Renal Cell Carcinoma Staging with Learnable Image Histogram-based Deep Neural Network

Mohammad Arafat Hussain[1], Ghassan Hamarneh[2], and Rafeef Garbi[1]

[1] BiSICL, University of British Columbia, Vancouver, BC, Canada
[2] Medical Image Analysis Lab, Simon Fraser University, Burnaby, BC, Canada
{arafat,rafeef}@ece.ubc.ca, hamarneh@sfu.ca

Abstract. Renal cell carcinoma (RCC) is the seventh most common cancer worldwide, accounting for an estimated 140,000 global deaths annually. An important RCC prognostic predictor is its 'stage' for which the tumor-node-metastasis (TNM) staging system is used. Although TNM staging is performed by radiologists via pre-surgery volumetric medical image analysis, a recent study suggested that such staging may be performed by studying the image features of the RCC from computed tomography (CT) data. Currently TNM staging mostly relies on laborious manual processes based on visual inspection of 2D CT image slices that are time-consuming and subjective; a recent study reported about ∼25% misclassification in their patient pools. Recently, we proposed a learnable image histogram based deep neural network approach (ImHistNet) for RCC grading, which is capable of learning textural features directly from the CT images. In this paper, using a similar architecture, we perform the stage low (I/II) and high (III/IV) classification for RCC in CT scans. Validated on a clinical CT dataset of 159 patients from the TCIA database, our method classified RCC low and high stages with about 83% accuracy.

1 Introduction

Renal cell carcinoma (RCC) is the 7th most common cancer in men and 10th most common cancer in women [1] accounting for an estimated 140,000 global deaths annually [2]. The natural growth pattern varies across RCC, which has led to the development of different prognostic models for the assessment of patient-wise risk [3]. Clinical RCC staging is vital for proper treatment planning and thus considered one of the important prognostic predictors of cancer specific survival [4].

The American Joint Committee on Cancer (AJCC)/Union for International Cancer Control (UICC) specifies the criteria for tumor-node-metastasis (TNM) staging of each cancer depending on the primary tumor size (TX, T0-4); number and location of lymph node involvement (NX, N1-2); and metastatic nature, i.e. tumor spreading to other organs (M0-1) [3,5]. Clinical guidelines require clinicians to assign TNM stages prior to initiating any treatment [5]. AJCC TNM

is currently a manual process that includes two separate staging processes, performed before treatment planning and during/after surgery, to reflect the time-sensitive staging mechanism [3]. Clinical staging is performed prior to treatment by expert radiologists via physical examination, CT image measurements, and tumor biopsies. Clinically determined TNM stages (e.g. T or M) are designated with prefix 'c' (i.e. cT and cM). Pathological staging on the other hand is based on the resected tumor pathology results either during or after surgery [5], and designated with prefix 'p' (i.e. pT and pM). Radiologists also use the TNM description to assign an overall 'Anatomical stage' from 1 to 4 using the Roman numerals I, II, III, and IV [3], see Table 1. Accurate clinical staging of RCC

Table 1. Staging of RCC (AJCC TNM classification of tumors).

Stage I	T1 (Tumor $\leq$7 cm)	N0	M0
Stage II	T2 (Tumor >7 cm but limited to kidney)	N0	M0
Stage III	T1-2, T3 (Tumour extends up to Gerota's fascia)	N1, Any	M0
Stage IV	T4, Any (Tumour invades beyond Gerota's fascia)	Any	M0-1

is vital for appropriate management decisions [6]. Partial nephrectomy (PN), also known as nephron-sparing surgery, is typically preferred for T1 and T2 tumors [3]. After studying 7,138 patients with T1 kidney cancer, Tan et al. [7] suggested that treatment with PN was associated with improved survival. In a similar study on pT2 tumor patients, Janssen et al. [4] showed that patients having PN had a significantly longer overall survival. Radical nephrectomy (RN), which refers to complete removal of kidney with/without the removal of the adrenal gland and neighboring lymph node, is generally reserved for T3 and T4 tumors [6].

The pre-surgery clinical tumor staging often suffers from misclassification errors. For example, in a recent study, Bradley et al. [6] reported 23 disagreement cases between cT and pT stages of 90 patients. The study further reported that 5 patients were miscassified with cT3 but later down-staged to pT2, while 6 patients were misclassified with cT2 but later up-staged to pT3 for the same patient cohort ($\sim$12%). In another study on 1,250 patients who underwent nephrectomy, Shah et al. [8] reported 11% (140 patients) upstaging of tumors from cT1 to pT3. In addition, there was tumor recurrence in 44 patients (31.4% of the pT3 upstaged cases), where most of these patients initially had PN. These alarming findings suggest that PN is associated with better survival in low stage tumors (T1 and T2), while RN is associated with reduce recurrence in high stage (T3 and T4) tumors. However, high stage tumors (T3-4) are often misclassified as low stage (T1-2) in the clinical staging phase. In addition, we see in the rows 1-3 of Table 1 that the tumor criterion is not well defined for stages T1, T2, and T3. In contrast, anatomical stages I-IV defines better discrimination among tumor stages (see Table 1).

