Improving Renal Cell Carcinoma Classification by Automatic Region of Interest Selection

Qaiser Chaudry, S. Hussain Raza, Yachna Sharma, Andrew N. Young, and May D. Wang

Abstract— **In this paper, we present an improved automated system for classification of pathological image data of renal cell carcinoma. The task of analyzing tissue biopsies, generally performed manually by expert pathologists, is extremely challenging due to the variability in the tissue morphology, the preparation of tissue specimen, and the image acquisition process. Due to the complexity of this task and heterogeneity of patient tissue, this process suffers from inter-observer and intra-observer variability. In continuation of our previous work, which proposed a knowledge-based automated system, we observe that real life clinical biopsy images which contain necrotic regions and glands significantly degrade the classification process. Following the pathologist's technique of focusing on selected region of interest (ROI), we propose a simple ROI selection process which automatically rejects the glands and necrotic regions thereby improving the classification accuracy. We were able to improve the classification accuracy from 90% to 95% on a significantly heterogeneous image data set using our technique.**

I. Introduction

Renal Cell Carcinoma (RCC) is the most common type of kidney cancer [1, 2]. It accounts for 90% of all kidney cancer. Every year, about 32,000 people in the United States are diagnosed with renal cell carcinoma. Like almost all cancers, renal cell cancer is most likely to be successfully treated if detected early. RCC includes several histopathological subtypes, defined by the World Health Organization (WHO) classification system [3]; the most common subtypes are Clear Cell RCC (83%), Papillary RCC (11%) and Chromophobe RCC (2%). Renal Oncocytoma is a benign renal epithelial tumor with several clinical and morphologic features in common with chromophobe RCC (4%). Different subtypes of RCC are treated differently by clinicians. Thus, identification of RCC subtypes is extremely important for renal cancer treatment. However, manual classification is time consuming and prone to user subjectivity.

Computer aided diagnostic (CAD) tools are gaining popularity since they reduce inter-user and intra-user variability and assist the pathologist in making a quick diagnosis. In [4], authors have concluded that CAD can be used to improve radiologists' performance in breast cancer diagnosis. An important requirement for these diagnostic tools is to provide accurate and consistent results for biological images with varying features, illumination and staining differences. We have designed an automated classification algorithm that classifies the renal cell carcinoma images into four subtypes with minimal user interaction. In this work, we show that automated region selection, when applied before classification, can improve classification accuracy to a great extent with reduced computation time. The classification accuracy achieved with our improved method is 95%, thus showing the potential for future clinical use.

II. Background

There have been several research endeavors for developing an automated diagnosis system, where the main focus is on separating the cancerous images from the normal images. For example, in [5], mathematical morphology has been used to classify the images as cancerous or non-cancerous. In [6], multi-spectral analysis has been used to classify the prostate biopsy images as containing stroma, benign prostatic hyperplasia, prostatic intraepithelial neoplasia and prostatic carcinoma. Features such as entropy, contrast, and angular second moment were derived using a co-occurrence matrix in [7]. These features based on texture analysis were used to classify the colon mucosa into cancerous and non-cancerous categories. In [8], an improvement in the classification accuracy of colon cancer is reported by using a fractal dimension along with conventional texture analysis. Another work [9] pertains to the development of a machine vision system using morphological and texture characteristics for quantifying tissue composition to aid in automatic identification of prostate lesions. The classification of prostate tissue was based on tissue morphological characteristics assuming larger lumen areas for normal tissue. Statistics from the gray level co-occurrence matrix (GLCM) have been used before for classification. For instance, texture classification based on a combination of wavelet statistical features and co-occurrence features has been reported in [10]. In other work [11], texture features are

Manuscript received June 15, 2008. This work was supported in part by grants from National Institutes of Health (Bioengineering Research Partnership R01CA108468, P20GM072069, Center for Cancer Nanotechnology Excellence U54CA119338), Georgia Cancer Coalition (Distinguished Cancer Scholar Award to Professor Wang), Hewlett Packard, and Microsoft Research. The authors would also like to thank Dr. R. Mitchell Parry for his valuable insight during the preparation of this manuscript.

