
| | |
|---|---|
| Project acronym | SIMBAD |
| Project full title | Beyond Features: Similarity-Based Pattern Analysis and Recognition |
| Deliverable Responsible | Department of Computer Science<br>ETH Zurich<br>Universitätstrasse 6<br>http://www.ml.inf.ethz.ch/ |
| Project web site | http://simbad-fp7.eu |
| EC project officer | Teresa De Martino |
| Document title | WP6: Analysis of tissue micro-array (TMA) images of renal cell carcinoma, final work package report |
| Deliverable n. | D6.2 |
| Document type | Final Report |
| Dissemination level | Public |
| Contractual date of delivery | M (month 35) |
| Project reference number | 213250 |
| Status & version | Definitive version |
| Work package | WP6 |
| Deliverable responsible (SHORT NAME) | ETHZ |
| Contributing Partners (SHORT NAME) | |
| Author(s) | Peter Schüffler, Sharon Wulff, Thomas Fuchs, Cheng Soon Ong, Volker Roth, Joachim M. Buhmann |
| Additional contributor(s) | |

# SIMBAD
# WP6: Analysis of tissue micro-array (TMA) images of renal cell carcinoma.
# D6.2: Final Work Package Report

Peter Schüffler, Sharon Wulff, Thomas Fuchs,
Cheng Soon Ong, Volker Roth, Joachim M. Buhmann
Department of Computer Science
ETH Zurich, Switzerland
peter.schueffler@inf.ethz.ch

September 18, 2011

## 1 Introduction

The clinical workflow of cancer tissue analysis is composed of several estimation and classification steps which yield a diagnosis of the disease stage and a therapy recommendation. This subproject proposes an automated system to model such a workflow which is able to provide more objective estimates of cancer cell detection and nuclei counts than pathologists achieved in this study. Our image processing pipeline is tailored to renal cell carcinoma (RCC), which is one of the ten most frequent malignancies in Western societies. The prognosis of renal cancer is poor since many patients suffer already from metastases at the time of first diagnosis. The identification of biomarkers for prediction of prognosis (prognostic marker) or response to therapy (predictive marker) is therefore of utmost importance to improve patient prognosis. Various prognostic markers have been suggested in the past, but conventional estimation of morphological parameters is still most useful for therapeutical decisions.

### 1.1 Goals of work package 6

There are two main goals of this work package in the SIMBAD project. First, we propose an automated pipeline to provide objective and reproducible diagnosis of renal cell carcinoma. This pipeline involves the following sub goals:

1. Nucleus detection, which comprises the identification of nuclei as well as the segmentation from the surrounding tissue.
2. Nucleus classification, which is based on the combination of various dissimilarity measurements between individual morphological structures across different features.
3. Survival analysis, which completes the TMA analysis pipeline.

This pipeline has been implemented as open source software and it is available on the SIMBAD website.

Second, our system provides a test bed for the novel methods and algorithms proposed in the SIMBAD project, in particular:

- WP4: Geometric embedding
- WP2: Deriving structural kernels

These methods are analyzed and discussed in more details in the various subsections.

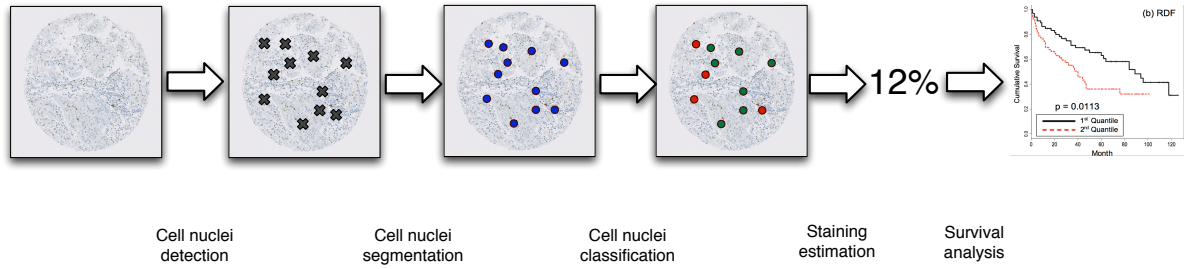



Figure 1: Automated pipeline for tissue microarray analysis. From left to right: (i) A TMA image is digitally stored for computational analysis. (ii) With object detection methods from computer vision, nuclei are identified in the image. (iii) The detected nuclei are cut out in image patches and segmentation algorithms discover the shape of the nuclei. (iv) The segmented nuclei are conducted to feature extraction (mainly histogram-like features) that are classified according to the clinical labels (here cancerous/benign). (v) Among the cancerous nuclei, the amount of stained nuclei is calculated. Staining discrimination is done via thresholding in color information. (vi) In a larger patient cohort, the pipeline (i.e. staining estimation) is validated regarding a survival analysis. Validation also comprises the comparison of pathologists' staining estimation vs. the prediction of the computational pathology algorithm.

## 2 Renal Cell Carcinoma Data

Clear cell RCC (ccRCC) is the most common subtype of renal cancer and it is composed of cells with clear cytoplasm and typical vessel architecture. ccRCC shows an architecturally diverse histological structure, with solid, alveolar and acinar patterns. The carcinomas typically contain a regular network of small thin-walled blood vessels, a diagnostically helpful characteristic of this tumor. Most ccRCC samples show areas with hemorrhage or necrosis (Fig. 2d), whereas an inflammatory response is infrequently observed. The cytoplasm is commonly filled with lipids and glycogen, which are dissolved in routine histological processing, creating a clear cytoplasm surrounded by a distinct cell membrane (Fig. 2d). Nuclei tend to be round and uniform with finely granular and evenly distributed chromatin. Depending upon the grade of malignancy, nucleoli may be inconspicuous and small, or large and prominent. Very large nuclei or bizarre nuclei may occur [DJSH04].

The tissue microarray (TMA) technology promises to significantly accelerate studies seeking for associations between molecular changes and clinical endpoints [KJ98]. In this technology, $0.6mm$ tissue cylinders are punched from primary tumor blocks of hundreds of different patients and these cylinders are subsequently embedded into a recipient tissue block. Sections from such array blocks can then be used for simultaneous in situ analysis of hundreds or thousands of primary tumors on DNA, RNA, and protein level (Fig. 2b,c). These results can then be integrated with expression profile data which is expected to enhance the diagnosis and prognosis of ccRCC [TM01], [MH99], [YA01]. The high speed of arraying, the lack of a significant damage to donor blocks, and the regular arrangement of arrayed specimens substantially facilitates automated analysis.

Although the production of tissue microarrays is an almost routine task for most laboratories, the evaluation of stained tissue microarray slides remains tedious, time consuming and prone to error. Furthermore, the significant intratumoral heterogeneity of RCC results in high inter-observer variability. The variable architecture of RCC also results in a difficult assessment of prognostic parameters. Current image analysis software requires extensive user interaction to properly identify cell populations, to select regions of interest for scoring, to optimize analysis parameters and to organize the resulting raw data. Because of these drawbacks in current software, pathologists typically collect tissue microarray data by manually assigning a composite staining score for each spot - often during multiple microscopy sessions over a period of days. Such manual scoring can result in serious inconsistencies between data collected during different microscopy sessions. Manual scoring also introduces a significant bottleneck that hinders the use of tissue microarrays in high-throughput analysis.

The prognosis for patients with RCC depends mainly on the pathological stage and the grade of the tumor at the time of surgery. Other prognostic parameters include proliferation rate of tumor cells

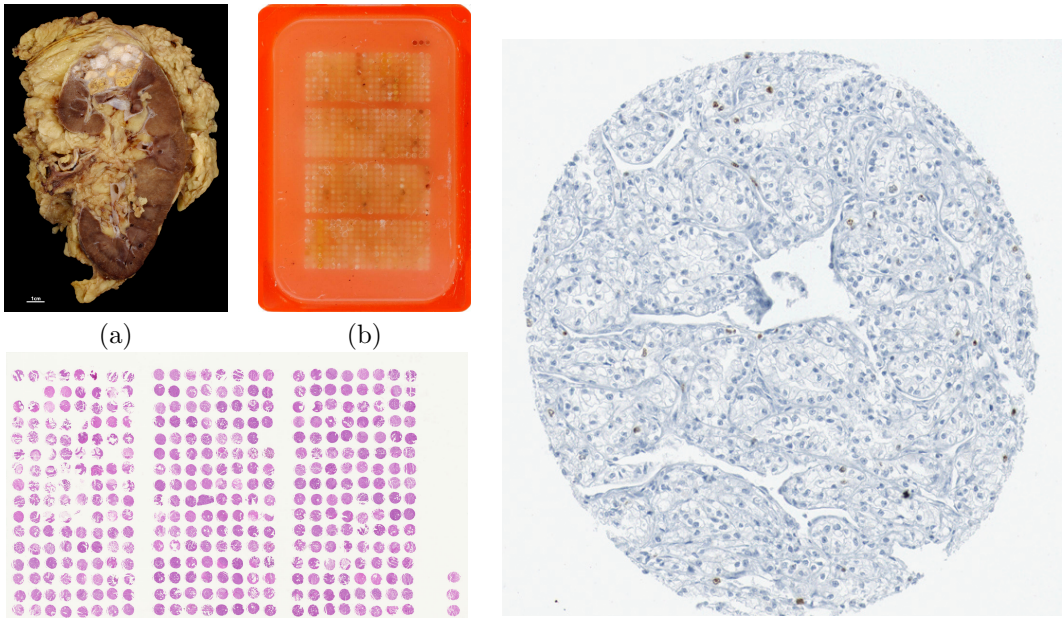



Figure 2: Tissue Microarray Analysis (TMA): Primary tissue samples are taken from a cancerous kidney (a). Then $0.6mm$ tissue cylinders are punched from the primary tumor block of different patients and arrayed in a recipient paraffin block (b). Slices of $0.6\mu m$ are cut off the paraffin block and are immunohistochemically stained (c). These slices are scanned and each spot, represents a different patient. Image (d) depicts a TMA spot of clear cell renal cell carcinoma from our test set stained with the MIB-1 (Ki-67) antigen.

and different gene expression patterns. Tannapfel et al. [TA96] have shown that cellular proliferation may prove to be another measure for predicting biological aggressiveness and, therefore, for estimating the prognosis. Immunohistochemical assessment of the MIB-1 (Ki-67) antigen indicates that MIB-1 immunostaining (Fig. 2d) is an additional prognostic parameter for patient outcome. TMAs are highly representative of proliferation index and histological grade using bladder cancer tissue [NA01].

