A Novel Classification Model for Cancer using Neural Network Classifiers

M.Dhivya,
PG SCHOLAR,
Department of ECE,
Excel Engineering College,
Komarapalayam.
dhivyabme@gmail.com
S.Anbumani,
ASSISTANT PROFESSOR,
Department of ECE,
Excel Engineering College,
Komarapalayam .
anbumaniexcel@gmail.com

ABSTRACT

Cancer is one of the most important health problems that threat the human life. The likelihood of curing cancer increases with its early diagnosis and correct grading, for which histopathological examination is routinely used. The developed novel model uses both structural and statistical pattern recognition techniques to locate and characterize the biological structures in a tissue image for tissue quantification. This approach mainly includes three steps. They are graph generation for tissue images and query glands, localization of key regions, and feature extraction from the key regions. Unlike conventional approaches, this model quantifies the located key regions with structural and textural features extracted from the images. Then based on the extracted key features it classifies the images into two groups low and high grade with the help of SVM (Support Vector Machine) classifiers. The developed model leads to higher classification accuracies, compared against the conventional approaches that use only statistical techniques for tissue quantification.

***Index terms*-histopathological examination, pattern recognition, graph generation, key features, and Support Vector Machine classifiers.**

1 INTRODUCTION

Cancer is a class of diseases characterized by out-of-control cell growth. There are over 100 different types of cancer, and each is classified by the type of cell that is initially affected.The likelihood of curing cancer increases with its early diagnosis and correct grading, for which histopathological examination is routinely used. The number of computational studies on histopathological image analysis is increasing over the past few years.

The main aim of these studies is to automate the diagnosis and grading process for reducing the subjectivity that can be observed in histopathological examination. These studies extract features from a histopathological tissue image and use the features in automated diagnosis and grading.

Digital pathology provides a digital environment for the management and interpretation of pathology information that is enabled by digital slides (virtual slides). The implementation of these systems typically requires a deep analysis of biological deformations from a normal to a cancerous tissue as well as the development of accurate models that quantify the deformations. These de-formations are typically observed in the distribution of the cells from which cancer originates, and thus, in the biological structures that are formed of these cells.

For example, colon adenocarcinoma, which accounts for 90%–95% of all colorectal cancers, originates from epithelial cells and leads to deformations in the morphology and composition of gland structures formed of the epithelial cells (Figure 1). Moreover, the degree of the deformations in these structures is an indicator of the cancer malignancy (grade). Thus, the correct identification of the deformations and their

accurate quantification are quite critical for precise modeling of cancer.

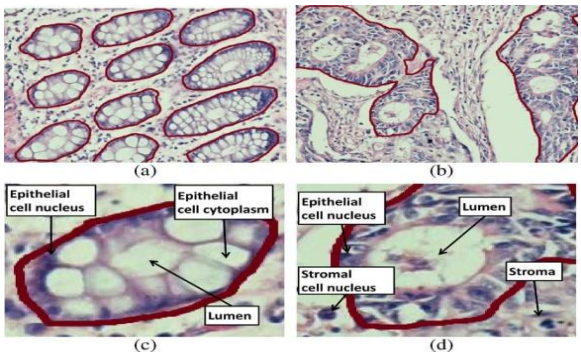

Fig 1 Colon adenocarcinoma changes the morphology and composition of colon glands.

Although digital pathology systems are implemented for different purposes, including segmentation, and retrieval, most of the research efforts have been dedicated to tissue image classification. Compared to traditional pathology, major advantages of digital pathology are that slides can be viewed via computer monitors, archived and retrieved easily and most importantly analysed using software algorithms rather than by manual analysis. Virtual microscopy is the technique for creating (via scanners) and viewing (via software) whole-slide images.

Pattern recognition is a field with in the area of machine learning, and it aims to classify data (patterns) based on either a priori knowledge or on statistical information extracted from the patterns. Most of the pattern recognition methods exist make use of procedures and algorithms.