Qaiser Chaudry is with the Georgia Institute of Technology, Atlanta, GA 30332 USA (phone: 404-542-2998; e-mail: qaiser@gatech.edu).

Syed Hussain Raza is with the Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: hussain.raza@gatech.edu).

Yachna Sharma is with the Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: ysharma3@gatech.edu).

Andrew N Young is with the Emory University, Atlanta, GA 30332 USA (e-mail: ayoun01@emory.edu,).

May Wang is with Georgia Tech and Emory University, Atlanta, GA 30332 USA (phone: 404-274-4625; e-mail: maywang@bme.gatech.edu).

calculated on the expansion wavelet coefficients of the images. In [12], the GLCM has been used for the classification of cell nuclei in Pap smear images. In this work, we focus on nuclear segmentation and derive features for each segmented nuclei with several angle and distance measures.

All the above mentioned works have mainly focused on binary classification of normal tissue from the cancerous regions; however, there are no significant efforts on cancer subtype classification. Our previous works [13, 14] attempted computer aided automatic RCC subtype classification. In [13], we used the knowledge-based RCC features to obtain high classification accuracy using our test data set which was selected by the pathologist as a good representative of each RCC subtype. However, in practice clinical image data is highly heterogeneous with significant variations in the images of each RCC subtype. Our algorithm in [13] gives reduced classification accuracy when used for significantly heterogeneous images within each subtype class. In addition to the heterogeneity, the tissue samples also contained necrotic regions which contribute to the reduced accuracy. In [14], we designed a new methodology to overcome the reduced accuracy in the presence of heterogeneous data. We extracted features using a combination of morphological analysis, wavelet analysis and texture analysis. We achieved classification accuracy of about 90% with a simple Bayesian classifier.

In our present work, we augment our knowledge-based classification system with automatic region of interest (ROI) selection from heterogeneous images. We have achieved an improvement in classification accuracy. In [14], the accuracy was 90% with features selected from textural, morphological and wavelet analysis. In this work, we report an accuracy of 95% with only four features and no wavelet analysis. Thus we have achieved higher accuracy with reduced computation time. Our work clearly demonstrates the importance of intelligent ROI selection to reduce computation time and increase classification accuracy.

III. Methodology

The image dataset consists of standard photo micrographs of hematoxylin & eosin (H & E) stained biopsy tissue sections as shown in Figure 1. A flowchart of the overall methodology is shown in Figure 2. We first perform color segmentation for each image and then convert this image into four-level grayscale images (one level for each color class). Automatic ROI segmentation of these images is performed to reject necrotic regions. The ROI masks and the grayscale images were used to compute statistical features. These features were then used to train the Bayesian classifier and classification of the unknown images into subtypes of the RCC. We will describe the detailed processing steps in the remainder of this section.

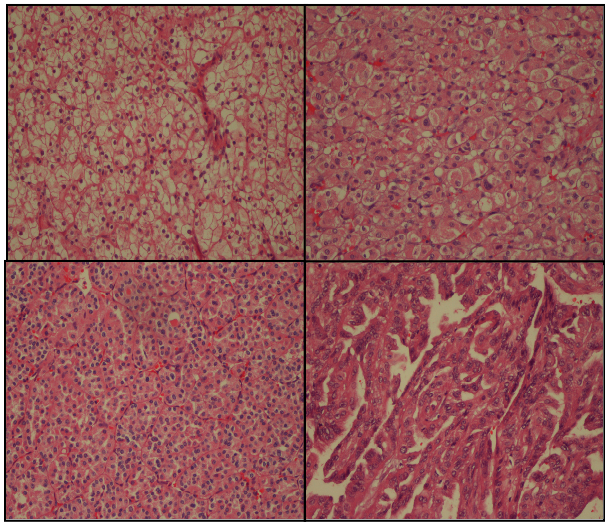

Fig. 1. Clockwise from top-left: Clear cell (CC), Chromophobe (CHR), Papillary (PAP), Oncocytoma (ONC).