In the domain of cytology, especially blood analysis and smears, automated analysis is already established [YMF05]. Histological tissues typically differs substantially from blood sample analysis with its homogeneous background on which the cells are clearly distinguishable and the absence of vessels and connection tissue. The isolation of cells simplifies the detection and segmentation process of the cells significantly. A similar simplification can be seen in the field of immunofluorescence imaging [MDK$^+$07]. Only the advent of high resolution scanning technologies in recent years rendered it possible to consider an automated analysis of histological slices. Cutting-edge scanners are now able to scan slices with resolution, comparable to a 40x lens on a light microscope. In addition the automated scanning of staples of slices enables an analysis in a high throughput manner.

## 2.1 Tissue Micro Arrays

Tissue micro arrays (TMA) are an important device for diagnosing and grading many tissue cancers, among these also renal clear cell carcinoma (RCC). The TMA glass plates carry small round tissue spots of prospective cancerous tissue samples with a thickness of one cell layer for each spot. After staining the spots with protein specific dyes, one can see under the microscope the cells' structure and morphology as well as the allocation and accumulation of the respective stained protein. For example, TMAs with RCC tissue might be stained with cell membrane and nucleus visualizing dyes, as well as with MIB1 specific staining. The first dye will enable to see the cells' structure under the light microscope, whereas the second staining will show the accumulation of protein MIB1 in the cell nuclei. MIB1 is a protein, which is highly accumulated in the nuclei of proliferating cells. As cancer cells are expected to have a

high proliferation rate in progressed cancer sites, positive MIB1 staining especially in tumor cells gives information about the stage of the tumor.

The TMA assessment will benefit in following ways from automatic processing: (i) the TMA estimation is reproducible and objective and (ii) grading can be done cheaper, faster and with a higher throughput. Pathologists have only to confirm the results and judge special cases, instead of manually go over each TMA image. The main steps of computer aided TMA estimation are: (1) Identification and detection of cell nuclei with in a high resolution TMA image, (2) segmentation of the nucleus, (3) estimating the nucleus' label, (4) calculating the percentage of tumor cells and protein expressing tumor cells. Big challenges for computational image processing tools are the nucleus detection and segmentation.

# 3 Nuclei Detection

From the raw TMA image, we detect the cell nuclei by learning an ensemble of binary decision trees, using manually annotated images.

## 3.1 Tree Induction

The base learners for the ensemble are binary decision trees, designed to take advantage of large feature spaces. With minor modifications, tree learning follows the original approach for random forests described in [Bre01]. A recursive formulation of the learning algorithm is given in procedure `LearnTree`. The sub procedure `SampleFeature` returns a feature consisting of two rectangles uniformly sampled within a predefined window.

In accordance with [Bre01] the Gini Index is used as splitting criterion, i.e. the Gini gain is maximized. At a given node, the set $S = s_1, \ldots, s_n$ holds the samples for feature $f_j$. For a binary response $Y$ and a feature $f_j$ the Gini Index of $S$ is defined as:

$$\widehat{G}(S) = 2\frac{N_{false}}{|S|}\left(1 - \frac{N_{false}}{|S|}\right), \qquad N_{false} = \sum_{s_i} I\left(f_j(s_i) = false\right), \tag{1}$$

where $|S|$ is the number of all samples at the current node and $N_{false}$ denotes the number of samples evaluated to $false$ by $f_j$. The Gini indices $\widehat{G}(S_L)$ and $\widehat{G}(S_R)$ for the left and right subset are defined similarly. The Gini gain resulting from splitting $S$ into $S_L$ and $S_R$ with feature $f_j$ is then defined as:

$$\widehat{\Delta G}(S_L, S_R) = \widehat{G}(S) - \left(\frac{|S_L|}{|S|}\widehat{G}(S_L) + \frac{|S_R|}{|S|}\widehat{G}(S_R)\right), \tag{2}$$

where $S = S_L \cup S_R$. From that follows, that the larger the Gini gain, the larger the impurity reduction. Recently [SaA07] showed that the use of Gini gain can lead to selection bias because categorical predictor variables with many categories are preferred over those with few categories. In the proposed framework this bias is not a problem due to the fact that the features are relations between sampled rectangles and therefore evaluate always to binary predictor variables.

## 3.2 Multiple Object Detection

For multiple object detection in a gray scale image every location on a grid with step size $\delta$ is considered as an independent sample $s$ which is classified by the ensemble. Therefore each tree casts a binary vote for $s$ being an object or background. The whole ensemble predicts the probability of being class 1: $RDF(s) = \sum_{i|t_i(s)=1} 1/\#\{i|t_i(s) = 1\}$, where $t_i$ denotes the $i$th tree. This procedure results in an accumulator or probability map for the whole image.

The final centroids of detected objects are retrieved by applying weighted mean shift clustering with a circular box kernel of radius $r$. During shifting, the coordinates are weighted by the probabilities of the accumulator map. While this estimate leads to good results in most cases, homogeneous ridges in the accumulator can yield multiple centers with a pairwise distance smaller than $r$. To this end we run binary mean shift on the detection from the first run until convergence. The radius is predefined by the average object size. If the objects vary largely in size the whole procedure can be employed for different scales. In accordance with [VJ01], not the image but the features respectively the rectangles are scaled.

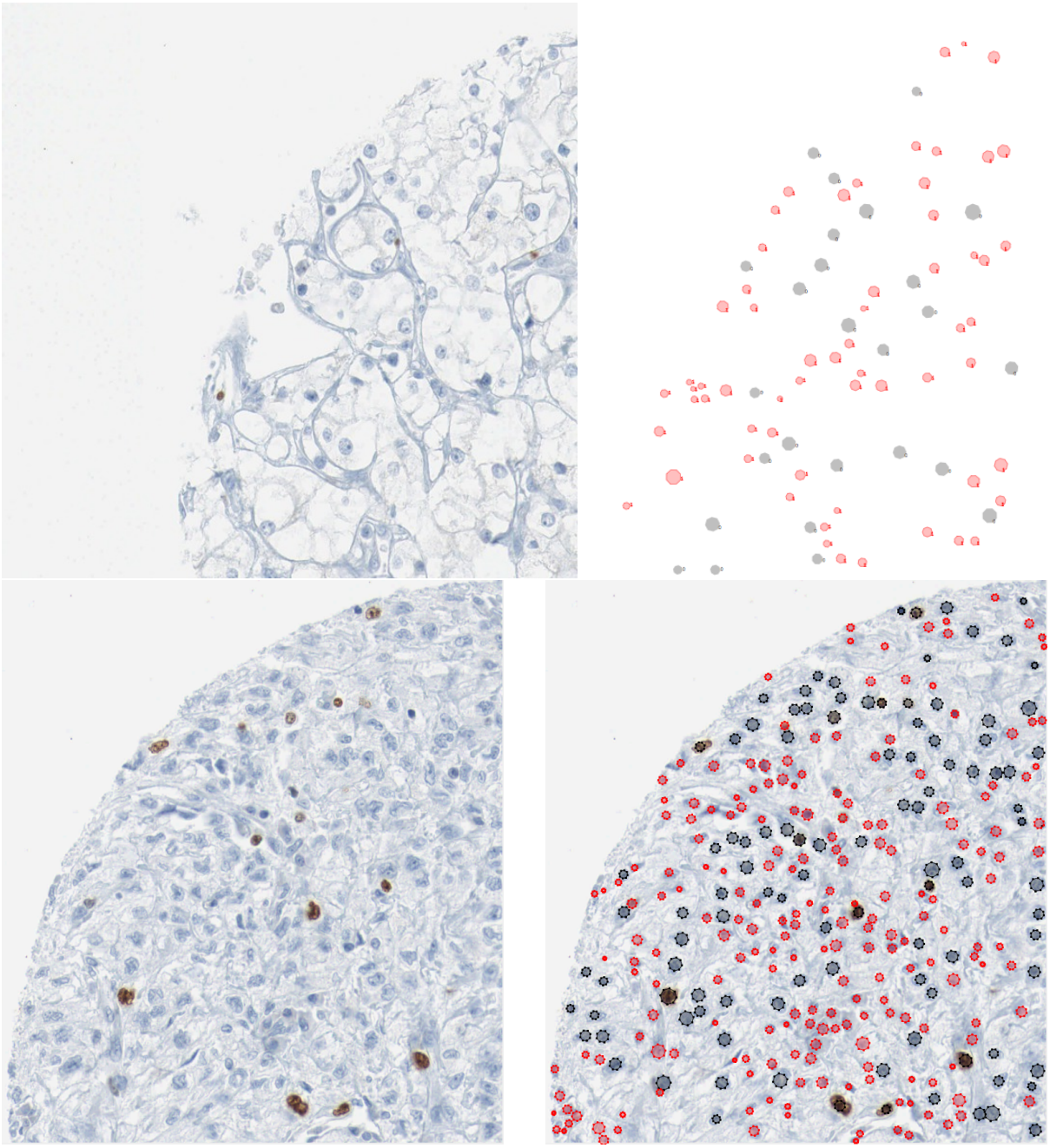


Figure 3: **Left:** Top left quadrant of a TMA spot from a ccRCC patient. **Right:** A trained pathologist labeled all cell nuclei and classified them into malignant (black) and benign (red). **Tissue Preparation and Scanning:** The TMA block was generated in a trial from the University Hospital Zurich. The TMA slides were immunohistochemically stained with the MIB-1 (Ki-67) antigen and scanned on a Nanozoomer C9600 virtual slide light microscope scanner from HAMAMATSU Photonics K.K.. The magnification of 40 times resulted in a per pixel resolution of $0.23\mu m$. Finally the spots of single patients were extracted as separate three channel color images of 3000 x 3000 pixels size.