The classification or description scheme usually uses one of the following approaches: statistical (or decision theoretic) or syntactic (or structural). Statistical pattern recognition is based on statistical characterizations of patterns, assuming that the patterns are generated by a probabilistic system and it is represented by d -features and attributes and viewed as a d-dimensional vector. Structural pattern recognition is based on the structural interrelationships of features recognition and represented as a symbolic data structures, such as strings, trees, or graphs.A wide range of algorithms can be applied for pattern recognition, from very simple Bayesian classifiers to much more powerful neural networks.

2 METHODOLOGY

It is a new approach to tissue image classification. This approach models a tissue image by constructing an attributed graph on its tissue components and describes what a normal gland is by defining a set of smaller query graphs. It searches the query graphs, which correspond to non deformed normal glands, over the entire tissue graph to locate the attributed sub graphs that are most likely to belong to a normal gland structure. Features are then extracted on these sub graphs to quantify tissue deformations, and hence, to classify the tissue. This approach includes three steps: graph generation for tissue images and query glands, localization of key regions (attributed sub graphs) that are likely to be a gland, and feature extraction from the key regions. The figure 2 shows the block diagram of entire system.

A. Tissue graph generation

Graphs are a general and powerful data structure for the representation of objects and concepts. In a graph representation, the nodes typically represent objects or parts of objects, while the edges describe relations between objects or object parts.

This model describes tissue image with an attributed graph G={V,E, μ} where V is a set of nodes , $E \subseteq V \times V$ is a set of edges, and $\mu : V \rightarrow A$ is a mapping function that maps each node $v_i \in V$ into an attributed node label $\alpha_i \in A$.This graph representation relies on locating the tissue components in the image, identifying them as the graph nodes, and as-signing the graph edges between these nodes based on their spatial distribution. However, as the exact localization of the components emerges a difficult segmentation problem, use an approximation that defines circular objects to represent the components.

In order to define these objects, first quantify the image pixels into two groups: nucleus pixels and non-nucleus pixels. For that, separate the hematoxylin stain using the deconvolution method proposed in and threshold it with the Otsu's method. Then, on each group of the pixels, locate a set of circular objects using the circle-fit algorithm. This approximation gives us two groups of objects: one group defined on the nucleus pixels and the other defined on the non - nucleus (whiter) pixels. These groups are herein referred to as "nucleus" and "white" objects.After defining the objects as the graph nodes, encode their spatial relations by constructing a tissue graph using Delaunay triangulation. For an example image given in Figure 3(a), the constructed tissue graph is illustrated in Figure 3(b) with the centroids of the nucleus and white objects (nodes) being shown as black and white circles, respectively.

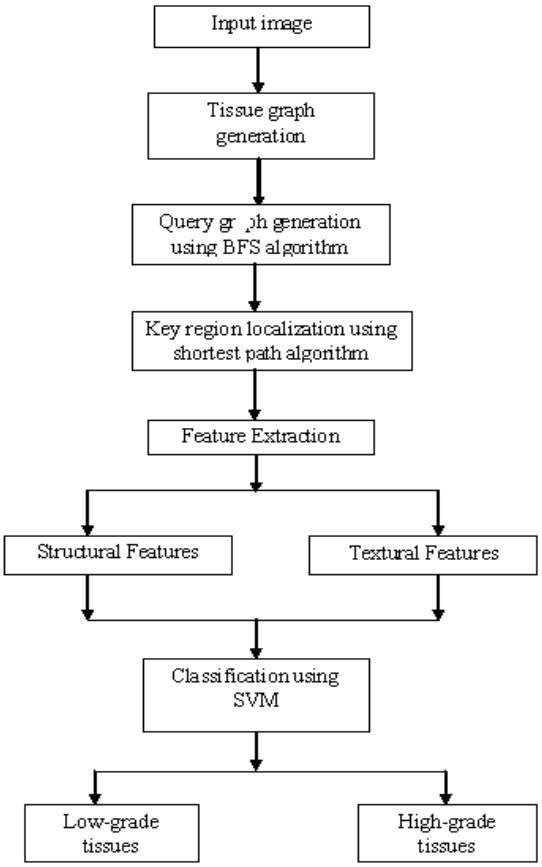

Fig 2 Block diagram of tissue image classification system.