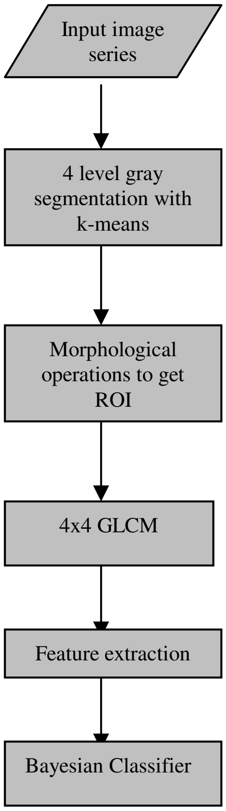
Input image series
4 level gray segmentation with k-means
Morphological operations to get ROI
4x4 GLCM
Feature extraction
Bayesian Classifier

Fig. 2. Flowchart of the overall methodology.

A. *Image Acquisition*

The tissue samples used in this study were obtained from renal tumors resected by total nephrectomy. Standard pathological procedures were followed to fix, process, section and stain the tissue. Microscopic sections were sliced after embedding the histological samples in paraffin. The sliced sections were then stained with hematoxylin & eosin. All tumors were diagnosed by board-certified anatomic pathologists using WHO histopathological criteria [3]. Representative photomicrographs were obtained at 200x total magnification and 1200 x 1600 pixels resolution for analysis.

For our study, we used a set of 48 images with 12 samples from each subclass; clear cell RCC (CC), papillary RCC (PAP), Chromophobe RCC (CHR), and renal Oncocytoma (ONC). A sample of representative images from each subclass is shown in Figure 1. The images were selected with a special emphasis on heterogeneity within each image which can be seen in the Figure 1.

B. *Image Segmentation*

The H&E staining in presence of red blood cells and the background reflects four distinct colors in the acquired images. The color and intensity of the images, however varies significantly based on the variations in sample preparation and the image acquisition process. Therefore, to be consistent with tissue staining, we first segmented the images into four-level grayscale images each level representing a mask for one category of objects. The four categories are nuclei, gland, cytoplasm and red blood cells. A large intra-sample color and intensity variation necessitates some intelligent processing to segment the RGB images into quantized grayscale images representing region masks. For this purposes, we used K-means clustering, a widely used algorithm for multispectral data [15]. K-Means clustering defines the cluster of colored pixels by reducing the objective function, given by

$$E = \sum_{i=1}^{k} \sum_{x_j \in s_i} \left(x_j - C_i \right)^2 \qquad (1)$$

where x_j is a pixel belonging to the cluster S_i with cluster mean C_i. For our RCC subtype images we start with the fixed initial values of the staining colors as the means of the k=4 clusters. The k-means algorithm adapts to the variation in the images by shifting the cluster means and updating the pixel assignments. Figure 3 illustrates the results of this segmentation.

