Mitochondria-based Renal Cell Carcinoma Subtyping: Learning from Deep vs. Flat Feature Representations

Peter J. Schüffler  
Department of Medical Physics  
Memorial Sloan Kettering Cancer Center  
New York City, NY, USA

Judy Sarungbam  
Department of Pathology  
Memorial Sloan Kettering Cancer Center  
New York City, NY, USA

Hassan Muhammad  
Weill Cornell Medicine  
New York City, NY, USA

Ed Reznik  
Computational Biology Program & Center for Molecular Oncology  
Memorial Sloan Kettering Cancer Center  
New York City, NY, USA

Satish K. Tickoo  
Department of Pathology  
Memorial Sloan Kettering Cancer Center  
New York City, NY, USA

Thomas J. Fuchs  
Department of Medical Physics  
Memorial Sloan Kettering Cancer Center  
New York City, NY, USA

Abstract

Accurate subtyping of renal cell carcinoma (RCC) is of crucial importance for understanding disease progression and for making informed treatment decisions. New discoveries of significant alterations to mitochondria between subtypes make immunohistochemical (IHC) staining based image classification an imperative. Until now, accurate quantification and subtyping was made impossible by huge IHC variations, the absence of cell membrane staining for cytoplasm segmentation as well as the complete lack of systems for robust and reproducible image based classification. In this paper we present a comprehensive classification framework to overcome these challenges for tissue microarrays of RCCs. We compare and evaluate models based on domain specific hand-crafted “flat”-features versus “deep” feature representations from various layers of a pre-trained convolutional neural network (CNN). The best model reaches a cross-validation accuracy of 89%, which demonstrates for the first time, that robust mitochondria-based subtyping of renal cancer is feasible.
Introduction

Renal cell carcinoma (RCC) arises in several subtypes that differ in histology, morphology and genetics. Clear cell (CC) RCC is the most common subtype, while relatively rare types include clear cell papillary (CCP) RCC, TCEB1-mutated RCC, and to a lesser extent MiTF/translocation family RCC and Tuberous sclerosis-associated RCC (Hakimi et al. (2015); Rohan et al. (2011); Argani et al. (2001)). These subtypes need to be differentiated from each other because of their different prognostic implications and different mechanisms for tumorigenesis.

Recent work by us (Reznik et al. (2016)) and others (Meierhofer et al. (2004); Nilsson et al. (2015); Davis et al. (2014)) has consistently shown that several subtypes of RCC show drastic changes in the number and activity of mitochondria. Alteration in mitochondrial copy number may play a significant role in cancer biology. CC RCC is for example genetically driven by the activation of hypoxia-inducible factor \( \text{HIF} \), a transcription factor that activates the cellular response to hypoxic stress. As a result, CC RCC is well-known to down-regulate the expression of metabolic and mitochondrial genes relative to normal tissue (Tickoo et al. (2000); The Cancer Genome Atlas Research Network (2013)).

In stark contrast, renal oncocytoma (ONC), a more rare benign tumor of the kidney, is uniquely characterized by the massive accumulation of mitochondria in the cytoplasm (Joshi et al. (2015)). Recent work has demonstrated that these mitochondria harbor mutations in their DNA (mtDNA), and that mitochondrial accumulation in ONC arises from defects in autophagy, which would normally clear defective mitochondria from the cell.

Notably, many rare subtypes of RCC remain genetically uncharacterized. Therefore, we examine in this study the histologic differentiation of ONC, CC, and CCP, which bears morphologic features resembling both the clear-cell and papillary subtypes of RCC (Rohan et al. (2011)). While CCP RCCs are distinguished from other RCCs by immunohistochemical staining patterns, disambiguation of CC from CCP tumors remains challenging (Sarungbam et al. (2016)).

For a systematic, objective and automated differentiation of these tumor subtypes, we use techniques drawn from computational pathology, which have shown promising predictive power in earlier studies.

Previous Work

RCC subtyping has been investigated previously on genetic profiles (e.g. Brannon et al. (2010)), tissue microarray readouts (e.g. Shi et al. (2004)), or H&E stained images features (e.g. Chaudry et al. (2009); Raza et al. (2009); Yeh et al. (2014)). To the best of our knowledge, there exists no subtyping based on mitochondria histology images.