**Procedure** LearnTree

**Input**: set of samples $S = \{s_1, s_2, \ldots, s_n\}$
**Input**: depth $d$; max depth $d_{max}$; features to sample $mTry$

**Init:** $\widehat{label} = null$; $g = -\inf$; $N_{left} = null$; $N_{right} = null$
**if** $(d = d_{max})$ *OR* $(\texttt{isPure}(S))$ **then**
  $\widehat{label} = \begin{cases} T & \textbf{if} \quad |\{s_j = T\}| > |\{s_j = F\}|; \quad j = 1, \ldots, |S| \\ F & \textbf{else} \end{cases}$
**else**
  **for** $i = 0, i < mTry, i++$ **do**
    $f_i = \texttt{SampleFeature()}$
    $S_L = \{s_j | f_i(s_j) = T\}; \quad S_R = \{s_j | f_i(s_j) = F\}; \quad j = 1, \ldots, |S|$
    $g_i = \widehat{\Delta G}(S_L, S_R)$
    **if** $g_i > g$ **then**
      $f^* = f_i; \quad g = g_i$
    **end**
  **end**
  $N_L = \texttt{LearnTree}(\{s_j | f^*(s_j) = T\})$
  $N_R = \texttt{LearnTree}(\{s_j | f^*(s_j) = F\})$
**end**

### 3.3 Performance Measure:

One way to evaluate the quality of the nuclei detection is to consider true positive ($TP$), false positive ($FP$) and false negative ($FN$) rates. The calculation of these quantities is based on a matching matrix where each Boolean entry indicates if a machine extracted nucleus matches a hand labeled one or not within the average nucleus radius. To quantify the number of correctly segmented nuclei, a strategy is required to uniquely match a machine detected nucleus to one identified by a pathologist. To this end we model this problem as a bipartite matching problem, where the bijection between extracted and gold-standard nuclei is sought inducing the smallest detection error [Kuh55]. This tuning prevents overestimating the detection accuracy of the algorithms. To compare the performance of the algorithms we calculated precision ($P = TP/(TP + FP)$) and recall ($R = TP/(TP + FN)$).

### 3.4 Implementation Details

The ensemble learning framework was implemented in C# and the statistical analysis was conducted in R [R D09]. Employing a multi threaded architecture tree ensembles are learned in real time on a standard dual core processor with 2.13 GHz. Inducing a tree for 1000 samples with a maximum depth of 10 and sampling 500 features at each split takes on average less than $500ms$. Classifying an image of $3000 \times 3000$ pixels on a grid with $\delta = 4$ takes approximately ten seconds using the non optimized C# code.

Three fold cross validation was employed to analyze the detection accuracy of RDFs. The nine completely labeled patients were randomly split up into three sets. For each fold the ensemble classifier was trained on data of six patients and tested on data of the other three. During tree induction, 500 features were sampled from the feature generator at each split. Trees were learned to a maximum depth of 10 and the minimum leave size was set to 1. The forest converges after 150 to an out of bag (OOB) error of approximately 2%. Finally, on the test images each pixel was classified and mean shift was run on a grid with $\delta = 5$. qFigure 4 shows precision/recall plot for single patients and the average result of the RDF object detector. The algorithm is compared to point estimates of several state of the art methods: SVM clustering was successfully employed to detect nuclei in H&E stained images of brain tissue by [GSC$^+$05]. SVMmorph [FLW$^+$08] is an unsupervised morphological approach [Soi03] for detection combined with an supervised support vector machine for filtering. The entry for the pathologists (red dot) shows the mean detection accuracy if alternately one expert is used as gold standard. On average the pathologists disagree on 15% of the nuclei.

Although only grayscale features were used for RDF it outperforms all previous approaches which also utilize texture and color. This observation can be a cue for further research that the shape information captured in this framework is crucial for good detection results.

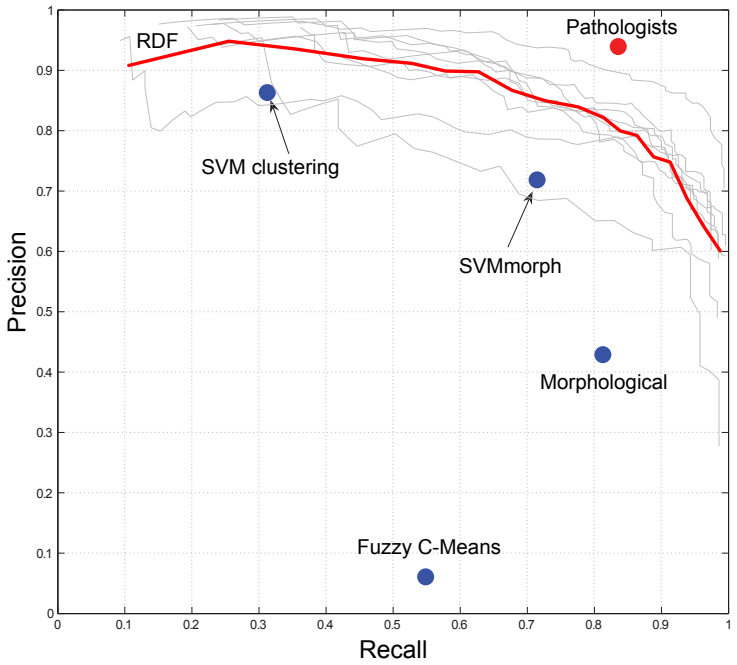



Figure 4: Precision/Recall plot of cross validation results on the renal clear cell cancer (RCC) dataset. Curves for the nine single patients and their average (bold) are depicted for Relational Detection Forests (RDF). RDF with the proposed feature base outperforms previous approaches based on SVM clustering [GSC+05], mathematical morphology and combined methods [FLW+08]. The inter-pathologists' performance is depicted in the top right corner.

### 3.5 Segmentation

Segmentation of the cell nuclei within the patches was generated with an adjusted graphcut method [Bag06, BK04, BVZ01, KZ04]. For that, patches were transformed to gray-scale and a low gray value was bound to the source node. The model was adjusted to prefer a roundish shape for a nucleus (the more a pixel is in the middle of the patch, the more it belongs to the source).

# 4 Nucleus Classification

Nucleus classification is an important issue in the computer-aided tissue micro array analysis. In short, this step comprises the decision that an image patch shows a single nucleus. The nucleus can be benign or cancerous which is to be classified by a subsequent algorithm. Of course, such a nucleus classification plays not only an important role in the automated TMA anlaysis of renal cell carcinoma, but also in a high variety of different cancers as well as in the entire clinical field of tissue pathology.

In the SIMBAD project, we investigated the performance of different nucleus classification approaches within our dataset of eight TMA image spots of human renal clear cell carcinomas. The cell nuclei in the images are bluish stained with hematoxylin. Nuclei that express the proliferation protein MIB-1 are further stained with a brown agent. Therefore, the cell nuclei to be classified can be blue or brown. TMA image analysis is difficult, also because (i) the dyes are inhomogenously dispersed in the image, (ii) the cell nuclei might be located very closely together and (iii) besides the nuclei, also other tissue fragments stain in the same color as the nuclei.

For these experiments, the cell nuclei of the eight TMA spots were identified and labeled by two pathologists, which enabled us to extract small image patches around the nuclei as samples. Each patch shows one nucleus in the middle. The patches are the bases for all classification experiments and serve as feature source for the experiments. In the following, we shortly outline four papers concerning the nucleus classification that were written in the context of SIMBAD.

## 4.1 Texture and shape features

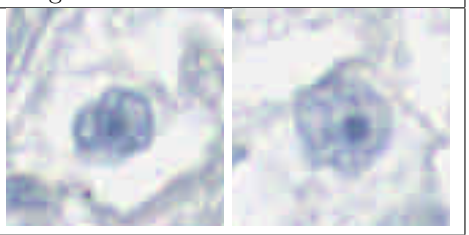
| | Non-Cancerous nucleus | Cancerous nucleus |
|---|---|---|
| Shape | Roundish | Irregular |
| Nucleus Membrane | Regular | Thick/thin irregular |
| Nucleus size | Smaller | Bigger |
| Nucleolus | None | Small dark spot in the nucleus |
| Nucleus texture | Smooth | Irregular |
| |  |  |

Table 1: Guidelines used by pathologists for identifying renal clear cell carcinoma nuclei.