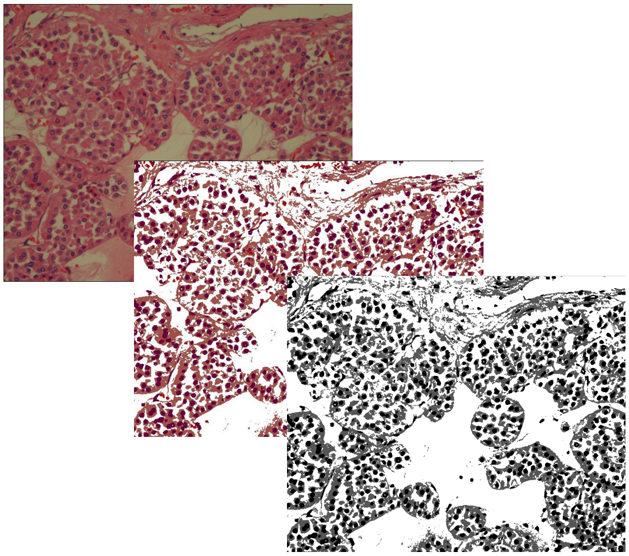

Fig. 3. K-Means segmentation 1) original image 2) color segmented pseudo image 3) gray level segmented image.

C. *Automated ROI Selection*

RCC tissue images generally focus on the viable malignant regions of the tissue. However, the malignant epithelial cells are sometimes integrated with necrotic or non-malignant regions (*e.g.,* cystic spaces, spaces between papillary structures, vascular structures, fibrosis, or benign kidney) in such a way that it is not possible to capture only viable carcinoma. Presence of necrotic or non-malignant areas significantly perturbs the image statistics, which subsequently results in classification errors. Several of these confounding factors are devoid of nuclei including cystic spaces, spaces between papillary structures, vascular or tubular lumina, necrosis or sectioning artifacts. Therefore, we design a set of morphological operations that focuses the analysis on regions of high nuclear density and removes the necrotic tissue regions in the images.

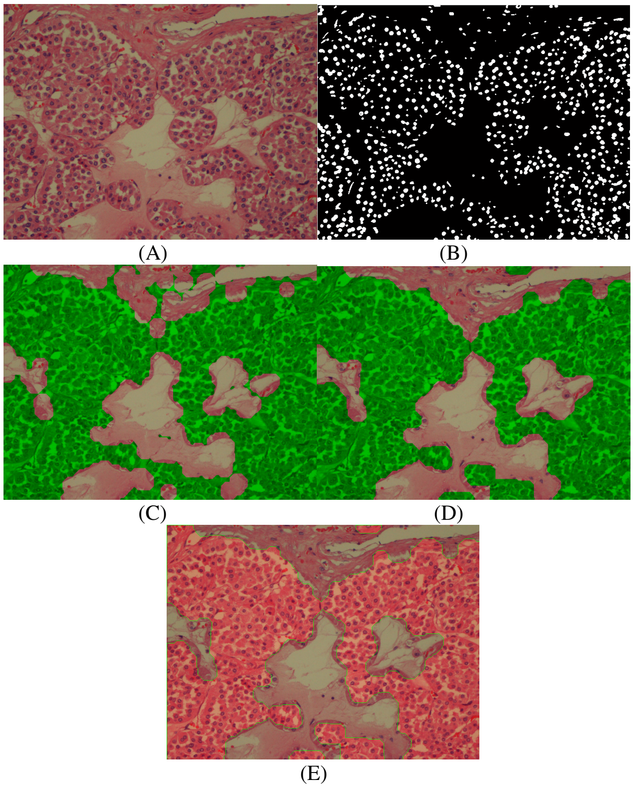

Fig. 4: Left → Right: Top→ Bottom: (A) Original image (B) nuclei mask. (C) Area mask after closing operation. (D) Mask after opening. (E) Final mask overlaid original image showing ROI segmentation.

The sequence of operations performed to extract the ROI is as follows:

a) Obtain nuclear mask N from gray level segmentation.

b) Perform image opening and closing with a one pixel structuring element to smooth out the nuclear mask by removing surface spikes.

$$N = N \bullet S \text{ (closing operation)} \qquad (2)$$

$$N = N \circ S \text{ (opening operation)} \qquad (3)$$

c) Estimate the average nuclei diameter, $\boldsymbol{R_{avg}}$, for the given image.

d) If X is the average inter-nuclear distance in the cancerous region of the image, then the final disk size to obtain relevant ROI is given by:

$$D(r) = R_{avg} \times X \tag{4}$$

e) Perform one step opening and closing to extract the ROI.