Fuchs and Buhmann (2011) introduced the field of computational pathology offering advanced pipelines for automated classification of histology images. Deep learning with convolutional neural networks (CNNs) gain more and more impact in this scenario. Their applications range from cell detection and classification (Ciresan et al. (2013); Wang et al. (2013)) to subtyping (Reznik et al. (2016)).
Figure 1: Scheme of the mitochondria based RCC subtype classification. **Top row:** in a classical structured approach, the cells are detected in the images to access their cytoplasm intensity histograms as feature vectors. **Bottom row:** In contrast to that complicated multistep approach, we test state-of-the-art CNN models on whole images and image patches for feature extraction from AlexNet’s fully connected layers fc6, fc7 and fc8. Feature vectors are then fed to the same cross-validation procedure for predicting CC, CCP or ONC subtypes.

We present in this study a new approach to extract high performing CNN features from histology images of a RCC tissue microarray (TMA) for subsequent subtype classification. We address several problems such as scaling and patching and compare the different network layers as feature vectors (c.f. Fig 1, bottom row). Further, we compare our method...
Figure 2: **Example image data.** Brownish *TOM20* staining indicates mitochondria in cytoplasm. A: CC type, B: CCP type, C: ONC type. ONC usually has extremely over-replicated mitochondria, whereas CC and CPP have an under-represented number. The automatic distinction of the three classes by mitochondria is a key task of this paper.

to a dedicated and more complex cytoplasm-based approach, which yields quantitative histogram features from the mitochondria, in contrast to the less intuitive CNN features (c.f. Fig 1, top row). We show that CNNs lack in information when they look at the downscaled whole images (and histogram information helps), but the CNN outperforms both methods when they look at dedicated image patches.

**Methods**

75 TMA specimen replicates from 28 RCC patients (10 CC, 9 CCP, 9 ONC) were stained at the specific mitochondria receptor unit *TOM20*, resulting in brown color response. Cell nuclei were counter-stained with bluish *Eosine*. See Fig. 2 for three example images and S1 for a TMA overview. The TMA was scanned with an Aperio AT2 digital slide scanner at 20x magnification, resulting in single spot images of 1500x1500px size.

**Automatic White Balancing:** Each image has been corrected for scanning bias to get a neutral, white background. For this, each pixel was divided by the image's representative background color found by a window shifting algorithm over the image: each window $W$ is gray-scaled to $W_G$, and the normalized intensity histogram $I$ of $W_G$ is calculated. The entropy $H(I)$ is defined as:

$$H(I) = - \sum_i I_i \log I_i$$

where $I_i$ is the $i_{th}$ bin of $I$. The average color of the window $W$ with the smallest entropy $H(I)$ represents the background color.
**Cytoplasm Histogram Based Approach (HIST)**

To quantify the mitochondria in the images, we describe their specific cytoplasm intensity histograms. The color deconvolution algorithm by Ruifrok and Johnston (2001) was used to separate bluish nuclei from mitochondria staining (Fig 3A-C). Using the complete mitochondria channel for histograms could exclude bright clear cells and potentially include non-cell structures in the histograms. Since we aim to include the information of all cancer cells in the image and not only stained ones, we suggest prior cell detection and cytoplasm segmentation.

Since the accurate cell segmentation of RCC can be very hard due to missing membrane information (and is not part of this study), we approximate the cytoplasm as rings around the nuclei. For this, the nucleus positions are determined on the nucleus channel by a watershed based algorithm introduced earlier (Schüffler et al. (2013)).

The cytoplasm on the mitochondria channel is then outlined as circular rings around the nuclei with thickness of 10px, as optically confirmed to fit our cells (Fig 3D, purple). RCC TMA images may be disturbed by tissue lesions in which the background is visible. Therefore, we use a background intensity threshold of \( t = 220 \) to exclude brighter image background pixels from the image and rings (cf. Fig 3D, green). Therefore, the rings do not close properly, when the cells lie at a tissue border.