Renal cell carcinoma revealed as one interesting aspect that the classification of cancerous cells can be achieved in a local fashion. There exist several guidelines followed by the pathologists (given in Table 1), which we use for the design of features employed in the machine learning approach. Testing of the classifiers was performed by leave-one-patient-out cross-validation. In this scenario, all patches from one TMA image were reserved for training in one cross-validation step and then classified by the resulting classifier. Eight classification errors (1-accuracy) for each cross-validation experiment (and eight patients) could then be visualized in a box.

## 4.2 Kernel based approach

Since the feature vectors consists of two different types, namely vectorial and histogram based features, we use the corresponding kernels or distance measures. Dissimilarity matrices were then centered to result in similarity matrices with zero mean. For kernel calculation, the histogram features were normalized to sum up to 1.

## 4.3 Multiple Kernel Learning for Cell Nucleus Classification [SUCM11, GUS+11]

We consider a Multiple Kernel Learning (MKL) framework for nuclei classification in tissue microarray images of renal cell carcinoma. Several features are extracted from the automatically segmented nuclei and we apply MKL and NLMKL (nonlinear MKL) for classification.

The main idea behind SVMs [Vap98] is to transform the input feature space to another space (possibly with a greater dimension) where the classes are linearly separable. After training, the discriminant function of SVM becomes $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \Phi(\boldsymbol{x}) \rangle + b$, where $\boldsymbol{w}$ is the vector of weights, $b$ is the threshold, and $\Phi(\cdot)$ is the mapping function. Using the dual formulation and the kernel trick, one does not have to define this mapping function explicitly and the discriminant function can be written as

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

where $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle$ is the kernel function that calculates a similarity metric between data instances. Selecting the kernel function is the most important issue in the training phase; it is generally handled by choosing the best-performing kernel function among a set of kernel functions on a separate validation set.

In recent years, MKL methods have been proposed [BLJ04, LCB+04], for learning a combination $k_\eta$ of multiple kernels instead of selecting only one:

$$k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\eta}) = f_\eta(\{k_m(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m)_{m=1}^P\}; \boldsymbol{\eta}) \tag{3}$$

where the combination function $f_\eta$ forms a single kernel from $P$ base kernels using the parameters $\boldsymbol{\eta}$. Different kernels correspond to different notions of similarity and instead of searching which works best, the MKL method performs the selection for us, or it may prefer a combination of kernels. MKL also allows us to combine different representations possibly coming from different sources or modalities.

**Linear Multiple Kernel Learning:** There is significant work on the theory and application of MKL and most of the proposed algorithms use a linear combination function such as convex sum or conic sum. Fixed rules use the combination function in (3) as a fixed function of the kernels, without any training. Once we calculate the combined kernel, we train a single kernel machine using this kernel. For example, we can obtain a valid kernel by taking the mean of the combined kernels.

Instead of using a fixed combination function, we can also have a function parameterized by a set of parameters and then we have a learning procedure to optimize these parameters as well. The simplest case is to parameterize the sum rule as a weighted sum:

$$k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\eta}) = \sum_{m=1}^{P} \eta_m k_m(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m)$$

with $\eta_m \in \mathbb{R}$. Different versions of this approach differ in the way they put restrictions on the kernel weights: [BLJ04, LCB+04, RBCG08]. For example, we can use arbitrary weights (i.e., linear combination), nonnegative kernel weights (i.e., conic combination), or weights on a simplex (i.e., convex combination).

**Nonlinear Multiple Kernel Learning:** A linear combination may be restrictive and nonlinear combinations are also possible [CMR10, GA08, LJN06]. Cortes [CMR10] developed a nonlinear kernel combination method based on kernel ridge regression (KRR) and polynomial combination of kernels. The nonlinear combination can be formulated as

$$k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{\boldsymbol{q} \in \mathcal{Q}} \eta_{q_1 q_2 \ldots q_P} k_1(\boldsymbol{x}_i^1, \boldsymbol{x}_j^1)^{q_1} k_2(\boldsymbol{x}_i^2, \boldsymbol{x}_j^2)^{q_2} \ldots k_P(\boldsymbol{x}_i^P, \boldsymbol{x}_j^P)^{q_P}$$

where $\mathcal{Q} = \{\boldsymbol{q} : \boldsymbol{q} \in \mathbb{Z}_+^P, \ \sum_{m=1}^P q_m \leq d\}$ and $\eta_{q_1 q_2 \ldots q_P} \geq 0$. The number of parameters to be learned is too large and the combined kernel is simplified in order to reduce the learning complexity:

$$k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{\boldsymbol{q} \in \mathcal{R}} \eta_1^{q_1} \eta_2^{q_2} \ldots \eta_P^{q_P} k_1(\boldsymbol{x}_i^1, \boldsymbol{x}_j^1)^{q_1} k_2(\boldsymbol{x}_i^2, \boldsymbol{x}_j^2)^{q_2} \ldots k_P(\boldsymbol{x}_i^P, \boldsymbol{x}_j^P)^{q_P}$$

where $\mathcal{R} = \{\boldsymbol{q} \colon \boldsymbol{q} \in \mathbb{Z}_+^P, \ \sum_{m=1}^P q_m = d\}$ and $\boldsymbol{\eta} \in \mathbb{R}^P$. For example, when $d = 2$, the combined kernel function becomes

$$k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{m=1}^{P} \sum_{h=1}^{P} \eta_m \eta_h k_m(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m) k_h(\boldsymbol{x}_i^h, \boldsymbol{x}_j^h). \tag{4}$$

The combination weights are optimized by solving the following min-max optimization problem:

$$\underset{\boldsymbol{\eta} \in \mathcal{M}}{\text{minimize}} \ \underset{\boldsymbol{\alpha} \in \mathbb{R}^N}{\text{maximize}} \ \boldsymbol{y}^\top \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^\top (\mathbf{K}_\eta + \lambda \mathbf{I}) \boldsymbol{\alpha}$$

where $\mathcal{M}$ is a positive, bounded, and convex set. Two possible choices for the set $\mathcal{M}$ are the $l_1$-norm and $l_2$-norm bounded sets defined as

$$\begin{aligned} \mathcal{M}_1 &= \{\boldsymbol{\eta} \colon \boldsymbol{\eta} \in \mathbb{R}_+^P, \ \ \|\boldsymbol{\eta} - \boldsymbol{\eta}_0\|_1 \leq \Lambda\} \\ \mathcal{M}_2 &= \{\boldsymbol{\eta} \colon \boldsymbol{\eta} \in \mathbb{R}_+^P, \ \ \|\boldsymbol{\eta} - \boldsymbol{\eta}_0\|_2 \leq \Lambda\} \end{aligned} \tag{5}$$

where $\boldsymbol{\eta}_0$ and $\Lambda$ are two model parameters. A projection-based gradient-descent algorithm can be utilized to solve this min-max optimization problem. At each iteration, $\boldsymbol{\alpha}$ is obtained by solving a KRR problem with the current kernel matrix and $\boldsymbol{\eta}$ is updated with the gradients calculated using $\boldsymbol{\alpha}$ while considering the bound constraints on $\boldsymbol{\eta}$ due to $\mathcal{M}_1$ or $\mathcal{M}_2$.

We formulate a variant of this method by replacing KRR with SVM as the base learner. In that case, the optimization problem becomes

$$\underset{\boldsymbol{\eta} \in \mathcal{M}}{\text{minimize}} \ \underset{\boldsymbol{\alpha} \in \mathcal{A}}{\text{maximize}} \ J_\eta = \mathbf{1}^\top \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^\top ((\boldsymbol{y}\boldsymbol{y}^\top) \odot \mathbf{K}_\eta) \boldsymbol{\alpha}$$

where $\mathcal{A}$ is defined as

$$\mathcal{A} = \{\boldsymbol{\alpha} \colon \boldsymbol{\alpha} \in \mathbb{R}_+^P, \ \ \boldsymbol{y}^\top \boldsymbol{\alpha} = 0, \ \ \boldsymbol{\alpha} \leq C\}.$$

We solve this optimization problem again using a projection-based gradient-descent algorithm. When updating the kernel parameters at each iteration, the gradients of $J_\eta$ with respect to $\boldsymbol{\eta}$ are used. These gradients can be written as

$$\frac{\partial J_\eta}{\partial \eta_m} = -\frac{1}{2} \sum_{h=1}^{P} \eta_h \boldsymbol{\alpha}^\top ((\boldsymbol{y}\boldsymbol{y}^\top) \odot \mathbf{K}_h \odot \mathbf{K}_m) \boldsymbol{\alpha}.$$

The data of 1273 nuclei samples are divided into ten folds (with stratification). We then train support vector machines (*svl, svp, svg*, see below) and MKL using these folds. We also combine the support vector machines using voting and report average accuracies using 10-fold CV. For the Gaussian kernel, $\sigma$ is chosen using a rule of thumb: $\sqrt{D}$ where $D$ is the number of features of the data representation. We compare our results using 10-fold CV $t$-test at $p = 0.05$.

As a summary, we have 9 representations (ALL, BG, COL, FCC, FG, LBP, PHOG, SIG and PROP), three different kernels (linear kernel: *svl*, polynomial kernel with degree 2: *svp*, and Gaussian kernel: *svg*), and five combination algorithms (RBMKL, SimpleMKL, GLMKL, NLMKL, and VOTE).