$$N = N \bullet D(r) \tag{5}$$

$$N = N \circ D(r) \tag{6}$$

Our ROI selection differs from a general pre-processing step since we are trying to select nuclei rich regions (corresponding to cancerous areas). The algorithm encapsulates this *a priori* intention by using the average nuclei diameter and inter-nuclear distance in ROI selection steps described above. The morphological processing is intentionally kept simple to keep the computational cost low which may not give precise region segmentation but it is still capable of rejecting large open spaces and necrotic regions in the tissue. Figure 4 shows some results obtained in the process of ROI segmentation.

D. Feature Extraction

After the ROI segmentation of high nuclear density areas, we use the four-level grayscale images for feature extraction. Our feature extraction algorithm is based on gray level co-occurrence matrix (GLCM), however, unlike [12], we compute GLCM for the selected ROIs only. The GLCM captures the frequency that a gray-level value occurs adjacent to another gray-level value [16]. This kind of information is not presented by histograms. As we already segmented the images into four gray level intensities, our GLCM is a 4x4 matrix. One GLCM matrix of size 4x4 represents one spatial relation (e.g horizontal) between the intensities of the image. Therefore we calculated four GLCM matrices to cover all the four spatial relations (horizontal, vertical, diagonal at the angle of 45 degree and diagonal at -45 degrees) between the intensities from the image and take the average of these four to present the overall spatial relation of the gray level intensities within the image. We use this average GLCM matrix to calculate the statistics such as contrast, correlation, energy, homogeneity and entropy. The *correlation* between neighboring pixels is given by

$$\sum_{i,j=0}^{N-1} P_{i,j} \frac{(i-\mu_i)(i-\mu_j)}{\sqrt{\sigma_i^2 \sigma_j^2}} \tag{7}$$

where μ is the mean and σ is the variance [16]. Energy, also known as uniformity or the angular second moment, provides the sum of squared elements in the GLCM and is given as

$$\sum_{i,j=0}^{N-1} P_{i,j}^2 \tag{8}$$

where $P_{i,j}$ is the probability of the i^{th} grayscale value occurring next to the j^{th} grayscale value in one of the four spatial relations. Entropy corresponding to the randomness between the elements of GLCM is given by

$$\sum_{i,j=0}^{N-1} P_{i,j}(-\ln P_{i,j}) \tag{9}$$

We represent four gray levels with numeric labels 1 to 4, corresponding to nuclei, cytoplasm, red blood cells and glands, respectively. We experimented with several different associations between morphological features using different GLCM elements such as GLCM(1,2), GLCM(2,3) etc. Among all the GLCM elements GLCM (2,4) provides the best classification results showing that the association between cytoplasmic staining and glands is an important feature for classification. In Table 1, the numeric values

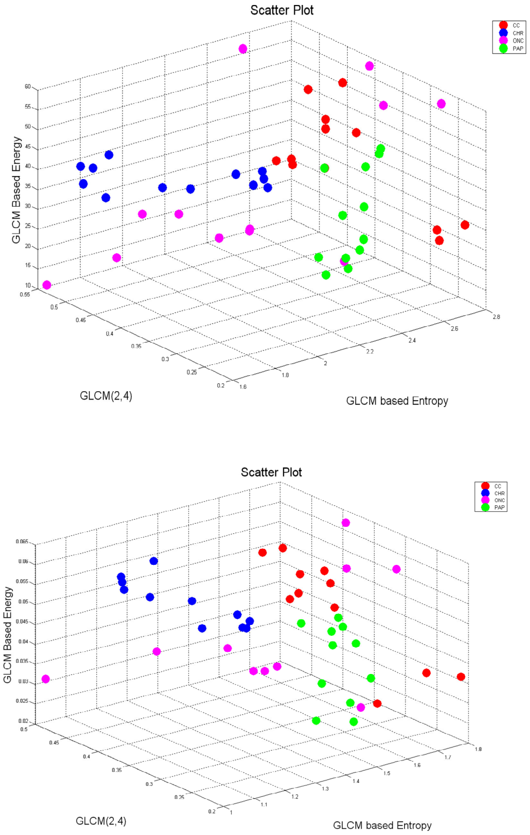

Fig. 5: Plot, showing distribution of subtypes using only three features (top) using the complete image, (bottom) using the masked region.