A normal gland is formed of a lumen surrounded by monolayer epithelial cells. The cytoplasms of normal epithelial cells are rich in mucin, which gives them their white-like appearance. Thus, in the ideal case, a query graph consists of many white objects at its center surrounded by a single layer nucleus objects.

Note that there may exist deviations from this ideal case due to noise and artifacts in an image as well as model approximations, as seen in Figure 2.2(c). Subgraphs generated from normal tissues show an object distribution similar to that of a query graph. On the other hand, the object distribution of cancerous tissue subgraphs becomes different since colon adenocarcinoma causes deviations in the distribution of epithelial cells and changes the white-look appearance of their cytoplasms (epithelial cells become poor in mucin). The graph edit distance features will be used to quantify this difference.

B.Query graph generation

Query graphs are the sub graphs that correspond to normal gland structures in an image. To define a query graph G_s on the tissue graph G of a given image, select a seed node (object) and expand it on the tissue graph G using the breadth first search (BFS) algorithm until a particular depth is reached.

In graph theory, breadth-first search (BFS) is a strategy for searching in a graph when search is limited to essentially two operations: (a) visit and inspect a node of a graph; (b) gain access to visit the nodes that neighbor the currently visited node. The BFS begins at a root node and inspects all the neighboring nodes. Then for each of those neighbor nodes in turn, it inspects their neighbor nodes which were unvisited, and so on. Compare BFS with the equivalent, but more memory efficient Iterative deepening depth-first searches and contrast withdepth-first search.

Then, take the visited nodes and the edges between these nodes to generate the query graph G. In this procedure, the seed node and the depth are manually selected, considering the corresponding gland structure in the image. Figure 3(c) shows this query graph generation on an example image; here black and white indicate the selected nodes and edges whereas gray indicates the unselected ones.

The Figure 3 shows that (a) is the example normal tissue image, (b) is the tissue graph generated for this image, and (c) is a query graph generated to represent a normal gland. The node labels are indicated using four different representations and the orders in which the nodes are expanded are given inside their corresponding objects.

Subsequently, the mapping function μ attributes each selected node with a label according to its object type and the order in which this node is expanded by the BFS algorithm. In particular, define four labels: $\alpha_{n\text{-}in}$ and $\alpha_{\omega\text{-}in}$ for the nucleus and white objects whose expansion order is less than the BFS depth and $\alpha_{n\text{-}out}$ and $\alpha_{n\text{-}out}$ for the nucleus and white objects whose expansion order is equal to the BFS depth.

The query graph generation and labeling processes are illustrated in Figure 4.In this figure, a query graph is generated by taking the dash bordered white object as the seed node and selecting the depth as 4. This illustration uses a different representation for the nodes of a different label: it uses black circles for $\alpha_{n\text{-}in}$ white circles for $\alpha_{\omega\text{-}in}$, black circles with green borders for $\alpha_{n\text{-}out}$, and white circles with red borders for $\alpha_{\omega\text{-}out}$. It also indicates the expansion order of the selected nodes

inside their corresponding circles; note that the order is not indicated for the unselected nodes.

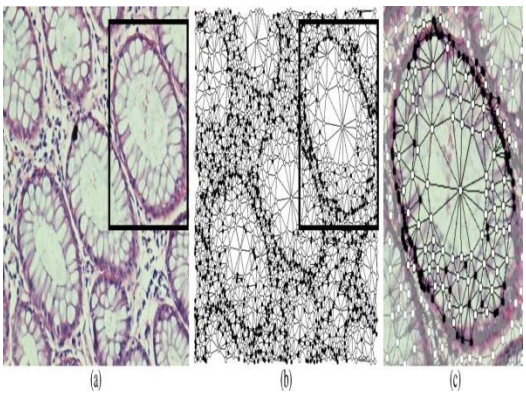

Fig 3 An illustration of the graph generation step.

