Searching Patterns in Glands for predicting Gastric Cancer Survival

Ricardo Moncayo[1], Sunny Alfonso[1], Angel Y. Sánchez[1], Carlos A. Parra[2], and Eduardo Romero[1]

[1]Computer Imaging and Medical Applications Laboratory CIM@LAB, School of Medicine, Universidad Nacional de Colombia, Bogotá, Colombia
[2]Department of Microbiology, Graduated School in Biomedical Sciences, Universidad Nacional de Colombia, Bogotá, Colombia

ABSTRACT

This article presents an entire framework for analyzing survival-related gland features in gastric cancer images. This approach builds upon a previous automatic gland detection, which partitions the tissue into a set of primitive objects (glands) from a binarized version of the hematoxylin channel. Next, gland shape and nuclei are characterized using local and contextual features that include relationships between color or texture from glands and nuclei (5.120 features). A mutual information max-relevance-min-redundancy (mRMR) approach selects hundred features that correlate with patient survival "survival vs not survival (first year)". Finally, ten statistically significant features (test t-student, $p < 0.05$) were used to set a "one-year" survival. Evaluation was carried out in a set of fourteen cases diagnosed with pre-cancerous gastric lesions or cancer, under a leave-one-out scheme. Results showed an accuracy of 78.57% when predicting the patient survival (less or more than a year), using a QDA Linear & Quadratic Discriminant Analysis. This approach suggests there exist morphometric gland differences among cases with gastric related pathology.

Keywords: Survival-related, local and contextual features, automatic gland detection, gastric cancer, mRMR, differences among cases.

1. INTRODUCTION

Gastric cancer (GC) incidence and mortality have been reduced over the past 70 years.[1] Despite a recent decline, worldwide it is still the fourth most common cancer and the seventh leading cause of cancer-related death,[2,3] Geographically, the highest GC incidence has been reported in Japon, Latin America, and the Caribbean.[4,5] In Colombia, GC is the first cause of cancer-related death, representing a 15% of all cancer deaths, with a high incidence in the Andean zone, especially in the departments of Nariño, Boyacá, and Cundinamarca. Currently, it is considered a major public health problem whose economic burden has reached the 47 million USD in the last five years.[6]

A GC diagnosis and stratification[7] is achieved by examining a biopsy tissue under a microscope.[8] This mainly relies upon certain level of expertise,[9] a limited resource in actual pathology laboratories. Overall, such diagnosis is not exempt of an inevitable observer bias and subjectivity. In this context, an automatic characterization of gastric glands may objectively support diagnosis and lead to devise more accurate indexes to predict the disease evolution.[10,11]

To the best of our knowledge, few investigations have aimed to determine survival in GC populations. Williams et al.[12] integrated multiple databases of patients diagnosed with GC including pathological, clinical, surgical and survival information. They applied a Machine Learning methodology to characterize subgroups of patients with gastric cancer by exploring all relationships between patient descriptors and systematically extracted over 450,000 logical associations. A subset of more than 1000 associations identified possible disease risk

Further author information: (Send correspondence to Eduardo Romero)
E-mail: edromero@unal.edu.co, Telephone: +57 3165000 ext. 15183

markers. Oh et al.[13] developed an automatic model to predict survival outcomes for patients with GC using a recurrent neural network (RNN). This study enrolled 1.243 cancer patients. Results showed a ROC AUC of 0.81 for the survival recurrent network (SRN) data test.

A main contribution of this work is an automatic characterization of gastric glands together with a set of features that might be associated with the disease aggressiveness. This set of characteristics correlates with the survival time ($\pm 1 year$) in a group of patients with gastric pathology. Furthermore, these discriminatory features are used in a classification task to build a model for predicting the patient survival time.

2. METHODOLOGY

A set of morphological and textural features are extracted from gastric glands automatically detected and their nuclei.[14] Using a max-relevance-min-redundancy (mRMR) criterion these features are reduced from about six thousand features to barely a hundred. These features are then statistically assessed to identify the ones that better express differences and these ones are then used to train Quadratic Discriminant Analysis (QDA) classifier.