#### 4.3.1 Nuclei classification using Linear MKL [SUCM11]

Our first paper is a preliminary study on the classification of nuclei using the linear MKL formulation of Bach [BLJ04]. The best accuracy using a single SVM is 76.9 %. For most representations (except PHOG and COL), the accuracies of different kernels are comparable.

Next, we use the same kernel and combine all the feature sets which we extracted. As shown in Table 2, we can achieve an accuracy of 81.3 % using the linear kernel, by combining all representations. This experiment shows that the combination of information from multiple sources might be important and, by using MKL, the accuracy can be increased around 5 %. We observe from the table also that we have a decrease in accuracy compared to the single best support vector machine, when we use all kernels with *svp*. This phenomenon is analogous to combining all classifiers in classifier combination. If

one has relatively inaccurate classifiers, combining all may decrease accuracy. Instead, it might be better to choose a subset. This effect also shows that from a medical viewpoint, almost all the information is complementary and should be used to achieve better accuracy. In Figure 5, we plotted the weights of MKL when we use the linear kernel. As expected, the two best representations PHOG and PROP have high weights. But the representation LBP that has very low accuracy when considered as a single classifier increases the accuracy when considered in combination. This shows that when considering combinations, even a representation which is not very accurate alone may contribute to the combination accuracy. From this, we also deduce that these three features are useful in discriminating between healthy and cancerous cells and we may focus our attention on these properties.

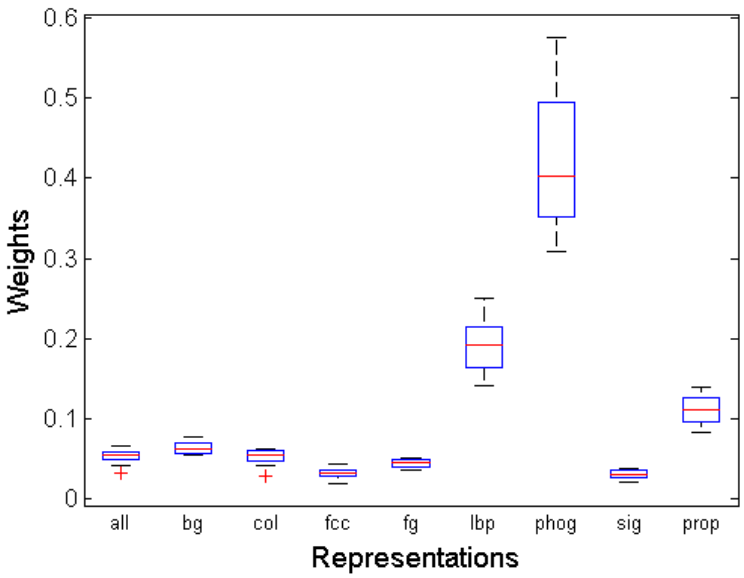



Figure 5: Combination weights in MKL using the linear kernel.

Table 2: MKL accuracies (in %). accuracy (± std) of combining all kernels.

| | *svl* | *svp* | *svg* |
|---|---|---|---|
| SINGLE-BEST | 76.5±3.7 (PHOG) | 75.6±2.6 (PROP) | 76.9±3.6 (PHOG) |
| MKL | **81.3±3.6** | 72.0±3.3 | 76.9±3.6 |
| VOTE | 70.0±0.2 | 71.3±1.7 | 72.4±1.2 |

#### 4.3.2 Nuclei classification using Nonlinear MKL [GUS+11]

Since our preliminary results on linear MKL's was a success, we formulated a nonlinear MKL variant derived from Cortes's formulation [CMR10]. In this paper, we compare our results with the single kernel svms and the four MKL variants. The results differ slightly with Section 4.3.1 because of small changes in segmentation.

Using four different MKL algorithms, we combined eight kernels calculated on the feature representations with the same kernel function. Table 3 lists the results of best single-kernel SVMs and four MKL algorithms trained. We can achieve an accuracy of 83.3 % by combining eight `GAU` kernels with `NLMKL`. This result is better than all other MKL settings and single-kernel SVMs. In the last column of Table 3, the results of combining all possible feature representation and kernel function pairs (i.e., 24 kernels)

in a single learner are shown. NLMKL is still the best MKL algorithm even though the average accuracy decreases to 83.1 %.

Table 3: MKL accuracies.

| | *svl* | *svp* | *svg* | *svl+svp+svg* |
|---|---|---|---|---|
| SVM | 76.0±3.4 | 72.7±3.8 | 76.9±2.7 | NA |
| RBMKL | 77.3±4.0 | 77.2±2.4 | 82.7±3.6 | 81.8±3.8 |
| SimpleMKL | 77.1±3.3 | 77.3±2.3 | 81.8±3.8 | 81.6±3.9 |
| GLMKL | 77.1±3.5 | 76.5±3.2 | 81.8±4.3 | 81.8±3.8 |
| NLMKL | 77.9±3.9 | 79.2±3.8 | **83.3±3.6** | 83.1±3.5 |

## 4.4 Hybrid Generative-Discriminative Nucleus Classification of Renal Cell Carcinoma [USB$^+$11, BUS$^+$11]

We propose a hybrid generative/discriminative classification scheme and apply it to the detection of renal cell carcinoma (RCC) on tissue microarray (TMA) images. In particular we use probabilistic latent semantic analysis (pLSA) as a generative model to perform generative embedding onto the free energy score space (FESS). Subsequently, we use information theoretic kernels on these embeddings to build a kernel based classifier on the FESS. We compare our results with support vector machines based on standard linear kernels and with the nearest neighbor (NN) classifier based on the Mahalanobis distance. We conclude that the proposed hybrid approach achieves higher accuracy, revealing itself as a promising approach for this class of problems. We demonstrate that the feature space created using pLSA achieves better accuracies than the original feature space and we get better accuracies when we apply IT kernels.

### 4.4.1 Generative model training

The generative model adopted is based on pLSA [Hof01], which was introduced in the text understanding community for unsupervised topic discovery in a corpus of documents, and subsequently largely applied by the computer vision community [BZM06], as well as in bioinformatics [BLOP10].

The basic idea underlying pLSA – and in general the class of the so-called topic models (of which another well-known example is the *latent Dirichlet allocation* model [BNJ03]) – is that each document is characterized by the presence of one or more topics (e.g. sport, finance, politics), which may induce the presence of some particular words. From a generative probabilistic point of view, pLSA generates a set of co-occurrences of the form $(d, w)$, where each of these pairs specifies the presence of a given word $w$ in a document $d$ (as in *bag-of-words* descriptions of documents). The generative model underlying these co-occurrence pairs is as follows: (i) obtain a sample $z$ from the distribution over the topics $P(z)$; (ii) given this specific topic sample, obtain a word sample from the conditional distribution of words given topics $P(w|z)$; (iii) given this specific topic sample, obtain a document sample (independently from the word sample) from the conditional distribution of documents given topics $P(d|z)$. The resulting distribution is

$$P(d, z) = \sum_z P(z)P(d|z)P(w|z), \tag{6}$$

where the sum ranges over the set of topics in the model. The parameters of this generative model may be obtained from a dataset using an expectation-maximization (EM) algorithm; for more details, the reader is referred to [Hof01].

In our approach, we simply assume that the visual features previously described are the words in the pLSA model, while the nuclei are the documents. The pLSA model learned from this data can be seen as defining *visual topics*. The representation of *documents* and *words* with topic models has one clear advantage: each topic is individually interpretable, providing a probability distribution over words that picks out a coherent cluster of correlated terms. This may be advantageous in the cancer detection context, since the final goal is to provide knowledge about complex systems, and to detect possible hidden correlations.

### 4.4.2 Generative Embedding

In this step, all the objects involved in the problem (namely training and testing patterns) are projected, through the learned model, onto a vector space. Different approaches have been proposed in the past, each one with different characteristics, in terms of interpretability, efficacy, efficiency, and others. Here we employ two schemes: the posterior distribution $P(z|d)$ – which was the first generative embedding based on pLSA models that was considered – and the *free energy score space* (FESS) – a novel method whose efficacy has been shown in different contexts [PCC$^+$09b, PCC$^+$09a].

In the posterior distribution embedding, a given nucleus (or document) $d$ is represented by the vector of posterior topic probabilities, obtained via the function $\psi$ defined as

$$\psi(d) = (P(z = 1|d), \ldots, P(z = T|d)) \in \mathbb{R}^T, \tag{7}$$

where we are assuming that the set of topics is indexed from 1 to $T$ (the total number of topics). Intuitively, the co-occurrences of visual features differ between healthy and cancerous cells and these co-occurrences are captured by the topic distribution $P(z|d)$, which should thus contain meaningful information for discrimination. This representation with the topic posteriors has been already successfully used in computer vision tasks [BZM06] as well as in medical informatics [BLOP10].

The FESS embedding [PCC$^+$09b, PCC$^+$09a] has been shown to outperform other generative embeddings (including those in [JH98] and [TKR$^+$02]) in several applications. This embedding expresses how well each data point fits different parts of the generative model, using the variational free energy as a lower bound on the negative log-likelihood. FESS embedding has been shown to yield highly informative and discriminative representations that lead to state-of-the-art results in several computational biology and computer vision problems (namely, scene/object recognition) [PCC$^+$09b, PCC$^+$09a]. Due to lack of space, the details of the FESS embedding are not reported here – please refer to [PCC$^+$09b, PCC$^+$09a] for a detailed presentation. The only important fact that needs to be pointed out here is that (as the posterior distribution embedding), the components of the FESS embedding of any object are all non-negative.