After this processing, the area on the rings defines the region of interest (ROI) of an image. The intensity histogram of the mitochondria channel in that ROI (cf. Fig 3E) is the basis of the feature vector for that image. The feature vector \( HIST \) has a length of 517 and consists of a pyramid of normalized ROI histograms with 256, 128, 64, 32, 16, 8, and 4 bins, the four 256-bin histogram quantiles \( q_1, q_2, q_3 \) and \( q_4 \), the mean and median intensity, skewness, kurtosis, and a histogram-based H-score defined as:

\[
H\text{-Score} = 100 \times \sum_{i=1}^{4} i \times q_i.
\]

![Figure 3](image)

Figure 3: Processing of histology images. From the white-balanced image (A shows detail), color deconvolution reveals the nuclei (B) and mitochondria (C). The cytoplasm approximated as all rings around the nuclei (D, purple) defines the ROI for the histogram (E). A simple threshold excludes white background (D, green).
Convolutional Neural Network

CNNs have seen striking advancements over recent years in the area of medical image classification. To compare this state-of-the-art technique with the more complex nucleus-based approach, we employ the use of CNNs to deconstruct our TMA images into feature vectors using two different methodologies: (i) we apply CNNs on the whole TMA spots and (ii) we incorporate a patching technique to sample small areas from the whole image using a mask to ignore areas without relevant information.

AlexNet by Krizhevsky et al. (2012) is a well studied and popular CNN pre-trained on natural images. It offers a much simpler structure than other nets (e.g. GoogleNet) and can be used for our purposes easily. Due to the relatively low number of samples, training a CNN from scratch would not be feasible for our study. Also, as our main goal is to explore the power of naive deep features, we do not fine-tune AlexNet for our purposes. Instead, we transfer the parameterization (tuning) step to random forests as primary classifiers.

The AlexNet architecture contains three fully connected layers ($f c_6$ and $f c_7$ with 4096 nodes (features), and $f c_8$ before a final softmax classifier, with 1000 nodes (features)). Note that softmax normalizes $f c_8$ to the probabilistic label vector of the CNN. Before that normalization, $f c_8$ can be interpreted as an important feature vector for classification. Because each layer represents a different set of abstract features, it is important to choose one that best contributes to a well-performing classifier. Feature vectors were extracted from all three fully connected layers with the Caffe Deep Learning framework (Jia et al. (2014)) for CNN on a cluster of 20 NVIDIA TitanX GPUs.

Whole Image Approach (CNN)

A main advantage of CNNs is the simplicity of their application on raw images which is why we test our approach with whole images. Since RCC subtyping is rotation invariant, and to increase the dataset for CNNs, we extend the original data set for whole image feature extraction by rotating the original images by $90^\circ$, $180^\circ$ and $270^\circ$ and flipping. The extended data set has $75 \cdot 4 \cdot 2 = 600$ images.

AlexNet has an input image size of 227x227px, and Caffe will downscale the whole image. This downscaling can cause loss of data due to a decrease in resolution. To explore this effect on classification performance, we introduce an image patching approach.

Image Patching Approach ($CNN_p$)

Each TMA spot was segmented into as many 227x227px partly overlapping patches that would fit within foreground areas of the TMA by a random sampling method according to following protocol (see Fig. 4):

To separate the background of an white-balanced image $I$, we defined a fixed threshold $t = 230$ which was applied on a Gaussian blurred ($\sigma = 2$), gray-scaled version of $I$, yielding the black/white foreground mask $BW$. To generate the patches, we first sampled 1000
random candidate patches of size 227x227px (AlexNet input size) uniformly over the whole image. A patch is then only accepted if:

- the corresponding foreground fills at least 80% of the patch (thus ignoring patches with large background content), and

- the patch has an overlap of the already patched area of at most 50% (thus reducing redundant patches, see Fig. 4), and

- the gray-scaled patch has an entropy $e \geq 4.6$, as empirically evaluated, excluding background structures not excluded by the background threshold (see Fig. 5).