2.1 Characterization of gastric glands

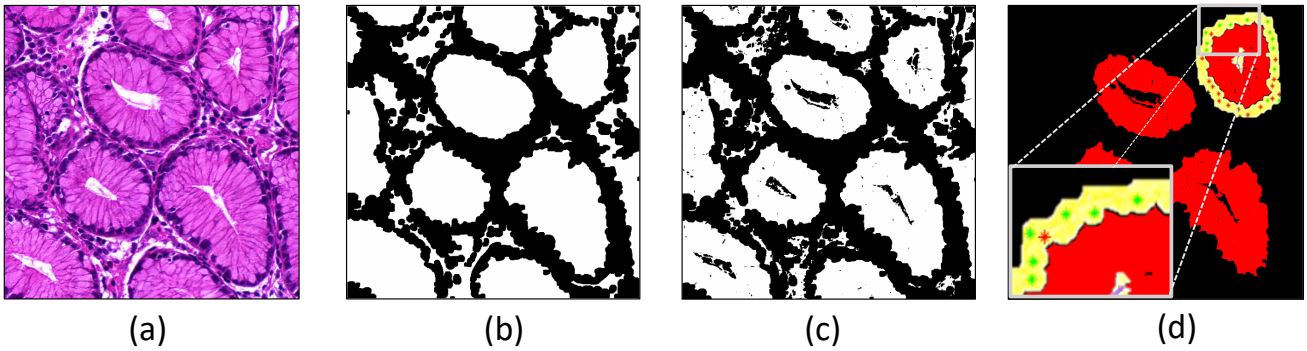

Figure 1. Proposed Methodology: a) Original Image, b) Gland binary mask, c) Gland candidates, d) Gland Nuclei

A coarse Gland binary mask is firstly constructed,[14] as illustrated in panel (b) of figure 1, by thresholding the hematoxylin channel, previously determined by a color deconvolution technique.[15] The original image in panel (a) is then thresholded and subtracted from a version of the image in panel (b) whose impulse noise has been filtered out by specific morphological operations, i.e., erosion and the area operator, which switches the binary value of all zones whose areas are smaller than a given value, see panel (c). Every gland intersecting the image border is excluded from this analysis.

A next step is the search of gland nuclei, a process starting by segmenting all nuclei and determining which of them belong to the gland. For doing so, gland candidates previously found are dilated by the maximum diameter of the largest nucleus (disk-structuring element of 60 pixels) and nuclei are evaluated by a Gradient-Boosted-Regression-Trees model to distinguish between gland-nuclei and non-gland-nuclei,[16] see panel (d). This classifier was trained using a set of 45.702 manually annotated gland nuclei, characterized by shape, texture and color features presented in table 1. This characterization also includes local and neighborhood analyses: while a local feature decomposes nuclei in terms of their geometric or physic characteristics, the neighborhood properties aim to capture nuclei in terms of their environment and population features. The neighborhood analysis is basically a spatial exploration of the region surrounding a nucleus and for doing so a set of circles with incremental radii of $k = dL \times 10$ pixels is placed at any nucleus center, starting with the average nuclei diameter ($dL = 20$ pixels) until $dL = 50$ pixels. Neighborhood features are computed from the nuclei inside the circles, in this case the

Features	Gland Nuclei			Gland
	Local	Neighborhood		
		Nuclei	Cytoplasm	
Shape	Area, perimeter, eccentricity, longness equiv-diameter, ratio between axes, Angle between axes, orientation, oval shape	Zernike moments, proximity to othe similar nuclei	Sum Inverse Distance of Nuclei, quantitative variance, areas and eccentricity variance, diameter variance, orientation variance, longness, ratio area	Area, relation between axes, perimeter, equiv-diameter, eccentricity, Zernike moments, mayor axes orientation.
Color	Max. intensity, Min. intensity, Mean intensity, Mean of red channel	Median red channel	Ratio Min. Intensity, Ratio RGB, Ratio Mean. Intensity, Ratio Red, Ratio Max. Intensity	Mean of red channel, Mean intensity, Mean intensity and variance of red channel, Max. intensity and Min. intensity, red entropy, EdgeMedIntensity
Texture	Entropy intensity, entropy of red channel	Entropy, Haralick features	Eosinophilic cytoplasm, Haralick features	Haralick features, entropy red channel, entropy intensity

Table 1. Features extracted from gland, gland nuclei and their cytoplasm.

nuclei and cytoplasm characteristics shown in table 1. A nucleus is represented by a vector with 52 features which correspond to such local and neighborhood descriptions.