### 4.4.3 Discriminative Classification

In a typical hybrid generative-discriminative classification scenario, the feature vectors resulting from the generative embedding are used to feed some kernel-based classifier, namely, a *support vector machine* (SVM) with simple linear or radial basis function (RBF) kernels. Here, we take a different approach. Instead of relying on standard kernels, we investigate the use of the recently introduced information theoretic (IT) kernels [MSX$^+$09] as a similarity measure between objects in the generative embedding space. The main idea is that, with such kernels, we can exploit the probabilistic nature of the generative embeddings, improving even more the classification results of the hybrid approaches – this has been already shown in other classification contexts [BPM$^+$10, MBM$^+$10].

More in details, given two probability measures $p_1$ and $p_2$, representing two objects, several information theoretic kernels (ITKs) can be defined [MSX$^+$09]. The Jensen-Shannon kernel (will be referred to as `JS`) is defined as

$$k^{\,\mathrm{JS}}(p_1, p_2) = \ln(2) - JS(p_1, p_2), \tag{8}$$

with $JS(p_1, p_2)$ being the Jensen-Shannon divergence

$$JS(p_1, p_2) = H\left(\frac{p_1 + p_2}{2}\right) - \frac{H(p_1) + H(p_2)}{2}, \tag{9}$$

where $H(p)$ is the usual Shannon entropy.

The Jensen-Tsallis (JT) kernel (will be referred to as `JT`) is given by

$$k_q^{\,\mathrm{JT}}(p_1, p_2) = \ln_q(2) - T_q(p_1, p_2), \tag{10}$$

where $\ln_q(x) = (x^{1-q} - 1)/(1 - q)$ is a function called the $q$-logarithm,

$$T_q(p_1, p_2) = S_q\left(\frac{p_1 + p_2}{2}\right) - \frac{S_q(p_1) + S_q(p_2)}{2^q} \tag{11}$$

is the Jensen-Tsallis $q$-difference, and $S_q(r)$ is the Jensen-Tsallis entropy, defined, for a multinomial $r = (r_1, ..., r_L)$, with $r_i \geq 0$ and $\sum_i r_i = 1$, as

$$S_q(r_1, ..., r_L) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{L} r_i^q\right). \tag{12}$$

In [MSX+09], versions of these kernels applicable to unnormalized measures were also defined as follows. Let $\mu_1 = \omega_1 p_1$ and $\mu_2 = \omega_2 p_2$ be two unnormalized measures, where $p_1$ and $p_2$ are the normalized counterparts (probability measures), and $\omega_1$ and $\omega_2$ arbitrary positive real numbers (weights). The weighted versions of the JT kernels are defined as follows:

- The weighted JT kernel (version A, will be referred to as `JT-W1`) is given by

$$k_q^A(\mu_1, \mu_2) = S_q(\pi) - T_q^\pi(p_1, p_2), \tag{13}$$

  where $\pi = (\pi_1, \pi_2) = \left(\frac{\omega_1}{\omega_1+\omega_2}, \frac{\omega_2}{\omega_1+\omega_2}\right)$ and

$$T_q^\pi(p_1, p_2) = S_q\left(\pi_1 p_1 + \pi_2 p_2\right) - \left(\pi_1^q S_q(p_1) + \pi_2^q S_q(p_2)\right).$$

- The weighted JT kernel (version B, will be referred to as `JT-W2`) is defined as

$$k_q^B(\mu_1, \mu_2) = \left(S_q(\pi) - T_q^\pi(p_1, p_2)\right)(\omega_1 + \omega_2)^q. \tag{14}$$

The approach herein proposed consists in defining a kernel between two observed objects $x$ and $x'$ as the composition of the generative embedding function $\psi$ (the posterior embedding or the FESS embedding) with one of the JT kernels presented above. Formally,

$$k(x, x') = k_q^i\big(\psi(x), \psi(x')\big), \tag{15}$$

where $i \in \{\text{JT, A, B}\}$ indexes one of the Jensen-Tsallis kernels (10), (13), or (14), and $\psi(x)$ is the generative embedding of object $x$. Notice that this kernel is well defined because all the components of $\psi$ are non-negative, as is clear from (7) for the posterior probability embedding and was mentioned above for the FESS embedding. Once the kernel is defined, SVM learning can been applied. Recall that positive definiteness is a key condition for the applicability of a kernel in SVM learning. It was shown in [MSX+09] that $k_q^A$ is a positive definite kernel for $q \in [0, 1]$, while $k_q^B$ is a positive definite kernel for $q \in [0, 2]$. Standard results from kernel theory [STC04, Proposition 3.22] guarantee that the kernel $k$ defined in (15) inherits the positive definiteness of $k_q^i$, thus can be safely used in SVM learning algorithms. Moreover, we also employ nearest neighbor (NN) classifiers, in order to clearly assess the suitability of the derived kernels.

In these experiments, we selected a subset of three patients preserving the cancerous/benign cell ratio on which the two pathologists agree on the label. We have 474 (79 % of the 600 segmented nuclei), with the following proportions: 321 (67 %) benign and 153 (33 %) malignant; all the experiments are performed on this set of 474 nuclei images, which is divided into ten folds (with stratification). For the first paper, we use the all representations as in [SUCM11] except Prop. For the second one, we only use Phog. For each representation and fold, we learn a pLSA model from the training set and apply it to the test set. The number of topics has been chosen using leave-another-fold-out (of the nine training folds, we used 9-fold cross validation to estimate the best number of topics) cross validation procedure on the training set. We applied the same partitioning scheme also to choose the $q$ parameter in IT kernels. All reported accuracies are percentual accuracies and are the averages over 10 folds. In all experiments the standard errors around the mean were inferior to 0.02.

#### 4.4.4 Classification on the generative embedding space using pLSA [USB+11]

In our first paper, we compare our results using the original space and the generative embedding space using pLSA. In the obtained space, different classifiers have been tried. The obtained results have been compared with those obtained with the same classifier working on the original histograms (namely without the intermediate generative coding). In particular we employed the following classifiers (where not explicitly reported, all parameters have been tuned via cross validation on the training set)

- (svl): support vector machines with linear kernel (this represents the most widely employed solution with hybrid generative-discriminative approaches).
- (svp): support vector machines with polynomial kernel: after a preliminary evaluation, the degree $p$ was set to 2.
- (svr): support vector machines with radial basis function kernel.
- (knn): k-nearest neighbor classifier

All results were computed by using PRTools [Dui05] MATLAB toolbox. They are reported in tables 4. The feature representations where the proposed approach overperforms the original space are marked with bold face (statistically significant difference with paired $t$-test, $p = 0.05$). In particular, results are averaged over ten runs.

Table 4: Accuracies with SVM. ORIG is the original histogram based feature space, whereas PLSA stands for the proposed approach.

| | *svl* | | *svp* | | *svr* | | *knn* | |
|---|---|---|---|---|---|---|---|---|
| | ORIG | PLSA | ORIG | PLSA | ORIG | PLSA | ORIG | PLSA |
| All | 68.36 | **74.26** | 65.40 | **75.06** | 74.47 | 75.11 | 72.35 | **73.44** |
| Bg | 72.88 | 70.82 | 66.79 | **71.50** | **74.22** | 71.92 | **74.25** | 71.29 |
| Col | 66.90 | 69.03 | 56.93 | **70.32** | 68.98 | 68.82 | 69.41 | 68.62 |
| Fcc | 67.30 | 67.72 | 66.89 | 67.92 | 67.95 | 68.57 | 66.66 | 67.71 |
| Fg | 70.68 | 71.97 | 64.12 | **72.62** | 70.49 | 71.09 | 69.79 | 70.48 |
| Lbp | 68.61 | 69.43 | 42.36 | **70.70** | 68.79 | 70.47 | 71.13 | 70.29 |
| Phog | 75.45 | **79.67** | 63.92 | **79.22** | 76.55 | 76.80 | 70.71 | ***74.69** |
| Sig | 67.72 | **68.34** | 58.64 | **67.69** | 67.72 | 67.72 | 63.50 | 67.72 |

Observing the Table 4, we can see that the best accuracy using a SVM is 75.45% whereas the best accuracy on the pLSA space is 79.22 %. For most representations (except Lbp, Phog and Col), the accuracies of different kernels on the original space do not have large differences. We also observe that the data set is a difficult data set because there are some classifiers which have accuracy equal to the prior class distribution of the data set (67 per cent). We see that except the support vector machine with rbf kernel, the space constructed by pLSA always supercedes the original space (except Bg on *svl*) in terms of average accuracy. The bold face in the table shows feature sets where pLSA space is more accurate than the original space using 10-fold CV paired $t$-test at $p = 0.05$.

#### 4.4.5 Application of IT kernels on generative embedding spaces [BUS+11]

Observing the success on generative embedding spaces, we conducted some experiments on these spaces using IT kernels developed in the context of WP2.

In this setup, pLSA is trained in an unsupervised way, *i.e.*, we learn the pLSA model ignoring the class labels. Table 5 presents the results using the posterior distribution (referred to as `PLSA`) and the FESS embedding with SVM classification; these results show that in the proposed hybrid generative-discriminative approach, the IT kernels outperform linear and RBF kernels. The first and second columns show the classification results of $\psi$ and FESS scores classified using linear and RBF kernels which allows us to show the contribution of the IT kernels.

