Semantic Interpretation of Robust Imaging Features for Fuhrman Grading of Renal Carcinoma

Andrew Champion, Guolan Lu, Marcus Walker, Sonal Kothari, Adeboye O. Osunkoya, and May D. Wang*

Abstract— **Pattern recognition in tissue biopsy images can assist in clinical diagnosis and identify relevant image characteristics linked with various biological characteristics. Although previous work suggests several informative imaging features for pattern recognition, there exists a semantic gap between characteristics of these features and pathologists' interpretation of histopathological images. To address this challenge, we develop a clinical decision support system for automated Fuhrman grading of renal carcinoma biopsy images. We extract 1316 color, shape, texture and topology features and develop one vs. all models for four Fuhrman grades. Our models are highly accurate with 90.4% accuracy in a four-class prediction. Predictivity analysis suggests good generalization of the model development methodology through robustness to dataset sampling in cross-validation. We provide a semantic interpretation for the imaging features used in these models by linking features to pathologists' grading criteria. Our study identifies novel imaging features that are semantically linked to Fuhrman grading criteria.**

I. INTRODUCTION

Pathological analysis of tissues is an important step for cancer diagnosis and treatment. Traditionally, pathologists examine specimens under microscopes and make judgments based on deviations in cellular structures, change in the distribution of cells across the tissue, and clinical information about the patients being treated. However, this process is time-consuming, subjective and inconsistent due to inter- and intra-observer variations [1]. Therefore, computer-aided histological image classification systems are highly desirable to provide efficient, quantitative, and reliable information for cancer diagnosis and treatment planning. The computer-based detection and analysis of cancer tissue represents a challenging, yet unsolved task because of the large volume of patient data and their complexity [2]. The goal of predictive modeling is to construct models that make sound, reliable predictions and help physicians improve their prognosis, diagnosis or treatment planning procedures. However, there are many challenges facing computer-based decision making in clinical medicine. One important challenge is a semantic gap, the lack of biological interpretation of quantitative image features [3]. Because of the semantic gap, decision support systems act as a black box for pathologists and are more susceptible to errors. The purpose of our work is to develop an accurate computer-based decision support system for renal carcinoma grading and identify important image features contributing to high grading accuracy. We discuss relationships of these image features with underlying biological features. With this work, we hope to reduce the semantic gap between pathologists' knowledge and informative image features.

Andrew Champion is with the School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

Marcus Walker is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

Guolan Lu and Sonal Kothari are with the Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332.

Adeboye O. Osunkoya is with the Department of Pathology, Emory University School of Medicine, Atlanta, GA 30322 USA.

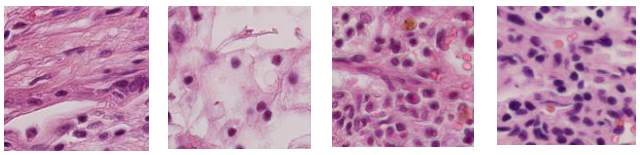
May D. Wang is with the Department of Biomedical Engineering, Winship Cancer Institute, Parker H. Petit Institute of Bioengineering and Biosciences, Institute of People and Technology, Georgia Institute of Technology and Emory University, Atlanta, GA (e-mail: maywang@bme.gatech.edu). .

II. BACKGROUND

The grading schema of renal cell carcinoma is based on the microscopic morphology of a hematoxylin and eosin (H&E) stained neoplasm. The most widely used schema is a nuclear grading system described in 1982 by Fuhrman et al. [4]. Pathological samples are classified by their disease stage, tumor size, cell arrangement, cell type, and nuclear grade. Four nuclear grades (1-4) are defined in order of increasing nuclear size, irregularity, and prominence (Fig. 1).

Nucleoli and other morphological features are important grade indicators, occurring in 94% of malignant cases [5]. Size and shape features capture these differences in nuclear and cellular structure and support discrimination between tissue cell subtypes [6]. Topological features have been shown to support accurate grade classification [5, 7]. Texture features contain information about the spatial distribution of gray tones related to tissue structure and markers in cytoplasm and nuclei. The gray-level co-occurrence matrix (GLCM) encodes properties of this distribution. The gray level run-length matrix (GLRL) captures texture features from contiguous, directional sequences of similar gray level intensities [8]. Local binary pattern (LBP) is a rotation-invariant feature useful for texture classification characterizing spatial structure and contrast [9].