A gland description is achieved by averaging these 52 features among the whole set of gland nuclei. The nuclei gland is then described by a vector with 104 characteristics composed of 52 feature averages, 52 standard deviations of these features and 24 gland characteristics.

So far exploration has been devoted to nuclei (local features) and neighborhood characteristics which are spatially extracted from a series of circles placed at the gland nuclei. Notice these features highly depend on the gland size and shape, two characteristics probably result of the biological sample treatment. These features in consequence are not absolute measures and therefore the relevant relative relationships were found out by a selection process. For doing so, an exhaustive computation of every relation between features was carried out, v.g. $\frac{\text{mean nuclei texture}}{\text{whole gland texture}}$, making the original 104 nuclei and 24 gland features are mapped to a new vector of $4,992$ relations, always respecting any relation is set between nuclei and gland characteristics. Finally, the gland feature vector corresponds to the concatenation of the original vector and the previously described relation feature vector, for a total of $5,120$ dimensions which is then pruned by the selection process.

A feature selection is performed in two steps: First, Minimum Redundancy Maximum Relevance (mRMR) approach is used to select relevant features[17] by minimizing the mutual information between features and maximizing the join probability of the selected features between classes "survival vs not survival (first year)". Afterwards, selected features correspond to those showing significant statistical differences (test t-student, $p < 0.05$) between the two survival groups.

Finally, the selected gland features are used to train a quadratic discriminant analysis (QDA) and obtain a model to predict patient-survival time.

3. EXPERIMENTATION AND RESULTS

3.1 Dataset Acquisition

This data-set was composed of 14 cases, description shown in table 2. Due to the high variability between individuals, applied inclusion and exclusion criteria are reported in table 3.
Gastric Cancer WSI were provided by Universidad Nacional de Colombia[18] and training glands were annotated

Data-set	Men	Women	Average Age	Min age	Max age
14	9	5	58	22	87

Table 2. Data-set Acquisition

Inclusion criteria
Patients older than 18 years of age.
Histopathological diagnosis of acute and chronic gastritis.
Intestinal metaplasia
Gastric cancer in-situ or advanced.
Without prior treatment and newly diagnosed.

Exclusion criteria
Women in gestation or lactation period
Underweight and/or malnutrition
In chemo-radiotherapy
Surgically operated in the last year
With autoimmune disease
Who have received blood transfusions in the last 6 months
Who have received treatment with antibiotics in the last 2 months
With recent infectious processes less than 2 months

Table 3. Inclusion and exclusion criteria

by one expert pathologist. Samples were obtained and digitized with a signed "informed consent" that followed the Helsinki protocol.[19] Table 4 shows the survival-time of 14 patients, 8 survived less than one year and the remaining 6 cases survived more than one year. The complete data set corresponds then to these 14 cases in which the survival-time was reported with their Kaplan-Meier curve in the figure 2 .

Cases	Diagnostic	Year of death	Years of survival
7	Adenocarcinoma	2014	0
1	Dysplasia	2014	0
2	Adenocarcinoma	2014	1
1	Gastritis	2015	1
1	Dysplasia	2016	2
1	Dysplasia	2018	4
1	Gastritis	2018	4

Table 4. All biopsies were taken in 2014, No survival (NS) less than 1 year=8, Survival (S)more than 1 year=6

3.2 Results

A total of 638 Fields of View (FoV) of 1024 × 1024 pixels at ×40 magnification were extracted from a set of H&E WSI, digitized from the 14 patients diagnosed with adenocarcinoma (n=9), gastritis (n=2), and dysplasia (n=3). From these FoV's 2.076 structures were found out and characterized by the model. The dimensionality reduction was achieved by a Minimum redundancy maximum relevance feature selection, finding the 100 most relevant features. Afterwards, statistical differences are computed to reduce the original set of 100 features to only the 10 most relevant characteristics, which are reported in table 5.