TABLE I

FEATURES SELECTED FOR CLASSIFICATION

	CC	CHR	ONC	PAP
GLCM(2,4)	49.5000 ±13.912	41.916 ±3.175	40.583 ±22.24	36.7500 ±11.794
Correlation	0.7491 ±0.0169	0.7405 ±0.046	0.7376 ±0.063	0.8376 ±0.033
Energy	0.2786 ±0.0382	0.4178 ±0.068	0.3599 ±0.092	0.2569 ±0.015
Entropy	1.5526 ±0.2032	1.9034 ±0.193	2.0968 ± .3211	2.3124 ±0.065

represent the mean value of the calculated statistic plus/minus the standard deviation.

Classification: We used the features (in Table I) extracted in the last step for classification into subtypes with a simple multi-class Bayesian classifier assuming Gaussian distribution for the extracted features. To estimate the accuracy of our system we used the leave-one-out cross validation resulting in about 95% (45/48) accuracy.

IV. RESULTS AND CONCLUSION

We have obtained about 95% classification accuracy for H&E stained RCC images with the features in Table I. Figure 5 is the scatter plot of these features showing the separation between the subtypes. We achieved 90% accuracy in our previous work [14] when the whole image was considered for classification and features from wavelet analysis were also used. By segmenting the desired ROIs, the classification accuracy improved to 95% with only four features. Thus, we have achieved faster classification with improved accuracy since the number of features is reduced and computation for wavelet analysis is omitted. Figure 6 provides an example CC image that was misclassified as ONC before ROI segmentation and correctly classified after ROI segmentation. Such accurate classification of difficult heterogeneous images was not achievable in our previous works [13, 14]. Thus, it can be concluded that using only relevant regions for classification results in better classification accuracy. This approach shows potential for clinical impact because it incorporates the pathologist's method of focusing on a region of interest.

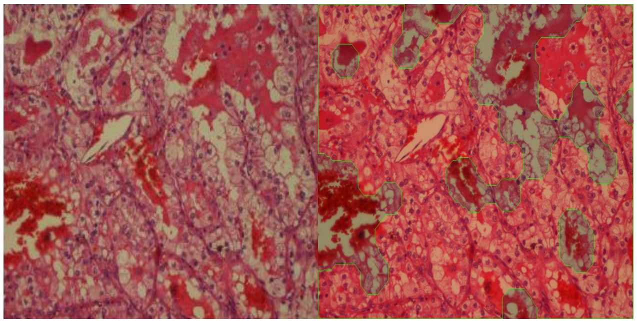

Fig. 6: (Left) CC image misclassified observing complete image. (Right) Segmented ROI classifies image correctly.

Figure 7 shows an example of misclassified and correctly classified Chromophobe images. The misclassified image has several glands which are masked out by our ROI selection. Thus, the cancerous area available for classification is much less as compared to the correctly classified images. Also, this small area resembles the selected region in Oncocytoma image resulting in misclassification of Chromophobe into Oncocytoma.