In this study, we have extracted a combination of color, shape, texture, and topology features. Our goal is to identify robust, informative features for Fuhrman grading and link



Grade 1 Grade 2 Grade 3 Grade 4

Figure 1. Representative Histology Images for Each Fuhrman Grade

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

them to pathologists' interpretation. Similar previous work discussed biological interpretation of statistically significant feature subsets for one vs. one Fuhrman grading models [10]. Compared to previous work, we use a different dataset, a different set of imaging features, and different prediction model design. Using one vs. all Fuhrman grade models, we are able to isolate not only feature subsets but also individual features useful for distinguishing a particular Fuhrman grade from others. Though extracting fewer features overall, our approach extracts all combinations of texture, color and shape features with all segmented regions, while prior work considers only a targeted subset of feature type and region combinations. Consequently, we identify new features semantically linked to Fuhrman grading classification.

III. Methods

A. Data

Our dataset consists of 160 RGB images of H&E stained renal carcinoma tissue samples. These images are manually curated 2048x2048, 24-bit RGB sections of whole-slide images. Three carcinoma subtypes are represented in the dataset: clear cell (47%), papillary (33%), and chromophobe (20%). Expert pathological labeling of the dataset yielded 13% grade 1, 31% grade 2, 39% grade 3, and 17% grade 4 using the Fuhrman grading criteria.

B. Feature Extraction

Images are first segmented via the segmentation classifier described in [11] into nuclear, red blood cells, cytoplasm, and background regions. Red blood cell regions are discarded, and morphological cleaning is applied to the three remaining regions. To remove small regions from cytoplasm and background, which may be spurious or contribute to noise in extracted features, regions having a radius smaller than 2 pixels are removed via morphological opening with a disk. Connected components comprised of fewer than 20 pixels are then removed. This cleaning is not applied to nuclear regions, where small components may be individual nuclei. From each image and binary segmentation, we extract 1316 color, texture, shape, and topological features for each image.

Color features include 16-bin histograms and intensity distribution properties for the grayscale image and each color channel, color channel intensity differences, red ratio [12], and each of these features repeated for each region of interest, totaling 704 features. For these and all other features, distribution properties are extracted as eight summary statistics: mean, median, standard deviation, skewness, kurtosis, minimum, maximum, and inter-quartile range [10].

We extract 22 Haralick features from the GLCM, 44 GRLR features, and 18 LBP features for the entire image and each region of interest [8, 9]. The image is quantized into 8 discrete levels prior to GLCM and GLRL feature extraction. In total, we extract 339 texture features.

To quantify whole image shape features, the total area, number of distinct objects, and total perimeter are extracted for each region of interest. Each individual segmented object is also analyzed to reveal information typically related to the nuclear and cytoplasmic structures. Statistical distribution properties of each region of interest are extracted for the area, convex hull area, eccentricity, Euler number, axis lengths, orientation, and solidarity of each connected component, yielding 225 shape features .

Topological features are extracted based on the centroids of connected components segmented as nuclei. We measure distribution statistics of region area, edge length, and region perimeter of the Voronoi diagram and area and edge length of Delaunay triangulation, as well as the edge lengths of the minimal spanning tree of the triangulation graph [7]. This yields 48 topological features.

C. Model Development

Model selection and evaluation is performed through nested cross-validation consisting of 10 iterations of 3 folds in both inner and outer cross-validation. Dataset sampling in both the outer and inner folds is stratified by grade to preserve proportional representation of each grade. Binary one-versus-all classifiers are developed in each inner fold for each grade. Within each inner fold features are ranked for each grade, then a grid search is performed over all (hyper)parameters jointly. Winning parameters are selected according to the maximal mean accuracy of each unique parameter tuple within the inner cross-validation. This winning parameter tuple is used for model training in the corresponding outer cross-validation fold. Multiclass classification is aggregated from the selected binary classifiers by labeling each testing set pattern with the grade whose corresponding binary classifier has the maximal positive class probability (with random tie breaking) [13]. This process is repeated for each outer cross-validation fold independently.

Features are ranked by minimum redundancy maximum relevancy (mRMR) feature selection [14], using mutual information difference as the kernel. mRMR is a supervised, sequential feature ranking process whose object is, for each sequential feature, to select the remaining feature that maximizes the mutual information (joint probability density) with the target labeling while minimizing the mutual information between the new features and all features already selected. The number of features selected for model training is a parameter of grid search rather than chosen by heuristic during ranking. Prior to ranking, features are z-scored, then discretized into half-standard score width bins. Features are also z-scored, but not discretized, for model development.