The search process for key region localization uses the same algorithm to obtain sub graphs to which a query graph is compared. However, these sub graphs are generated by taking each object as the seed node and selecting the depth as the same with that of the query graph. Thus, the search process involves no manual selection. The search process is detailed in the key region localization process.

C.Key region localization

The localization of key regions in an image includes a search process. This process compares each query graph G with sub graphs G_t generated from the tissue graph G of the image and locates the ones that are the N- most similar to this query graph.

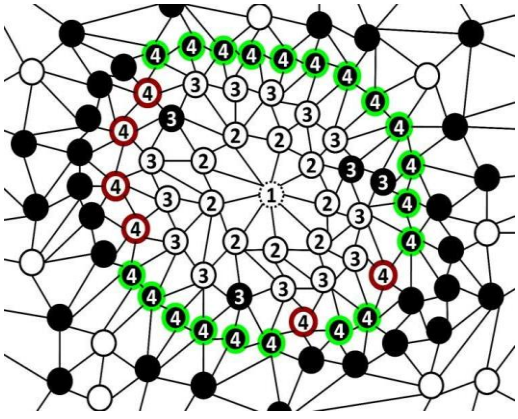

Fig 4 An illustration of generating a query graph.

. The regions corresponding to the located subgraphs are then considered as the key regions. Since a query graph is generated as to represent a normal gland, the located subgraphs are expected to correspond to the regions that have the highest probability of belonging to a normal gland.

Typically, the subgraphs located on a normal tissue image are more similar to the query graph than those located on a cancerous tissue image. Thus, the similarity levels of the located subgraphs together with the features extracted from their corresponding key regions are used to classify the tissue image.

The search process requires inexact graph matching between the query graph and the subgraphs, which is known to be an NP-complete problem. Thus, use an approximation together with heuristics on the subgraph definition to reduce the complexity due to polynomial time.

1) Query graph search

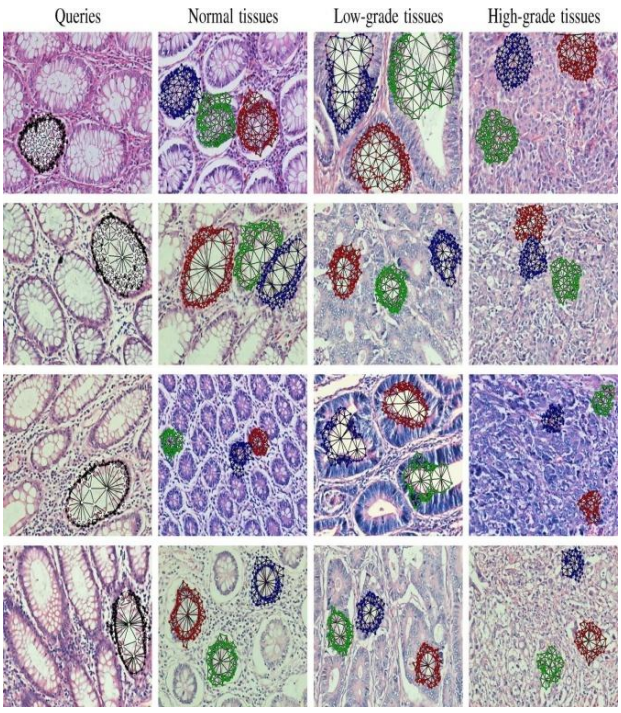

Fig 5 The query graphs generated as a reference for a normal gland structure and the subgraphs located in example normal, low-grade cancerous and high-grade cancerous tissue images.

Let $G_s=\{V_s,E_s,\mu_s\}$ be a query graph and $v_s \epsilon\ V_s$ be its seed node from which all the nodes in V_s are expanded using the BFS algorithm until the graph depth d_s is reached. In order to search this query over the

entire tissue graph G={V,E, μ} , first enumerate candidate subgraphsG_t={V_t,E_t,μ} from the graph G.