V. DISCUSSION

With our improved classification system, we are highly motivated to apply this system for classification of other cancer types. To further improve the accuracy of our algorithm, we are exploring techniques for better ROI segmentation. For example, we can remove other irrelevant features such as necrotic and interstitial regions. However, some textural information such as distinct light yellow longitudinal streaks in papillary RCC (representing the spaces between malignant papillary projections), as shown in Figure 8, are significant for feature extraction and subsequent classification. This requirement of preserving distinct subtype features while still omitting unnecessary regions for faster and accurate classification, rules out the possibility of using cancer region classification algorithms such as described in [5]. These algorithms will isolate the cancerous areas only and thus we will not have enough image information to generate reliable feature statistics. For example, we will lose the papillary streaks to the non-cancer region and will not be able to distinguish papillary subtype from other subtypes. This makes the design of ROI selection process a non-trivial task for RCC classification purposes. Given these challenges, we are further improving our system to perform better ROI segmentation as well as target other unwanted tissue structures, for instance strands of connective tissues that could also be eliminated for even more accurate classification results. We also plan to generate classification results using other classifiers such as support vector machines. The outcome of our algorithm depends on the heterogeneity of images and better results can be obtained if dataset excludes difficult heterogeneous images. However, our objective for this work is to achieve as much accuracy as possible with the heterogeneous datasets since heterogeneity cannot be overruled in biological images. It is quite clear that the high classification accuracy is obtained due to intelligent region selection prior to classification. We used leave one out cross validation method in our previous work [14], however the accuracy was not as high as in present work.

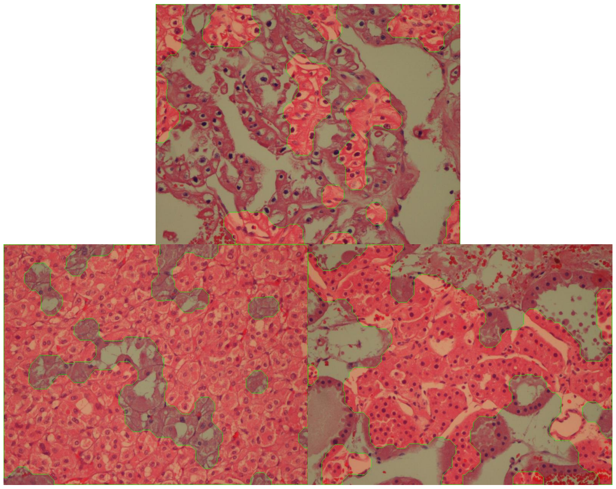

Fig. 7: (Top) misclassified CH image (Bottom left) correctly classified CH image (Bottom Right) Correctly classified Oncocytoma image.

VI. Future Trends

Computer aided diagnosis is gaining acceptance by the clinical community since it provides faster diagnosis with reduced subjectivity. Works such as in [17-20], have clearly demonstrated the need for digital diagnosis. In [17] and [18], diagnostic disagreement among pathologists for melanoma is studied. To decrease the diagnosis variability, the authors have suggested the use of panels of pathologists to study individual specimens. However, access to a panel of pathologists may not be feasible especially in developing and third world countries. Computer aided diagnosis with added pathologist opinion and sharing of digital results with other pathologists when needed will be an important component of future medicinal practices.

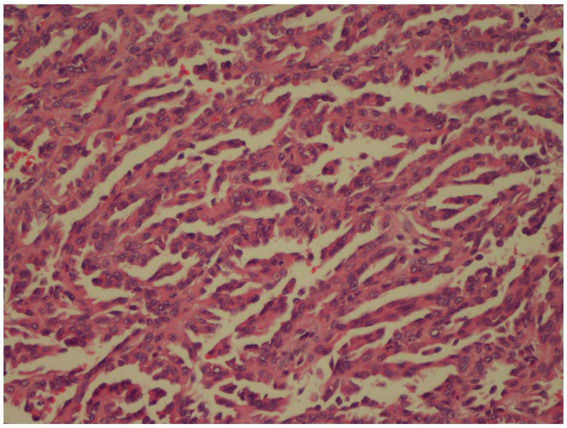

Fig. 8: Papillary image.

VII. References

[1] http://www.emedicinehealth.com/renal_cell_cancer/article_em.htm.