The results of the NN classifier are shown in Table 6. Although NN is not a good choice for this experiment (baseline NN accuracy using Mahalanobis distance on the original data is 64.57%), we still see the advantage of the IT kernels on the generative approach. We can achieve 72.74% and 72.53% using pLSA and FESS embeddings, respectively, when we use the similarities computed by the IT kernels in the NN classifier.

Table 5: Average accuracies (in percentage) using pLSA and FESS embeddings with SVMs. ORIG shows the baseline accuracies on the original feature space.

| | LIN | RBF | JS | JT | JT-W1 | JT-W2 |
|---|---|---|---|---|---|---|
| PLSA | 76.78 | 76.99 | 79.31 | 80.17 | 74.22 | 80.17 |
| FESS | 77.41 | 76.17 | 73.21 | 78.87 | 72.31 | 79.96 |
| ORIG | 75.45 | 76.55 | | N/A | | |

Table 6: Average accuracies (in percentage) using pLSA and FESS embeddings with NN classifiers. ORIG shows the baseline accuracies on the original feature space with Mahalanobis distances.

| | MB | JS | JT | JT-W1 | JT-W2 |
|---|---|---|---|---|---|
| PLSA | 66.41 | 68.97 | 72.53 | 72.74 | 68.75 |
| FESS | 67.11 | 67.08 | 72.53 | 71.27 | 71.08 |
| ORIG | 64.57 | | N/A | | |

# 5 Survival Analysis

Recall that the ultimate goal of TMA analysis is to determine the prognosis of the patient or to diagnose different cancer subtypes. The analysis of the proliferation marker MIB-1 allows the search for subgroups of patients which show different survival outcomes. Hence, the results of the previous two steps (Cell detection and classification) can be used to estimate the proportion of cells with particular properties (reflected by their staining with different antibodies), and ultimately their effect on patient prognosis.

## 5.1 Staining Classification

To differentiate a stained cell nucleus from a non-stained nucleus a simple color model was learned. Based on the labeled nuclei color histograms were generated for both classes based on the pixels within the average cell nuclei radius. To classify a nucleus on a test image the distance to the mean histograms of the both classes is calculated.

## 5.2 Kaplan-Meier estimates

The patients are split in two (1/2 : 1/2) groups based on the estimated percentage of cancerous nuclei which express MIB-1. Then the Kaplan-Meier estimator is calculated for each subgroup. This involves first ordering the survival times from the smallest to the largest such that $t_1 \leq t_2 \leq t_3 \leq \ldots \leq t_n$, where $t_j$ is the $j$th largest unique survival time. The Kaplan-Meier estimate of the survival function is then obtained as

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \left(1 - \frac{d_j}{r_j}\right) \tag{16}$$

where $r_j$ is the number of individuals at risk just before $t_j$, and $d_j$ is the number of individuals who die at time $t_j$ .

To measure the goodness of separation between two or more groups, the log-rank test (Mantel-Haenszel test) is employed which assesses the null hypothesis that there is no difference in the survival experience of the individuals in the different groups. The test statistic of the log-rank test (LRT) is $\chi^2$ distributed: $\hat{\chi^2} = \left[\sum_{i=1}^{m} (d_{1i} - \hat{e}_{1i})\right]^2 / \sum_{i=1}^{m} \hat{v}_{1i}$ where $d_{1i}$ is the number of deaths in the first group at $t_i$ and $e_{1i} = n_{1j}\frac{d_i}{n_i}$ where $d_i$ is the total number of deaths at time $t_{(i)}$, $n_j$ is the total number of individuals at risk at this time, and $n_{1i}$ the number of individuals at risk in the first group.

## 5.3 Survival Estimation

The only objective and undisputed target in the medical domain relates to the survival of the patient. The experiments described in Section 1 show the large disagreement between pathologists for the estimation of staining. Therefore, the adaption of an algorithm to the estimates of one pathologist or to a consensus voting of a cohort of pathologist is not desirable. Hence we validate the proposed algorithm against

the right censored clinical survival data of 133 patients. In addition these results were compared to the estimations of an expert pathologist specialized on renal cell carcinoma. He analyzed all spots in an exceptional thorough manner which required him more than two hours. This time consuming annotation exceeds the standard clinical practice significantly by a factor of 10-20 and, therefore the results can be viewed as an excellent human estimate for this dataset.

Figure 6 shows Kaplan-Meier plots of the estimated cumulative survival for the pathologist and RDF. The father the survival estimated of the two groups are separated the better the estimation. Quantifying this difference with log-rank test shows that the proposed algorithm is significantly ($p = 0.0113$) better than the trained pathologist ($p = 0.0423$).

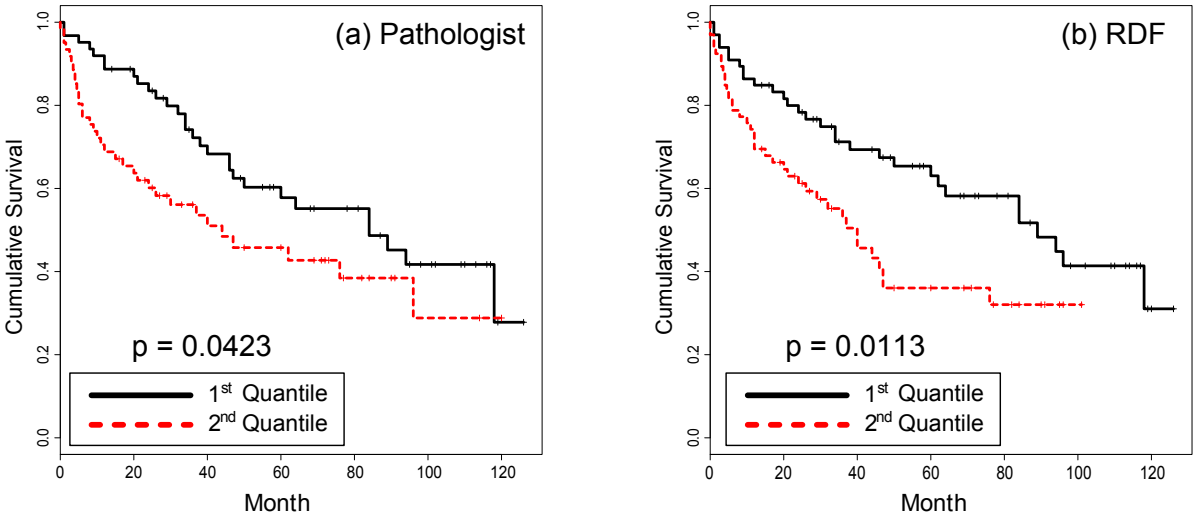



Figure 6: Kaplan-Meier estimators showing significantly different survival times for renal cell carcinoma patients with high and low proliferating tumors. Compared to the manual estimation from the pathologist (a) ($p = 0.04$), the fully automatic estimation from the algorithm (b) performs better ($p = 0.01$) in terms of survival prediction on the partitioning of patients into two groups of equal size.

# 6 Description of Software

The tissue micro array pipeline has been implemented in MATLAB as functions and scripts. Survival analysis and statistics have been performed in R. Calcuations were performed on a high performance computing cluster to capture the need of computational power. The implementation follows strictly the proposed pipeline, as described below.

Eight quarters of TMA images in .TIF file format were labeled by 2 pathologists (labels as .SVG file). The eight images serve as training and testing set for the generated code.

For the comprehensive nucleus detection procedure as proposed in Figure 1 (II), see 3.4.

A patch-generator (see Figure 1 (III)), implemented in MATLAB can extract squared image patches from the TMA images, showing one nucleus in the middle. Also available is a background patch-generator that extracts patches between the pathologists' labels showing surrounding tissue as background. These patches can be used for the nucleus detection task discriminating nuclei from background. All eight quarter TMA images together showed 1633 nuclei.

Graph-cut [Bag06] has been used for nucleus segmentation within the image patches. Parameters for graph-cut have been optimized to give optically the best segmentations around the nuclei. In general, intensity values were weights for the pixels and the distance from the center was the weight for the source and sink node.

A feature-extractor (see Figure 1 (IV)) transforms image patches in feature matrices. Features like intensity histograms, shape descriptors for segmentation (freeman chain code, 1D signature, MATLAB region properties), PHOG, local binary patterns and color information have been used for description of the patches.

In the similarity based pipeline, dissimilarity matrices are calculated using the feature vectors and various dissimilarity measures (both provided on the SIMBAD page). The combinatorically high number of dissimilarity matrices - 1633x1633 nuclei x around 10 features x around 20 dissimilarity measures - made the use of a high performance computing cluster unavoidable. One dissimilarity matrix is about 20-30 mb in size. To perform further classification tasks with support vector machines, the dissimilarity matrices are embedded to be positive semidefinite and so transformed to kernel matrices.

Classifiers are then trained on the similarity matrices and tested in a 8-fold cross validation. For classifation, we used the freely available lib-svm package for MATLAB [CL11]. Performances have been presented in this report.

The staining estimation (see Figure 1 (V)) is implemented as simple threshold method over the color space of the center of the nucleus patch. Brown nuclei are stained, whereas blue nuclei are not stained. All code is available on the SIMBAD site or at http://ml2.inf.ethz.ch/simbad/.

# 7 Conclusion

We have proposed an automated pipeline to provide objective and reproducible diagnosis of renal cell carcinoma. This pipeline involves three main components: cell nuclei detection from tissue microarray images, nucleus segmentation and classification into cancerous and healthy cells, and summarizing this information and analysing its effect on patient survival. This pipeline has been implemented as open source software and is available on the SIMBAD website.