For accurate staging of RCC before treatment planning, contrast-enhanced abdominal CT is considered essential [3]. Although tumor staging is believed to be dependent on the tumor size, by studying the pT stages of 94 kidney samples, Bradley et al. [6] argued that stages > T3 does not always correlate with tumor size and suggested to use CT image features to improve tumor staging.

Supervised deep learning using convolutional neural networks (CNN) have gained popularity for automatic feature learning and classification. However, the learned features of a classical CNN tend to ignore diffuse textural features that are often important for applications such as our tumor staging problem. Very recently, we proposed ImHistNet, a deep neural network for end to end texture-based image classification [9]. In [9] we showed ImHistNet to be capable of learning complex and subtle task-specific textural features directly from images completely avoiding any pre-segmentation of the RCC as the learnable image histogram can stratify tumor and background textures well.

In this paper, we propose automatic low stage (I-II) and high stage (III-IV) RCC classification using ImHistNet. We demonstrate that RCC stages can be determined from the CT textural features of the tumor. Our approach learns a histogram directly from the CT data and deploys it to extract representative discriminant textural tumor features to correlate to RCC stages.

2 Materials and Methods

2.1 Data

We used CT scans of 159 patients from the TCIA database [10]. These patients were diagnosed with clear cell RCC, of which 95 were staged low (I-II) and 66 were staged high (III-IV). The images in this database have variations in CT scanner models, contrast administration, field of view, and spatial resolution. The in-plane pixel size ranged from 0.29 to 1.87 mm and the slice thickness ranged from 1.5 to 7.5 mm. We divided the dataset for training/validation/testing as 77/3/15 and 48/3/15 for stage low and stage high, respectively. Note that typical tumor radiomic analysis comprises [11]: (i) 3D imaging, (ii) tumor detection and/or segmentation, (iii) tumor phenotype quantification, and (iv) data integration (i.e. phenotype + genotype + clinical + proteomic) and analysis. Our approach falls under step-iii. The input data to our method are thus 2D image patches of size 64×64 pixels, taken from kidney+RCC (i.e. both mutually inclusively present) bounding boxes. We do not require any fine pre-segmentation of the RCC rather only assume a kidney+RCC bounding box, generated in step-ii. Given data imbalance where samples for stage high are fewer than for stage low, we allowed more overlap among adjacent patches for the stage low dataset. The amount of overlap is calculated to balance the samples from both cohorts.

2.2 Learnable Image Histogram for RCC Stage Classification

Learnable Image Histogram: Our learnable image histogram (LIH) [9] stratifies the pixel values in an image x into different learnable and possibly overlapping intervals (bins of width w_b) with learnable means (bin centers β_b). The

feature value $h_b(x) : b \in \mathcal{B} \rightarrow \mathcal{R}$, corresponding to the pixels in x that falls in the b^{th} bin, is estimated as:

$$h_b(x) = \Phi\{H_b(x)\} = \Phi\{\max(0, 1 - |x - \beta_b| \times \widetilde{w}_b)\}, \tag{1}$$

where $\mathcal{B}$ is the set of all bins, Φ is a global pooling operator, $H_b(x)$ is a piece-wise linear basis function that accumulates positive votes from the pixels in x that fall in the b^{th} bin of interval $[\beta_b - w_b/2, \beta_b + w_b/2]$, and $\widetilde{w}_b$ is the learnable weight related to the width w_b of the b^{th} bin: $\widetilde{w}_b = 2/w_b$. Any pixel may vote for multiple bins with different $H_b(x)$ since there could be an overlap between adjacent bins in our learnable histogram. The final $|\mathcal{B}| \times 1$ feature values from the learned image histogram are obtained using a global pooling Φ over each $H_b(x)$ separately. This pooling can be a 'non-zero elements count' (NZEC), which matches the convention of a traditional histogram, or can be an 'average' or 'max' pooling, depending on the task-specific requirement. The linear basis function $H_b(x)$ of the LIH is piece-wise differentiable and can back-propagate (BP) errors to update β_b and $\widetilde{w}_b$ during training. Readers are referred to the original manuscript on the ImHistNet [9] for details.