Radial-basis kernel support vector machines (RBF SVM) are the classification algorithm for all models presented in this work [15]. The misclassification penalty hyperparameter, C, and radial basis function width, γ, are optimized jointly with a number of selected model features via grid search. Parameter search for γ is conducted in $\{2^{-16}, 2^{-14.5},\ldots, 2^{-1}\}$, C in $\{2^{0}, 2^{1.5},\ldots,2^{12}\}$, and number of selected features in $\{1, 2, \ldots, 10^{1.6}, 10^{1.8}, \ldots, 10^{3}\}$. In all, we consider 4653 parameter combinations and develop 17.2 million models.

IV. Results and Discussion

A. Grade Classification

Our model selection methodology resulted in an outer cross-validation accuracy of 90.43±4.43%, as shown in the confusion matrix in Fig. 2. Recall is best for grade 2 and worst for grades 1 and 4. For grades 1, 2, and 3, most

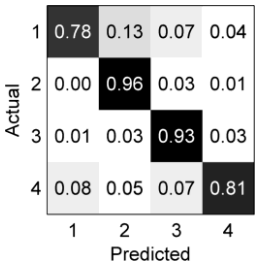


Figure 2. Confusion matrix of grade labeling on outer cross-validation

confusion occurs between neighboring grades, which is expected given the progressive, ordinal nature of the Fuhrman schema. However, for grade 4, most mislabeling (8%) occurs as grade 1, despite the marked dissimilarity of grade 1 and grade 4 tissue. Overall, most mislabeling occurs from a true lower grade to a predicted higher grade, which may be desirable from a clinical standpoint as the risk of underestimating the severity of illness is typically greater than that of overestimating it.

To evaluate the robustness of our model, Fig. 3 gives an analysis of the predictivity of selected model parameters, i.e., the correlation of model performance in internal and external validation. Although aggregate grade classification variance in outer validation is greater than in internal validation, 70% of outer folds have higher accuracy than the mean inner accuracy and 43% exceed the maximum inner accuracy. In a multiclass setting with high dimensional patterns, the 50% increase in training images in outer folds may increase model accuracy more than bias towards the selected inner CV parameters decreases it. This low bias and the low predictivity dispersion jointly indicate model parameters do not overfit inner cross-validation data.

The distribution of predictivity of all model parameters in Fig. 3 reflects the accuracy and relative difficulty of each grade classification task in Fig. 2. Selected binary models for all four grades have greater predictivity than nearly all other model parameters (grayscale density). For grade 2 many model parameterizations yield classifiers accurate in both internal and external validation, resulting from a large region of high accuracy in parameter space. In contrast, grade 4 has few parameterizations yielding accurate classifiers in either internal or external validation, with dense regions in Fig. 3 oriented perpendicular to the line perfect predictive correlation indicating high bias-variance tradeoff. Therefore, selecting models with both high internal accuracy and high predictivity is more difficulty for grade 4.

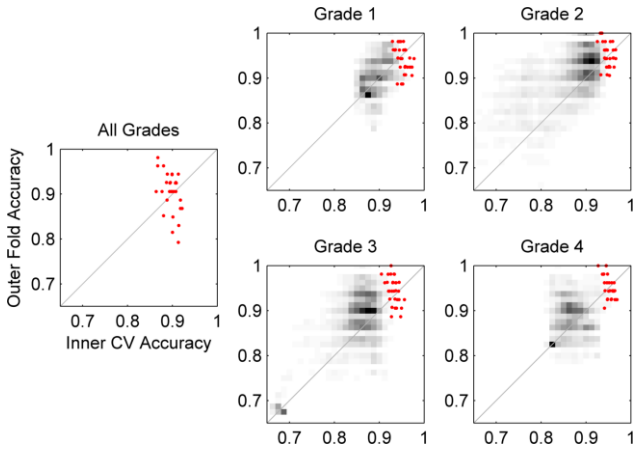


Figure 3. Predictivity analysis of selected models (red points) and all models in the parameter grid search (grayscale density)

B. *Informative Features and Semantic Interpretation*

The ranking frequency of the four feature categories aggregated across all inner cross-validation iterations is shown in Fig. 4, which summarizes the relative contribution of different feature types to each grading task. Frequency is normalized by the prevalence of each feature type so that relative performance can be assessed. First color features, then shape features are prevalent at the best ranks for all grades. Since the Fuhrman schema defines grade by nuclear morphology and density, the high ranking of shape and color features suggests that these features can capture cytoplasmic prevalence and nuclear density. Texture, which has previously been suggested to be more pertinent to tissue classification where cytoplasm properties are important indicators, is ranked lower on average than the other three categories. Following the first few ranks dominated by color and shape features, topology is highly ranked in grades 1 and 4. This suggests that topology features can capture sparse cell density in grade 1 and dense nucleation in grade 4.