For that follow a procedure similar to the one that used to generate the query. Particularly, take each node$v_i \epsilon$ V that has the same label with $v_s$as a seed node and expand this node using the BFS algorithm until the query depth d_s .Note that the use of the same BFS algorithm with the same type of the seed node and the same graph depth, which is used in generating the query graph G_s, prunes many possible candidates, and thus, yields a smaller candidate set. The nodes of the candidate subgraphs are also attributed with the labels in A={ $\alpha_{n\text{-}in}$, $\alpha_{\omega\text{-}in}$, $\alpha_{n\text{-}out}$, $\alpha_{\omega\text{-}out}$ } using the mapping function μ,which was used to label the nodes of the query graphs.

In the Figure 5 first image of each row shows the query graph on the image from which it is taken where as the remaining ones show the three-most similar subgraphs to the corresponding query graph. In the Figure 5, the subgraphs of the same image are shown with different colours (red for the most similar subgraph, green for the second-most similar subgraph, and blue for the third-most similar subgraph). After they are obtained, each of the candidate subgraphs G_t is compared with the query graph G using the graph edit distance metric and the most similar N nonoverlapping subgraphs are selected. To this end,start the selection with the most similar subgraph and eliminate other candidates if their seed node is an element of the selected subgraph.

Then repeat this process N times until the N-most similar subgraphs are selected. For different query graphs, Figure 2.4 presents the selected subgraphs in example tissue images; here only three-most similar subgraphs are shownNote that although there may not exist N .normal gland structures in an image, our algorithm locates the N-most similar sub-graphs, some of which may correspond to either more deformed gland structures or false glands.In this model do not eliminate these glands(subgraps) since the edit graph distance between the query graphs and the subgraphs of more deformed glands are expected to be higher and this will be an important feature to differentiate normal and cancerous tissue images.

Indeed, our experiments reveal that this feature is especially important in the correct classification of high- grade cancerous tissues since subgraphs generated from these tissues are expected to look less similar to a query graph, leading to higher graph edit distances. These higher distances might be effective in defining more distinctive features. Also note that sometimes there may existN normal gland structures in an image but the algorithm may incorrectly locate subgraphs that correspond to nongland benign tissue regions.

2) Graph edit distance calculation

To select the subgraphs G_t={V_t,E_t,μ} that are most similar to a query graph G_s={V_s,E_s,μ} ,the proposed model uses the graph edit distance algorithm, which gives error-tolerant graph matching. The edit graph distance quantifies the dissimilarity between a source graph G_s and a target graph G_t by calculating the minimum cost of edit operations the dissimilarity between a source graph G_s to transform it into G_t.This algorithm defines three operations : insertion ($\varepsilon \rightarrow v_t$) that inserts a target node into G_s,delection ($v_s \rightarrow \varepsilon$) that deletes a source node from G_s and substitution ($v_s \rightarrow v_t$) that changes the label of a source node in G_s to that of a target node in G_t. Note that these operations allow matching different sized graphs G_s and $G_t$with each other. As illustrated in Figure 3.3, the proposed graph representations together with this graph edit distance algorithm make it possible to match the query gland regions with the regions of different sizes and orientations.

Let (e_1 , ...e_i , ... , e_n) ϵ β denote a sequence of operations e_i that transforms G_s into G_t .

The graph edit distance *dist* (G_s,G_t) is then defined as,

$$dist G_s, G_t = \min_{(e1,\ldots..ei,\ldots,en)\in\beta} \sum_n \cos(ei) \qquad (1)$$

Where $\cos(ei)$ is the cost of the operation e_i. Since finding the optimal sequence requires an exponential number of trials with the number of G_s and G_t.This algorithm decomposes the graphs G and G_t into a set of subgraphs each of which contains a node in the graph and its immediate neighbors. Then, the algorithm transforms the problem of graph matching into an assignment problem between the subgraphs of G and G_t and solves it using the Munkres algorithm.

Then shortest path can be obtained by using dijkstra shortest path algorithm.