Due to the high number of initial patch candidates, we cover as much as possible of the foreground area. Note that systematic sampling in a grid would include too many white patches, especially when the tissue is broken. On average, 29 patches were extracted per image (min 4, max 38 patches), or 76 patches per patient (min 7, max 106). In total, 1936 patches were sampled (see Fig. 6 for example patches). Patches were labeled with their parents’ class, and feature vectors were extracted for each patch from $fc_6$, $fc_7$, $fc_8$ analog to the whole image approach.
Figure 6: Examples of 227x227px image patches. A+B: CC cases, C+D: CCP cases, E: ONC case.

Classification and LOPO Cross-Validation

To evaluate the classification performance of the proposed methods HIST, CNN and CNNP, and their different feature vectors, we trained random forest models (Breiman (2001); Amit and Geman (1997)) with 50 trees on the generated feature vectors from each approach separately and combined to predict the cancer subtype CC, CCP or ONC. We chose 50 trees, as a larger number of trees did not show better performance (see S2). All other parameters (depth, number of features, etc) were not optimized and taken from the standards of the random forest implementation in R. Feature combination was done by concatenation of the vectors being simple and justified by the use of random forests, since they are not highly affected by different dynamic ranges or correlation.

Each classifier has been cross-validated in a leave-one-patient-out (LOPO-CV) scenario, training on the images of 27 patients and predicting the left-out patient’s images. A patient’s cancer subtype is determined as the class majority of his or her images. The multiple images per patient allow for an entropy-based confidence score for the patient’s class.

Results & Discussion

Cytoplasm Intensity Histograms

The 75 cytoplasm intensity histograms of the TMA showed differential shape among the three cancer groups (see Fig. 7): while the CC types have a camel like shape (two peaks), the CCP cases show a dromedary like shape (one peak) and usually little dark pixels. The ONC subtypes, in contrast, have a clear main emphasis on dark pixels. By the histograms alone, one could conclude the CC cases are a mixture of CCP and ONC.

Classification Performance

The mean error rate per class (balanced error, BE) is reported for patient classification into CC, CCP or ONC in Fig. 8. HIST has a BE of 21.5% in LOPO-CV (Fig. 8A). CNN reaches a BE of 27.8% (Fig. 8B) due to a higher number of misclassified CCs, which can be improved by adding HIST to 14.4%, (Fig. 8C). However, the best model is CNNP (fc6) with a BE of 11.1% (Fig. 8D), only misclassifying CCP.
An overview of all experiments with patient and image classification error is given in Fig. 9. It demonstrates that the histogram-based method (middle, blue) performs better than CNN on whole images, and can add to those features. But on the other hand, the superior models are CNNP’s without histograms. It is remarkable that the CNN\((fc6+fc7+fc8)\) and \((fc8)\) show considerable performance of 25-35% classification error, although the images are highly scaled down. To test if the CNN still learns from tissue information, or if the mitochondrial average stain in the (downscaled) image is correlated with the class, we learned a baseline random forest on the average staining intensity per image. The performance of that classifier with per patient error 39% and per image error 43% is lower than that of the CNN, supporting the hypothesis that the network can learn from the scaled images.

Further, the patching method works significantly better with \(fc6\) than e.g. for \(fc6+fc7+fc8\). An explanation can be that the concatenated feature vector will be too large (9k) compared to the number of samples (2k patches or 600 images) considering the fact it is better on the many patches than on the few whole images.

In general, ONC is easier to separate than CC and CCP. This is supported by a clustering of the intensity histograms clearly separating ONC while CC and CCP overlap with each other (shown in S3). Our best model CNNP \((fc6)\) has an AUC of 1 for ONC classification, and very high AUC (0.91 and 0.86) for CC and CCP classification (Fig. 9 top).

Figure 8: Selected confusion matrices of A HIST, B CNN (fc8), C HIST+CNN (fc8) and D CNNP (fc6) classifiers. Percentage of predicted patients per class shown. While the whole image CNN is improved by histogram information, the patched CNNP has access to more information and subtle details in the images.
Figure 9: Result of LOPO-CV. Generally, histogram features $HIST$ (middle) are better than $CNN$ on whole images (right side), but worse than $CNN_P$ on patches. The ROC of our best model $CNN_P(f \epsilon 6)$, has an AUC of $0.91$ (CC), $0.86$ (CCP) and $1$ (ONC).