This last feature selection is used to train a QDA model and predict the survival time of a patient ($\pm 1 year survival time$). The model is trained using a leave-one-case out scheme validation, due to the small number of cases. That is to say, set aside one case for testing and use the remaining 13 GC cases for training, a task repeated 14 times. The final survival label is given by establishing the majority vote of the predictions for all the found glands of the test case. This model demonstrated an accuracy of 78.57% for the survival prediction task. Additionally, these 10 features are used in a multivariate regression COX model to determine the hazard

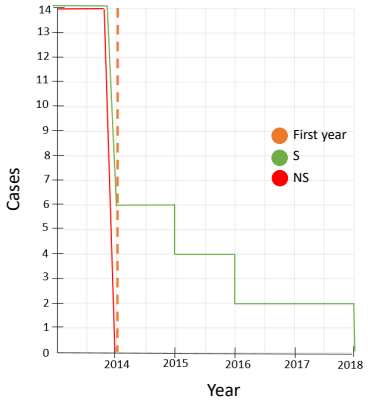

Figure 2. Survival time - Kaplan meier plot

ratio in each variable of both groups, thereby establishing a risk factor for each of the computed features. This risk ratio, or Hazard ratio, is a relative measure of how relevant a characteristic may be. For instance, for characteristics number 2, 3, 7, 8 and 9 in table 5, a value greater than one suggests that a change in the nuclei texture, cell proliferation per unit of area, the loss of cytoplasm- nuclei relation or nuclei hyperchromatism are linked with aggressiveness of the tumor. All these features have been widely described in most pathology manual as being important to describe aggressiveness.

4. CONCLUSIONS

This work has proposed a complete framework to determine a set of features suitable for predicting survival time in GC patients. Yet 14 GC cases are a small sample, this work suggests they are discriminant enough as to separate aggressive cases from those with a more benign biological pattern. Interestingly, most relevant features highlight relations between morphometry and nuclei texture and between nuclei and glandular texture. Future extension of this work includes the use of an extensive database as The Carcinoma Genome Atlas.[20]

Acknowledgments

This work was funded by Fundación CEIBA, Becaté Nariño and partially supported by COLCIENCIAS: contract 844-2017, code 110177758253, Phase I clinical study of immunotherapy with personalized synthetic vaccines in patients with triple-negative breast cancer and two grants from the Universidad Nacional de Colombia: (i) Biomarkers and vaccines for the management of breast cancer in Colombia, Hermes 42207 and (ii) Towards the implementation of different cancer immunotherapy strategies in Colombia, Hermes 41790.

REFERENCES

[1] Tekesin, K., Gunes, M. E., Tural, D., Akar, E., Zirtiloglu, A., Karaca, M., Selcukbiricik, F., Bayrak, S., and Ozet, A., "Clinicopathological characteristics, prognosis and survival outcome of gastric cancer in young patients: A large cohort retrospective study," *future* **1**, 3 (2019).

[2] Parkin, D. M., "Global cancer statistics in the year 2000," *The lancet oncology* **2**(9), 533–543 (2001).

[3] Parkin, D. M., "International variation," *Oncogene* **23**(38), 6329 (2004).

[4] Karimi, P., Islami, F., Anandasabapathy, S., Freedman, N. D., and Kamangar, F., "Gastric cancer: Descriptive epidemiology, risk factors, screening, and prevention," *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **23**, 700–713 (Mar. 2014).

[5] Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D., Forman, D., and Bray, F., "Cancer incidence and mortality worldwide: Iarc cancerbase no. 11." http://globocan.iarc.fr (2013). Accessed on 25/06/2018.

[6] Ministerio de Salud y Protección Social Instituto Nacional de Cancerología de Colombia, E., [*Plan Decenal para el Control de Cncer en Colombia, 2012-2021*] (2012).

[7] Gunduz-Demir, C., Kandemir, M., Tosun, A. B., and Sokmensuer, C., "Automatic segmentation of colon glands using object-graphs," *Medical image analysis* **14**(1), 1–12 (2010).

[8] Campagnola, P. J. and Loew, L. M., "Second-harmonic imaging microscopy for visualizing biomolecular arrays in cells, tissues and organisms," *Nature biotechnology* **21**(11), 1356 (2003).

[9] Gotink, A. W., ten Kate, F. J., Doukas, M., Wijnhoven, B. P., Bruno, M. J., Looijenga, L. H., Koch, A. D., and Biermann, K., "Do pathologists agree with each other on the histological assessment of pt1b oesophageal adenocarcinoma?," *United European gastroenterology journal* **7**(2), 261–269 (2019).