The images and comprehensive annotations by two pathologists provide a rich resource for future medical imaging research. Our publicly available data enables objective benchmarking of methods and algorithms. Furthermore, the predictions can be validated against the human annotations, leading to a deeper understanding of the variations between pathologists and its impact on designing tools to overcome this variability.

As a demonstration of the usefulness of our benchmark, the novel approaches developed in the SIMBAD project have been benchmarked on various parts of the dataset constructed in work package 6. As discussed in this document, the cell nuclei data have been used to demonstrate the usefulness of Ricci flow embedding as well as to develop new kernel combination methods for nucleus classification. These approaches have resulted in several publications, showing the value of our data resource.

# References


[Bag06] Shai Bagon. Matlab wrapper for graph cut, December 2006.

[BK04] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, September 2004.

[BLJ04] Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the 21st International Conference on Machine Learning*, pages 41–48, 2004.

[BLOP10] Manuele Bicego, Pietro Lovato, Barbara Oliboni, and Alessandro Perina. Expression microarray classification using topic models. In *Proceedings of the 2010 ACM Symposium on Applied Computing, SAC '10*, pages 1516–1520, New York, NY, USA, 2010.

[BNJ03] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[BPM+10] M. Bicego, A. Perina, V. Murino, A.F.T. Martins, P.M.Q. Aguiar, and M.A.T. Figueiredo. Combining free energy score spaces with information theoretic kernels: Application to scene classification. In *Proceedings of the IEEE International Conference on Image Processing*, pages 2661–2664, 2010.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[BUS+11] Manuele Bicego, Aydın Ulaş, Peter J. Schüffler, Umberto Castellani, Pasquale Mirtuono, Vittorio Murino, Pedro M. Q. Aguiar André Martins, and Mário A. T. Figueiredo. Renal cancer cell classification using generative embeddings and information theoretic kernels. In Marco Loog et al., editor, *IAPR International Conference on Pattern Recognition in Bioinformatics, PRIB '11*, volume 7036 of *Lecture Notes in Bioinformatics*, page accepted. Springer Berlin / Heidelberg, November 2011.

[BVZ01] Yuri Boykov, Olga Veksler, and Ramin Zabih. Efficient approximate energy minimization via graph cuts. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 20(12):1222–1239, November 2001.

[BZM06] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Scene classification via pLSA. In *Proceedings of the European Conference on Computer Vision, ECCV '06*, pages 517–530, 2006.

[CL11] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[CMR10] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Learning non-linear combinations of kernels. In *Advances in Neural Information Processing Systems 22*, pages 396–404, 2010.

[DJSH04] Grignon DJ, Eble JN, Bonsib SM, and Moch H. Clear cell renal cell carcinoma. *World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs.*, IARC Press, 2004.

[Dui05] Robert P. W. Duin. Prtools, a matlab toolbox for pattern recognition version 4.0.14. available at http://www.prtools.org/, 2005.

[FLW+08] Thomas J. Fuchs, Tilman Lange, Peter J. Wild, Holger Moch, and Joachim M. Buhmann. Weakly supervised cell nuclei detection and segmentation on tissue microarrays of renal cell carcinoma. In *Pattern Recognition. DAGM 2008*, volume 5096 of *Lecture Notes in Computer Science*, pages 173–182. Springer Berlin / Heidelberg, 2008.

[GA08] Mehmet Gönen and Ethem Alpaydın. Localized multiple kernel learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 352–359, 2008.

[GSC+05] D. Glotsos, P. Spyridonos, D. Cavouras, P. Ravazoula, P. Arapantoni Dadioti, and G. Nikiforidis. An image-analysis system based on support vector machines for automatic grade diagnosis of brain-tumour astrocytomas in clinical routine. *Medical Informatics and the Internet in Medicine*, 30(3):179–193(15), September 2005.

[GUS+11] Mehmet Gönen, Aydın Ulaş, Peter J. Schüffler, Umberto Castellani, and Vittorio Murino. Combining data sources nonlinearly for cell nucleus classification of renal cell carcinoma. In Marcello Pelillo and Edwin Robert Hancock, editors, *Proceedings of the International Workshop on Similarity-Based Pattern Analysis, SIMBAD '11*, volume 7005 of *Lecture Notes in Computer Science*, pages 250–260. Springer Berlin / Heidelberg, September 2011.

[Hof01] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1–2):177–196, 2001.

[JH98] Tommi S. Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Proceedings of the conference on advances in neural information processing systems, NIPS '98*, volume 11, pages 487–493, Cambridge, MA, USA, 1998.

[KJ98] et. al. Kononen J, Bubendorf L. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med.*, Jul;4(7):844–7, 1998.

[Kuh55] Harold W. Kuhn. The hungarian method for the assignment problem:. *Naval Research Logistic Quarterly*, pages 2:83–97, 1955.

[KZ04] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts? *IEEE transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, February 2004.

[LCB+04] Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[LJN06] Darrin P. Lewis, Tony Jebara, and William S. Noble. Nonstationary kernel combination. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 553–560, 2006.

[MBM+10] A.F.T. Martins, M. Bicego, V. Murino, P.M.Q. Aguiar, and M.A.T. Figueiredo. Information theoretical kernels for generative embeddings based on hidden Markov models. In E.R. Hancock, R.C. Wilson, T. Windeatt, I. Ulusoy, and F. Escolano, editors, *Proceedings of the International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, volume 6218 of *Lecture Notes in Computer Science*, pages 463–472. Springer, 2010.

[MDK+07] Kirsten D. Mertz, Francesca Demichelis, Robert Kim, Peter Schraml, Martina Storz, Pierre-Andre Diener, Holger Moch, and Mark A. Rubin. Automated immunofluorescence analysis defines microvessel area as a prognostic parameter in clear cell renal cell cancer. *Human Pathology*, 38(10):1454–1462, October 2007.

[MH99] et. al. Moch H, Schraml P. High-throughput tissue microarray analysis to evaluate genes uncovered by cdna microarray screening in renal cell carcinoma. *Am J Pathol.*, Apr;154(4):981–6, 1999.

[MSX+09] Andre F. T. Martins, Noah A. Smith, Eric P. Xing, Pedro M. Q. Aguiar, and Mario A. T. Figueiredo. Nonextensive information theoretic kernels on measures. *Journal of Machine Learning Research*, 10:935–975, 2009.

[NA01] et. al. Nocito A, Bubendorf L. Microarrays of bladder cancer tissue are highly representative of proliferation index and histological grade. *J Pathol.*, Jul;194(3)::349–57, 2001.

[PCC+09a] Alessandro Perina, Marco Cristani, Umberto Castellani, Vittorio Murino, and Nebojsa Jojic. Free energy score space. In *Proceedings of the conference on advances in neural information processing systems, NIPS '09*, volume 22, pages 1428–1436, 2009.

[PCC+09b] Alessandro Perina, Marco Cristani, Umberto Castellani, Vittorio Murino, and Nebojsa Jojic. A hybrid generative/discriminative classification framework based on free-energy terms. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV '09*, pages 2058–2065, 29 2009-oct. 2 2009.

[R D09] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[RBCG08] Alain Rakotomamonjy, Francis R. Bach, Stephané Canu, and Yves Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[SaA07] Carolin Strobl and Anne-Laure Boulesteix andThomas Augustin. Unbiased split selection for classification trees based on the gini index. *Computational Statistics & Data Analysis*, 52(1):483–501, 2007.

[Soi03] Pierre Soille. *Morphological Image Analysis: Principles and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2003.

[STC04] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[SUCM11] Peter J. Schüffler, Aydın Ulaş, Umberto Castellani, and Vittorio Murino. A multiple kernel learning algorithm for cell nucleus classification of renal cell carcinoma. In *Proceedings of the International Conference on Image Analysis and Processing, ICIAP '11*, page accepted, September 2011.

[TA96] et. al. Tannapfel A, Hahn HA. Prognostic value of ploidy and proliferation markers in renal cell carcinoma. *Cancer*, Jan 1;77(1):164–71, 1996.

[TKR+02] Koji Tsuda, Motoaki Kawanabe, Gunnar Rätsch, Sören Sonnenburg, and Klaus-Robert Müller. A new discriminative kernel from probabilistic models. *Neural Computation*, 14:2397–2414, October 2002.

[TM01] et. al. Takahashi M, Rhodes DR. Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification. In *Proc Natl Acad Sci U S A.*, volume Aug 14;98(17), pages 9754–9, 2001.

[USB+11] Aydın Ulaş, Peter J. Schüffler, Manuele Bicego, Umberto Castellani, and Vittorio Murino. Hybrid generative-discriminative nucleus classification of renal cell carcinoma. In Marcello Pelillo and Edwin Robert Hancock, editors, *Proceedings of the International Workshop on Similarity-Based Pattern Analysis, SIMBAD '11*, volume 7005 of *Lecture Notes in Computer Science*, pages 77–88. Springer Berlin / Heidelberg, September 2011.

[Vap98] Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998.

[VJ01] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[YA01] et. al. Young AN, Amin MB. Expression profiling of renal epithelial neoplasms: a method for tumor classification and discovery of diagnostic molecular markers. *Am J Pathol.*, May;158(5):1639–51, 2001.

[YMF05] Lin Yang, Peter Meer, and David J. Foran. Unsupervised segmentation based on robust estimation and color active contour models. *IEEE Transactions on Information Technology in Biomedicine*, 9(3):475–486, 2005.