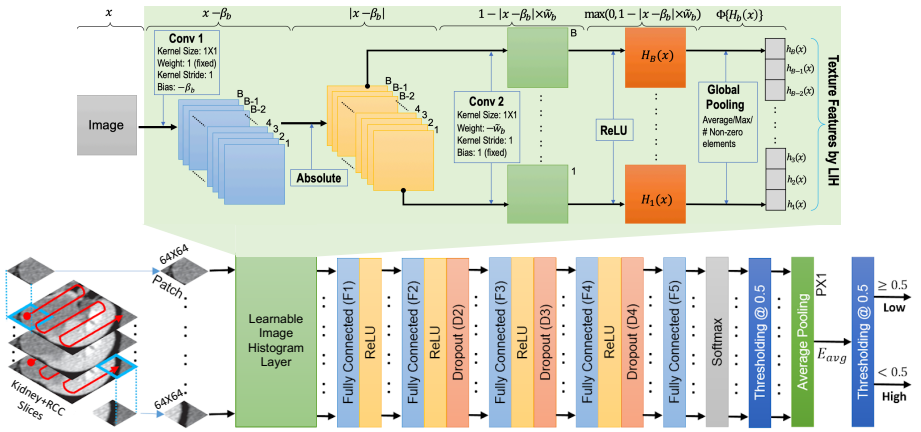

Fig. 1. Multiple instance decisions aggregated ImHistNet for stage classification.

Design of LIH using CNN Layers: The LIH is implemented using CNN layers as illustrated in Fig. 1. The input of LIH is a 2D image, and the output is a $|\mathcal{B}| \times 1$ histogram feature vector. The operation $x - \beta_b$ for a bin centered at β_b is equivalent to convolving the input by a 1×1 kernel with fixed weight of 1 (i.e. with no updating by BP) and a learnable bias term β_b ('Conv 1' in Fig. 1). A total of $B = |\mathcal{B}|$ number of similar convolution kernels are used for a set of $\mathcal{B}$ bins. Then an absolute value layer produces $|x - \beta_b|$. This is followed by a set of convolutions ('Conv 2' in Fig. 1) with a total of B separate (non-shared

across channels) learnable 1×1 kernels and a fixed bias of 1 (i.e. no updating by BP) to model the operation of $1 - |x - \beta_b| \times \widetilde{w}_b$. We use the rectified linear unit (ReLU) to model the $\max(0, \cdot)$ operator in Eq. 1. The final $|\mathcal{B}| \times 1$ feature values $h_b(x)$ are obtained by global pooling over each feature map $H_b(x)$ separately.

ImHistNet Classifier Architecture: The classification network comprises ten layers: the LIH layer, five (F1-F5) fully connected layers (FCLs), one softmax layer, one average pooling (AP) layer, and two thresholding layers (see Fig. 1). The first seven layers contain trainable weights. The input is a 128×128 pixel image patch extracted from the kidney+RCC slices. During training, randomly shuffled image patches are individually fed to the network. The LIH layer learns the variables β_b and $\widetilde{w}_b$ to extract representative textural features from image patches. In implementing the proposed ImHistNet, we chose $B = 128$ and 'average' pooling at $H_b(x)$. We set subsequent FCL (F1-F5) size to 4096×1. The number of FCLs plays a vital role as the overall depth of the model has been shown to be important for good performance [12]. Empirically, we achieved good performance with five FCL layers. Layers 8, 9 and 10 of the ImHistNet are used during the testing phase and do not contain any trainable weights.

Training: We trained our network by minimizing the multinomial logistic loss between the ground truth and predicted labels (1: stage low, and 0: stage high). We employed a Dropout unit (Dx) that drops 20%, 30%, and 40% of units in F2, F3 and F4 layers, respectively (Fig. 1) and used a weight decay of 0.005. The base learning rate was set to 0.001 and was decreased by a factor of 0.1 to 0.0001 over 250,000 iterations with a batch of 128 patches. Training was performed on a workstation with Intel 4.0 GHz Core-i7 processor, an Nvidia GeForce Titan Xp GPU with 12 GB of VRAM, and 32 GB of RAM.

RCC Stage Classification: After training ImHistNet (layers 1 to 7) by estimating errors at layer 7 (i.e. Softmax layer), we used the full configuration (from layer 1 to 10) in the testing phase. Although we used patches from only RCC-containing kidney slices during training and validation, not all the RCC cross-sections contained discriminant features for proper grade identification. Thus our trained network may miss-classify the interrogated image patch. To reduce such misclassification, we adopt a multiple instance decision aggregation procedure similar to our work in [13]. In this approach, we feed randomly shuffled single image patches as inputs to the model during training. During inference, we feed all candidate image patches of a particular kidney to the trained network and accumulate the patch-wise binary classification labels (0 or 1) at layer 8 (the thresholding layer). We then feed these labels into a $P \times 1$ average pooling layer, where P is the total number of patches of an interrogated kidney. Finally, we feed the estimated average (E_{avg}) from layer 9 to the second thresholding layer (layer 10), where $E_{avg} \geq 0.5$ indicates the stage 'low', and stage 'high' otherwise (see Fig. 1).