The 5 most frequently selected and best ranked features for each grade in outer cross-validation, in Table 1, suggest features that are most consistent with the Fuhrman schema. In the Fuhrman schema, increasing nuclear size is the primary grade indicator. Table 1 suggests that minimum nuclear major axis length and median nuclear area, which contribute to accurate classification at both extremes of disease progression in grade 1 and grade 4, can capture nuclear size and elongation. The top ranked feature in grade 4, the lowest bin of the red channel histogram, may increase when nuclei are numerous and dense and little cytoplasm is present.

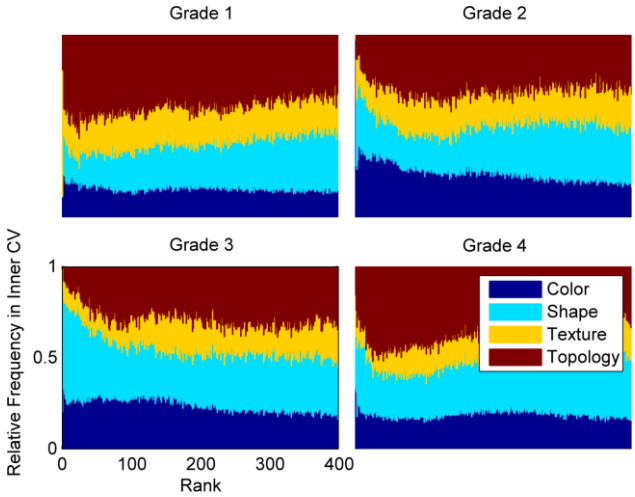


Figure 4. Ranking frequency of color, shape, texture, and topology features

TABLE I. MOST SELECTED FEATURES

	Feature	Type	Frequency
Grade 1	Background Red Histogram 240-255	Color	100%
	Nuclear Major Axis Length (Minimum)	Shape	100%
	Grayscale LBP (14/18)	Texture	100%
	Background Red Histogram 224-239	Color	97%
	Red Histogram 240-255	Color	93%
Grade 2	Red-Blue Channel Difference (Skewness)	Color	100%
	Red Ratio (Mean)	Color	100%
	Red-Green Channel Difference (Mean)	Color	97%
	Nuclear Red Ratio (Median)	Color	90%
	Cytoplasm Red Histogram 112-127	Color	87%
Grade 3	Blue Histogram 192-207	Color	100%
	Nuclear Eccentricity (Inter-Quartile Range)	Shape	100%
	Nuclear Eccentricity (Std. Dev.)	Shape	100%
	Background Eccentricity (Mean)	Shape	100%
	Cytoplasm Red Ratio (Kurtosis)	Color	100%
Grade 4	Red Histogram 0-15	Color	100%
	Nuclear Area (Median)	Shape	100%
	Grayscale Histogram 0-15	Color	100%
	Nuclear Convex Hull Area (Median)	Shape	100%
	Nuclear Major Axis Length (Minimum)	Shape	100%

Likewise, red-blue channel difference skewness is likely to associate with nuclear/cytoplasmic ratios, relating to increasing cell density in grade 2. Selected features also agree well with existing literature. For example, markedly increased nuclear/cytoplasmic ratios and nuclear eccentricity are useful in the separation of low-grade transitional cell carcinoma from benign urothelium [16]. Eccentric nuclei also indicate malignant rhabdoid tumors [17]. Feature type and region combinations not considered in previous work are also highly selected. Notably, regional color features appear for three grades in Table 1. Color properties of nuclear regions may indicate the presence nucleoli, which appear in higher Fuhrman grade samples. The red channel in cytoplasm regions can relate to cell density. In summary, Table 1 lists key image features affected by increasing nuclear size, irregularity, and prominence. Frequent selection of these feature subsets makes them an efficient choice for semantic interpretation of features that most impact the models. Categorical analyses like Fig. 4 can provide broad insight for the remaining features where specific interpretation is impractical.