D.Feature extraction and classification

Feature extraction is the transformation of the original data (using all variables) to a dataset with a reduced number of variables. In feature extraction, all available variables are used and the data are transformed (using a linear or nonlinear transformation) to a reduced dimension space. First characterize a tissue image I by extracting two types of local features and classify the image using a linear kernel support vector machine (SVM) classifier. The first type is used to quantify the

structural tissue deformations observed in the image .To quantify them, graph matching's used. However, as a standard SVM classifier does not work with the graphs, we embed the graph edit distances of the matching's in a feature vector D.

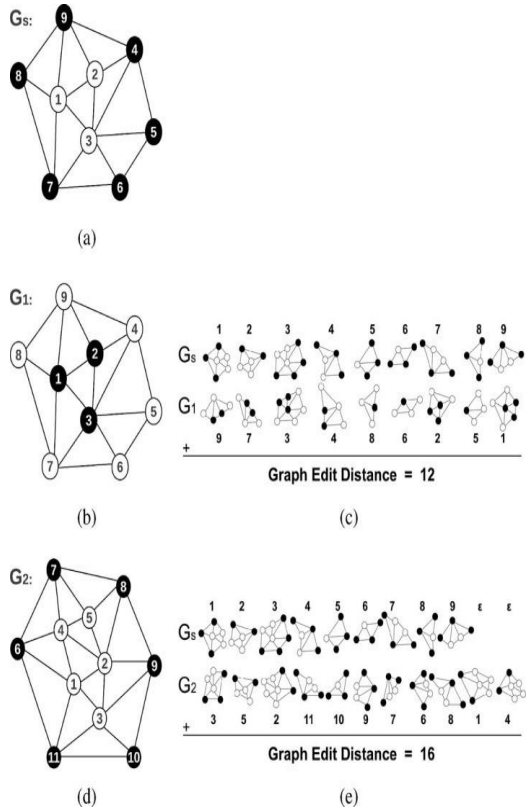

Fig 6 An illustration of minimum distance calculation.

To do that, for each query graph G_s, calculate the average of the be a set of the N-most similar subgraphs of the image I. Let Gs= $\{Gst\}$be a set of the N-more similar subgraph for the query graph G . The average graphs edit distance d_s for this query graph is,

$$d_s = \frac{1}{N}\sum_{t=1}^{N} dist(Gs, Gst) \qquad (2)$$

Where dist (Gs, Gst)is the graph edit distance between the query graph G_S the subgraph G_{st} .Then ,the structural tissue deformations in the image I are characterized by defining the feature vector D=[$d_1 \ldots d_s \ldots d_S$]'.he second feature type is used to quantify textural changes observed in the key regions. In our model; we focus on the outer parts of the key regions. The motivation behind this is the fact that changes caused by colon adenocarcinoma are typically observed in epithelial cells, which are lined up at gland boundaries.

To extract the second type of features, locate a window on the outer nodes of the subgraphs and extract four simple features on the window pixels that are quantized into three colors using k-means. The first three features are the histogram ratios of the quantized pixels and the last one is a texture descriptor (J-value) that quantifies their uniformity. Note that the three colors correspond to white, pink, and purple, which are the dominant colors in a tissue stained with hematoxylin-and-eosin.

3 RESULTS AND DICUSSION

In this model the structural features are represented by graphs and textural features represented by key features as d-dimensional vectors. Mainly the developed model uses both structural and textural feature to classify the images into two groups, low and high graded tissue images based on various features extracted from the images. The features extracted from the images are contrast, energy, homogeneity, correlation. Then the image classification is mainly performed with the help of SVM classifiers. If the SVM classifier output is classes zero means it is a low graded image otherwise it is classified as a high graded tissue image. From classification results the low graded images are seems to be a normal images and high graded images are deviate from normal images.

The accuracy of the image also calculated based on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative values. The above values are defined based on confusion matrix .The simulation result can be obtained with the help of image processing tool. The table 3.1 and 3.2 shows that the feature and accuracy value comparisons.

4 CONCLUSION

The developed novel model that makes use of both structural and statistical pattern recognition techniques for tissue image classification. This model represent a tissue image as an attributed graph of its components and characterize the image with the properties of its key regions. The main contribution of this work is on the localization and characterization of the key regions. The proposed model uses inexact graph matching to locate the key regions.

