normal cell-graphs as well as the random graphs. We use a multilayer perceptron with 5 hidden units, the accuracy values on the trainig and test sets are given in Table 2.

Table 1: The accuracy values on the training and test sets in the classification of inflammation and tumor cells.

	Average	Standard deviation
Training set	91.23	0.08
Test set	76.83	0.10

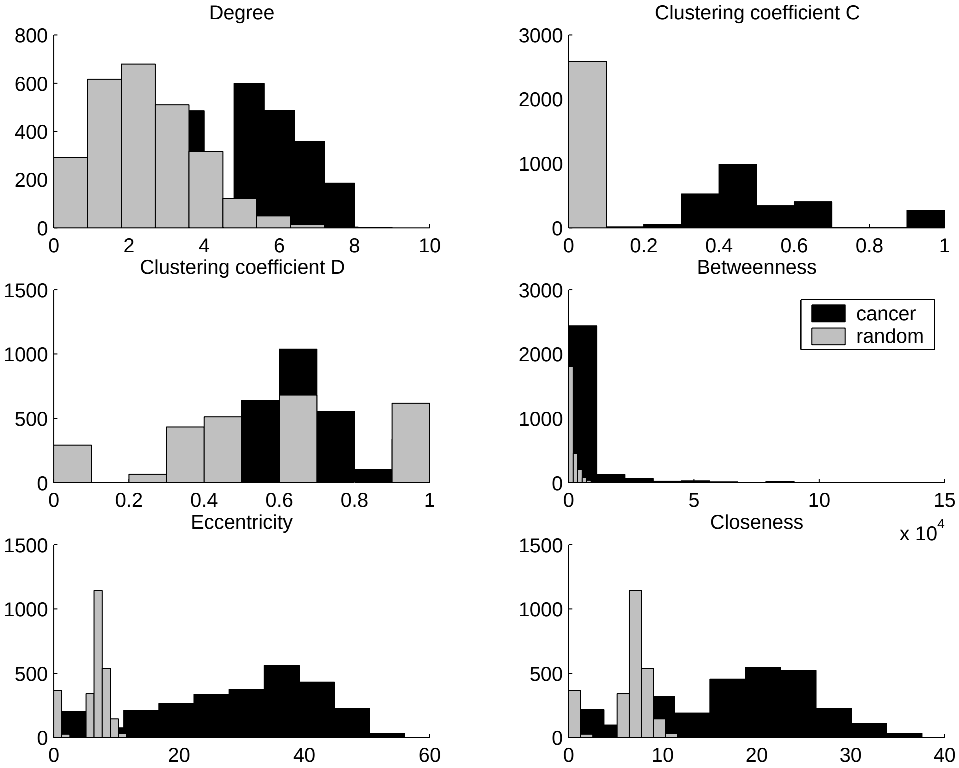

Figure 10: Histograms of the metrics computed for tumor cell-graph and random graph of the same size.

Table 2: The accuracy values on the training and test sets for three classes: normal, cancer, and random.