[2] "Cancer Facts & Figures," *American Cancer Society*, Inc., No. 500807, 2007.

[3] "WHO Classification of Tumours. Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs," Edited by J. N. Eble, G. Sauter, J. I. Epstein and I. A. Sesterhenn, *IARC Press*, Lyon France, February 2004, ISBN 92 8322 415 9.

[4] Y. Jiang, R.M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, K. Doi, "Improving breast cancer diagnosis with Computer-aided diagnosis," *Acad Radiol*, 6: 22-33, 1999.

[5] J. P. Thiran and B. Macq, "Morphological Feature Extraction for the classification of digital images of cancerous tissues," *IEEE Transactions on Biomedical Engineering*, 43 (1996), Page(s): 1011-1020.

[6] M. A. Roula, J. Diamond, A. Bouridane, P. Miller, and A. Amira, "A multispectral computer vision system for automatic grading of prostatic neoplasia," *IEEE International Symposium on Biomedical Imaging,* Washington D. C., 2002, pp. 193- 196.

[7] A. N. Esgiar, 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 Transactions on Information Technology in Biomedicine*, Volume 2, Issue 3, Sept. 1998 Page(s):197 - 203.

[8] A. N. Esgiar, R. N. G. Naguib, B. S. Sharif, M. K. Bennett, A. Murray, "Fractal analysis in the detection of colonic cancer images," *IEEE Transactions on Information Technology in Biomedicine*, Volume 6, Issue 1, March 2002, Page(s): 54-8.

[9] J. Diamond, N. Anderson, P. Bartels, R. Montironi, and P. Hamilton, "The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia," *Human Pathology*, vol. 35, pp. 1121-1131, 2004.

[10] S. Arivazhagan and L. Ganesan, "Texture classification using wavelet transform," *Pattern Recogn. Lett.* , vol. 24, no. 9-10, pp. 1513–1521, 2003.

[11] M. Fern´andez and A. Mavilio, "Texture analysis of medical images using the wavelet transform," *AIP Conference Proceedings*, vol. 630, October 2002, pp. 164–168.

[12] R.F. Walker, P.T. Jackway, B. Lovell, "Classification of cervical cell nuclei using morphological segmentation and textural feature extraction," *Proc., of the 2nd Australian and New Zealand Conference on Intelligent Information Systems*, 297-301, 1994.

[13] S. Waheed, R. A. Moffitt, Q. Chaudry, A. N. Young, M. D. Wang, "Computer Aided Histopathological Classification of Cancer," *Proc. IEEE Bioinformatics and Bioengineering*, 2007, Page(s):503 – 508.

[14] Q. Chaudry, S. H. Raza, A. N. Young, M. D. Wang, "Automated Renal Cell Carcinoma Subtype Classification Using Morphological, Textural and Wavelet-based Features," *Journal of Signal Processing: Special Issue on Biomedical Imaging*, 2008.

[15] A.R. Weeks and G.E. Hague, "Color Segmentation in the HSI Color Space Using the k-means Algorithm," *Proc. of the SPIE - Nonlinear Image Processing VIII*, 1997, pp. 143-154.

[16] M. Hall-Beyer, "The GLCM Tutorial Home Page," Version 2.10, Feb. 2007.[Online].Available: http://www.fp.ucalgary.ca/mhallbey/tutorial.htm [Accessed: Aug. 13, 2008]

[17] K. A. Fleming, "Evidence-based pathology," *Journal of Pathology* 1996, 179: 127-8.

[18] E. R. Farmer, R. Gonin, M. P. Hanna, "Discordance in the histopathologic diagnosis of melanoma and melanocytic nevi between expert pathologists," *Human Pathology*, 1996 27: 528-31.

[19] A. B. Ackerman, "Discordance among expert pathologists in diagnosis of melanocytic neoplasms," *Human Pathology*, 1996 27:1115-6.