[10] Sun, G., Cheng, C., Li, X., Wang, T., Yang, J., and Li, D., "Metabolic tumor burden on postsurgical pet/ct predicts survival of patients with gastric cancer," *Cancer Imaging* **19**(1), 18 (2019).

[11] Ekundina, V. and Eze, G., "Common artifacts and remedies in histopathology (a review)," *African Journal of Cellular Pathology* , 1–7 (2015).

[12] Williams, C., Polom, K., Adamczyk, B., Afshar, M., D'Ignazio, A., Kamali-Moghaddam, M., Karlsson, N., Guergova-Kuras, M., Lisacek, F., Marrelli, D., et al., "Machine learning methodology applied to characterize subgroups of gastric cancer patients using an integrated large biomarker dataset," *European Journal of Surgical Oncology* **45**(2), e79 (2019).

[13] Oh, S., Choi, M., Seo, S., Sohn, T., Bae, J., and Kim, S., "Prediction of overall survival and novel classification of patients with gastric cancer using the survival recurrent network," *European Journal of Surgical Oncology* **45**(2), e79–e80 (2019).

[14] Alfonso, S., Corredor, G., Moncayo, R., Barrera, C. R., Sanchez, A. Y., Toro, P., and Romero, E., "A method to detect glands in histological gastric cancer images," in [*14th International Symposium on Medical Information Processing and Analysis*], **10975**, 109750X, International Society for Optics and Photonics (2018).

[15] Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, C., and Thomas, N. E., "A method for normalizing histology slides for quantitative analysis," in [*2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*], 1107–1110, IEEE (2009).

[16] Barrera, C., Corredor, G., Alfonso, S., Mosquera, A., and Romero, E., "An automatic segmentation of gland nuclei in gastric cancer based on local and contextual information," in [*Sipaim–Miccai Biomedical Workshop*], 75–81, Springer (2018).

[17] Peng, H., Long, F., and Ding, C., "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis & Machine Intelligence* (8), 1226–1238 (2005).

[18] Morales Álvarez, A. et al., *Inmuno-monitoreo del componente de células presentadoras de antígeno (APC) y células T en distintos estadios del desarrollo de cáncer gástrico de tipo intestinal*, PhD thesis, Universidad Nacional de Colombia-Sede Bogotá.

[19] Association, W. M. et al., "World medical association declaration of helsinki. ethical principles for medical research involving human subjects.," *Bulletin of the World Health Organization* **79**(4), 373 (2001).

[20] Tomczak, K., Czerwińska, P., and Wiznerowicz, M., "The cancer genome atlas (tcga): an immeasurable source of knowledge," *Contemporary oncology* **19**(1A), A68 (2015).

N	Feature relevant	Relation	Clinical interpretation	Hazard Ratio
1	Nuclei Intensity mean vs glands texture Variance	Color / Texture	Nuclei Hyperchromatism	0.832
2	Nuclei Inverse Difference Moment Std	Nuclei Texture	Nuclei inflammatory changes associated by tumor	1.134
3	SumAverages between nuclei vs SumAverages between glands	Texture / Texture	Lymphoid aggregates	1.047
4	Nuclei Orientation vs glands Area	Orientation / Shape	De-differentiation - Loss of normal structure	0.043
5	Nuclei EntropyIntensity Std	Nuclei Color	Nuclei Heterogeneity	0.038
6	Eosin, Inverse Difference Moment Mean	Cytoplasm Texture	Larger nuclei and glands	0.991
7	Nuclei Info.MeasuresCorrelation 2 mean vs glands Info.Measures Correlation 2	Texture / Texture	Cell proliferation per unit area	1.013
8	Eosin Info.MeasuresCorrelation 2 Mean	Cytoplasm Texture	Loss of cytoplasm- nuclei relation	1.085
9	Eosin, Entropy mean	Cytoplasm Texture	Nuclei Hyperchromatism	1.017
10	Eosin SumAverage mean	Cytoplasm Texture	Invasion of nuclei in the tumor	0.926

Table 5. Selected Features using mRMR and statistical test with Clinical interpretation