	Average	Standard deviation
Training set	94.98	0.05
Test set	94.52	0.08

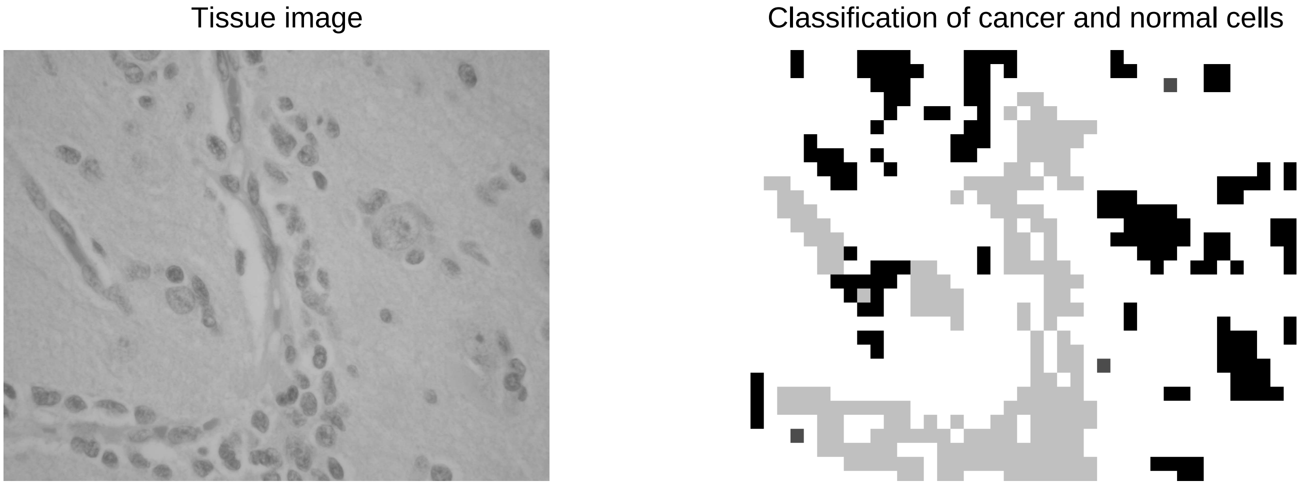

Figure 11: Classification of unlabeled tissue image. In this graph, black regions indicate normal cells, whereas gray ones show cancer cells.

4.4 Mixed Data

Some of the tissue images contains both cancer and normal cells. We also tested our algorithm on these images. This images are not used in training of either k-means algorithm nor multilayer perceptrons. A sample image can be seen in Figure 11. In this figure gray cells represent for cancer cells, where the black ones represent for normal cells.

5 Conclusion and Discussion

This work presents a novel approach for mathematical modeling of diffuse gliomas based on graph theory. It advances the current computational and mathematical modeling approaches by scaling up the cell-graphs with large number of vertices. Our results are preliminary but promising. The graph theoretical model is scalable and used by a machine learning algorithm which can distinguish (1) gliomas from surronding normal tissue, and (ii) gliomas from inflammation. We tested the model and algorithms on real data to validate the proposed approach.

References

[1] E. Rubin and J.L. Farber, *Pathology*, 2nd Ed., Lippincott,PA 1994.

[2] *Cancer Modeling* ed: J. Thompson and B. Brown, Marcel Dekker, Inc. 1987.

[3] M. A. J. Chaplain, "The Mathematical Modelling of Tumor Angiogenesis and Invasion", *Acta Biotheoret.*, 43:387-402, 1995.

[4] A. Anderson, M. Chaplain, E. Newman, R. Steele and A. Thompson, "Mathematical Modelling of Tumor Invasion and Metastasis", *J. Theor. med.* 2:129-165, 2000.

[5] B. Bollabas, Random Graphs (Academic Press, London, 1985).

[6] D. Drasdo, R. Kree and J. S. McCaskill, " Monte-Carlo Approach to Tissue Cell Populations", *Phys. Rev E*, 52(6B):6635-6657, 1995.

[7] S. Turner and J. Sherratt, “Intercellular Adhesion and Cancer Invasion: A Discrete Simulation Using the Extended Potts model”, *J. Theor. Biol.*, 216:85-100, 2002.

[8] J. A. Hartigan and M. A. Wong, “A K-Means Clustering Algorithm”, *Applied Statistics*, vol. 28, pp. 100-108, 1979. *Advances in Physics*, cond–mat/0106144, 2002.

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

[10] A. K. Jain, J. Mao and K. M. Mohiuddin, “Artificial Neural Networks: A Tutorial”, *Computer*, Vol. 29, No. 3, pp. 31–44, 1996.

[11] S. N. Dorogovtsev and J. F. F. Mendes, “Evolution of Networks”, *Advances in Physics*, cond–mat/0106144, 2002.

[12] P. Lantos, D.N. Louis, M. K. Rosenblum, P. Kleihuis, “ Tumors of the Nervous System”, in *Greenfield’s Neuropathology*, 7th Ed. Vol. 2 pp 767-1052 Eds: D. Graham & P. Lantos, Oxford University Press, London 2002.

[13] M. E. J. Newman, “Who is the Best Connected Scientist? A Study of Scientific Coauthorship Networks”, *Phys.Rev.*, cond–mat/0011144, 2001.