Conclusion

The separation of RCC subtypes based on mitochondria quantification is an ongoing topic in cancer research, as mitochondria alterations seem to play a completely different role in RCC as in other types of cancer. We proposed a new solution to solve this difficult task based on structurally derived intensity histograms and state-of-the-art CNN feature vectors from three different layers of the well-studied AlexNet. We call these CNN features naïve, since the CNN has not been tuned to our dataset. Instead, the tuning process lies in the training of the random forest alone.

We compared these different approaches in a random forest cross-validation scenario and found CNNs on informative patches work best for this task. CNNs on whole images, however, may suffer from downscaling of the images (but still have moderate performance).

Finally, a classical histogram-based approach with hand-crafted features gives valuable results, even if worse than CNNs ($f \epsilon 6$). This more complicated pipeline might offer three benefits: (i) the nuclei are only sampled on tissue (foreground), while the CNN would input the whole image, regardless of semantics (patching solves this to a certain extend, but introduces complexity), (ii) nucleus detection allows for explicit nucleus classification (Schueffler et al. (2010)), introducing further semantics, and (iii) Handcrafted features might be better interpretable than deep features.

We showed that a pre-trained CNN can already have highly predictive power on a highly specific histology problem even without CNN tuning. It might be expected that tuning further improves the classification.
References


1. Tissue Microarray

Figure 10 shows the tissue micro array where the individual spots have been identified and numbered automatically. 96 RCC patients with specimen triplicates are present, of which 28 could be used for this study. A large part of these patients have unclassified tumors. The diagnosis of the patients was done by trained pathologists on the underlying whole tissue slides. The TMA cores have been selected by pathologists as typical representative sites of the corresponding cancer.

Figure 10: Overview of the tissue microarray of 92 RCC patients. 28 of these patients were diagnosed by the interesting RCC subtypes clear cell (CC), clear cell papillary (CCP) or oncocytoma (ONC). Each patient has specimen triplicates from primary tumor tissue, although individual tissue cores might be destroyed during preparation.
2. Number of Trees for Classification

We explored the random forest models’ number of trees effecting the classification performance during cross-validation. The classification error does not decrease significantly after 50 trees (see Fig 11). Still, the computation time increases (especially for the CNN features, data not shown). We chose 50 trees to be consistently fixed for all for our experiments as a empirically good trade-off of performance and computation time. E.g. the training of one random forest on $CNN_P^{fc6}$ layer (4096 features, 1845 samples (patches)), takes 4.0 minutes of a standard desktop PC, with 50 trees. With 100 trees, the computation time takes 8.8 minutes.

![Classification Error in Dependence of Number of Trees](image)

Figure 11: Classification error of a LOPO-CV in dependence of the number of trees in the forest. Classification accuracy per class shown. **Left:** LOPO-CV of $HIST(517$ features), **Right:** One fold of $CNN_P(4096$ features). We chose 50 trees for all experiments (gray vertical line).

3. Histogram Clustering

To explore the potential of histograms to separate the cases, we calculated the Kullback-Leibler divergence $KL$ Kullback and Leibler (1951) between all histograms and the mean histograms. As we do not have a ground truth distribution but only observed distributions, we symmetrize $KL$ between two histograms $I$ and $J$ by

$$KL_{sym} = \frac{KL(I, J) + KL(J, I)}{2}$$

As we hypothesize the histograms to be similar in each class, we expect the divergence to be small within a subtype and larger between the classes. The multidimensional scaling of the dissimilarity matrix, which is shown in Figure 12, confirms that hypothesis as the Kullback-Leibler intra-class divergence is smaller than the inter-class divergence. Still, certain overlap of the classes indicates that individual cases from different classes show similar histograms. Also here, we observe the CC class to be clustered in between CCP and ONC.
Figure 12: **Left:** Multidimensional scaling of the Kullback-Leibler divergence between all histograms of all 75 images plus the three mean histograms for the individual subtypes CC, CCP and ONC. Blue, CC; red, CCP; black, ONC. The intra-class divergence is generally smaller than the inter-class divergence, and classes can largely be separated. **Right:** Clustering of cRCC subtypes based on their mitochondrial staining by the multidimensional scaling of the CNN features. Red, CC. Blue, CCP. Green, ONC.