To this end, it defines a set of query graphs as a reference to a normal gland structure and specifies the key regions as the sub graphs of the entire tissue graph that are structurally most similar to the quergraphs.Then, our model characterizes the key regions using the graph edit distances between the query graphs and their most

similar subgraphs as well as extracting textural features from the outer parts of the key regions.

Table 1 Feature value comparison

Extracted feature values	Low grade	High grade
Contrast	0.8593	0.4672
Correlation	0.8866	0.84430
Energy	0.0818	0.8861
Homogensity	0.1332	0.8256

Table 2 Accuracy comparison

Output image	Sensitivity	Specificity	Accuracy
Lowgrade	80	50	75
Highgrade	90	50	83

Then the classification is performed with the help of SVM classifiers. The proposed model provide improved classification accuracies compare to conventional classification methods, that uses only statistical pattern recognition for tissue image classification.

4 FUTURE WORK

Let us hope for further contribution can be implemented with feed forward neural network classifier, to improve the speed of operation and also improve number of classification range.

REFERENCES

[1] Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in Proc.Biomed Imag.: From Nano to Macro, 2008, pp. 496–499.

[2] D. Altunbay, C. Cigir, C. Sokmensuer, and C.Gunduz-Demir, "Color graphs for automated cancer diagnosis and grading," IEEE Trans.Biomed. Eng., vol. 57, no. 3, pp. 665–674, Mar. 2010.

[3] A.N. Basavanhally, S. Ganesan, S. Agner, J. P. Monaco, M. D. Feldman, J. E. Tomaszewski, G. Bhanot, and A. Madabhushi, "Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology," IEEE Trans.Biomed. Eng., vol. 57, no. 3, pp. 642– 653, Mar. 2010.

[4] ErdemOzdemir and CigdemGunduz-Demir*, Member, "A Hybrid Classification Model for Digital Pathology Using Structural and Statistical Pattern Recognition,'' IEEE transactions on medical imaging, vol. 32, no.2, february 2013.

[5] P.-W. Huang and C.-H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," IEEE Trans. Med. Imag., vol. 28, no. 7, pp. 1037–1050, Jul. 2009.

[6] M. Jondet, R. Agoli-Agbo, and L. Dehennin, "Automatic measure-ment of epithelium differentiation and classification of cervical intra-neoplasia by computerized image analysis," Diagnostic Pathol., vol. 5, no. 7, 2010.

[7] A. Noma, A. B. V. Gracianoa, R. M. Cesar, L. A. Consularo, and I. Bloch, "Interactive image segmentation by matching attributed rela-tional graphs," Pattern Recognit., vol. 45, pp. 1159–1179, 2012.

[8] E.Ozdemir, C. Sokmensuer, and C. Gunduz-Demir, "A resampling-based Markovian model for automated colon cancer diagnosis," IEEETrans. Biomed. Eng., vol. 59, no. 1, pp. 281–289, Jan. 2012.

[9] K. Riesen and H. Bunke, "Approximate graph edit distance computa-tion by means of bipartite graph matching," Image Vis. Comput., vol. 27, p. 950, 2009.

[[10] O. Sertel, J. Kong, H. Shimada, U. V. Catalyurek, J. H. Saltz, and M. N.Gurcan, "Computer-aided prognosis of neuroblastoma on whole slide images: Classification of stromal development," Pattern Recognit., vol. 42, no. 6, pp. 1093–1103, 2009.

[11] B.Tosun and C. Gunduz-Demir, "Graph run-length matrices for histopathological image segmentation," IEEE Trans. Med. Imag., vol. 30, no. 3, pp. 721–732, Mar. 2011.

[12] Y. Wang, D. Crookes, O. S. Eldin, S. Wang, P. Hamilton, and J. Diamond, "Assisted diagnosis of cervical intraepithelial neoplasia (CIN)," IEEE J. Sel. Topics Signal Process., vol. 3, no. 1, pp. 112121, Feb.2009.