V. CONCLUSION

The models developed by our methodology can classify renal carcinoma images into one of four Fuhrman grades with 90.4% accuracy. Predictivity analysis shows this methodology to be robust to sampling effects and selection set bias, indicating likely generalizability to other datasets in future work. Fuhrman grading criteria mainly depend on progressive growth and deformation of nuclei in carcinoma. Leveraging this fact, we summarize which parsimonious set of image features can best capture these nuclear characteristics. Some of the emerging features explicitly capture shape properties, while others only implicitly relate to pathologists' judgments, e.g., LPB features and color properties. We report only top 5 features but in fact many more may be used in some complex decision models. In future we will further analyze the model performance with small interpretable feature sets and suggest more granular image feature types that best capture pathologists' judgments. Using such imaging features will result not only in robust but interpretable decision support systems.

ACKNOWLEDGMENT

The authors would like to thank Dr. Todd H. Stokes for his assistance in image data acquisition.

REFERENCES

[1] S. M. Ismail, A. B. Colclough, J. S. Dinnen, D. Eakins, D. M. Evans, E. Gradwell, J. P. O'Sullivan, J. M. Summerell, and R. G. Newcombe, "Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia," *BMJ,* vol. 298, pp. 707-10, Mar 18 1989.

[2] A. Belle, M. A. Kon, and K. Najarian, "Biomedical Informatics for Computer-Aided Decision Support Systems: A Survey," *The Scientific World Journal,* vol. 2013, 2013.

[3] S. Kothari, J. H. Phan, T. H. Stokes, and M. D. Wang, "Pathology imaging informatics for quantitative analysis of whole-slide images," *J Am Med Inform Assoc,* vol. 20, pp. 1099-108, Nov-Dec 2013.

[4] S. A. Fuhrman, L. C. Lasky, and C. Limas, "Prognostic significance of morphologic parameters in renal cell carcinoma," *Am J Surg Pathol,* vol. 6, pp. 655-63, Oct 1982.

[5] M. Varma, M. W. Lee, P. Tamboli, R. J. Zarbo, R. E. Jimenez, P. G. Salles, and M. B. Amin, "Morphologic criteria for the diagnosis of prostatic adenocarcinoma in needle biopsy specimens. A study of 250 consecutive cases in a routine surgical pathology practice," *Arch Pathol Lab Med,* vol. 126, pp. 554-61, May 2002.

[6] S. Kothari, J. H. Phan, A. N. Young, and M. D. Wang, "Histological image classification using biologically interpretable shape-based features," *BMC Med Imaging,* vol. 13, p. 9, 2013.

[7] J. Sudbø, R. Marcelpoil, and A. Reith, "New algorithms based on the Voronoi Diagram applied in a pilot study on normal mucosa and carcinomas," *Analytical Cellular Pathology,* vol. 21, pp. 71-86, 2000.

[8] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics and Image Processing,* vol. 4, pp. 172-179, 1975.

[9] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, pp. 971-987, 2002.

[10] S. Kothari, J. H. Phan, A. N. Young, and M. D. Wang, "Histological Image Feature Mining Reveals Emergent Diagnostic Properties for Renal Cancer," *Proc IEEE Int Conf Bioinformatics Biomed,* pp. 422-425, 2011.

[11] S. Kothari, J. H. Phan, R. A. Moffitt, T. H. Stokes, S. E. Hassberger, Q. Chaudry, A. N. Young, and M. D. Wang, "Automatic batch-invariant color segmentation of histological cancer images," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, 2011, pp. 657-660.

[12] A. Tabesh, M. Teverovskiy, H. Y. Pang, V. P. Kumar, D. Verbel, A. Kotsianti, and O. Saidi, "Multifeature prostate cancer diagnosis and Gleason grading of histological images," *IEEE Trans Med Imaging,* vol. 26, pp. 1366-78, Oct 2007.

[13] R. Rifkin and A. Klautau, "In Defense of One-Vs-All Classification," *J Mach Learn Res,* vol. 5, pp. 101-141, 2004.

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans Pattern Anal Mach Intell,* vol. 27, pp. 1226-38, Aug 2005.

[15] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," ed, 2003.

[16] J. H. Hughes, S. S. Raab, and M. B. Cohen, "The cytologic diagnosis of low-grade transitional cell carcinoma," *Am J Clin Pathol,* vol. 114 Suppl, pp. S59-67, Nov 2000.

[17] M. B. Morgan, L. Stevens, J. Patterson, and M. Tannenbaum, "Cutaneous epithelioid malignant nerve sheath tumor with rhabdoid features: a histologic, immunohistochemical, and ultrastructural study of three cases," *J Cutan Pathol,* vol. 27, pp. 529-34, Nov 2000.