3 Results and Discussion

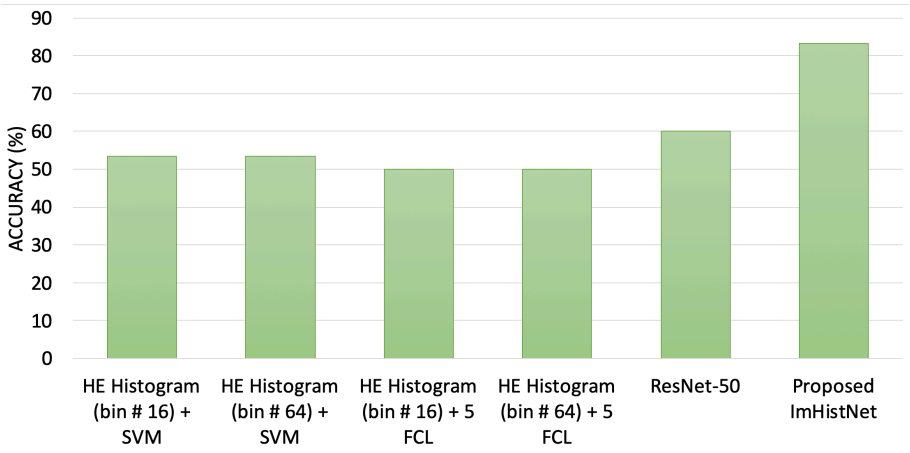

Fig. 2. Automatic RCC stage classification performance comparison. Acronyms used - HE: hand engineered, SVM: support vector machines.

We compared our RCC stage classification performance in terms of accuracy (%) to a wide range of methods in Fig 2. To our knowledge, there is no automatic and/or machine learning-based approach for RCC stage classification. Therefore, we compare the RCC staging performance of different methods by implementing those in our own capacity. Note that for all our implementations, we trained models with shuffled single image patches, and used multiple instance decision aggregation per kidney during inference. We fixed our patch size to 128×128 pixels across all contrasted methods except ResNet-50.

First, in order to compare performance to that of traditional hand-engineered (HE) feature based machine learning approaches, we evaluated an SVM employing a conventional image histogram of 16 and 64 bins. Fig. 2 shows a resulting poor performance at 53% accuracy. Next, to contrast the performance of SVM against DNN, we fed the conventional histogram (16 and 64 bins) features to a DNN of 5 FCL with weight sizes (4096×1)-(4096×1)-(4096×1)-(4096×1)-(2×1). We chose this FCL configuration for fairer comparisons since our ImHistNet contains the same. Fig. 2 shows that the FCL with conventional histogram performed the worst achieving a 50% accuracy. Next, we used ResNet-50 with transfer learning in order to test the performance of high performing modern CNN (see Fig. 2). We used full kidney+RCC slices of size 224×224 pixels as input. As mentioned in Sect. 1, a classical CNN typically fails to capture textural features, which is evident from our results where ResNet-50 performed poorly in learning the textural features of RCC, resulting in 60% accuracy. Finally, we show the performance of our proposed method in Fig. 2 where ImHistNet achieved the highest accuracy (83%) among all contrasted methods.

4 Conclusions

We proposed a powerful automatic RCC stage classification method that uses a learnable image histogram based deep neural network framework we recently proposed for tumor grading. Our approach learns a histogram directly from the image data and deploys it to extract representative discriminant textural image features. We also used multiple instance decision aggregation to further robustify binary classification. Our proposed ImHistNet outperformed competing approaches for this task including including SVM classification, deep learning with hand crafted traditional histogram features, as well as currently top performing deep CNNs. ImHistNet appears to be very well-suited for radiomic studies, where learned textural features using the learnable image histogram can aid in improving diagnosis accuracy.

Acknowledgement: We thank NVIDIA Corporatoin for supporting our research through their GPU Grant Program by donating the GeForce Titan Xp.

References

1. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2016. CA: a cancer journal for clinicians **66**(1) (2016) 7–30
2. Ding, J., Xing, Z., Jiang, Z., Chen, J., Pan, L., Qiu, J., Xing, W.: CT-based radiomic model predicts high grade of clear cell renal cell carcinoma. European journal of radiology **103** (2018) 51–56
3. Escudier, B., Porta, C., Schmidinger, M., Rioux-Leclercq, N., Bex, A., Khoo, V., Gruenvald, V., Horwich, A.: Renal cell carcinoma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology **27**(suppl_5) (2016) v58–v68
4. Janssen, M., Linxweiler, J., Terwey, S., Rugge, S., Ohlmann, C.H., Becker, F., Thomas, C., Neisius, A., Thüroff, J., Siemer, S., et al.: Survival outcomes in patients with large ($\geq$ 7cm) clear cell renal cell carcinomas treated with nephron-sparing surgery versus radical nephrectomy: Results of a multicenter cohort with long-term follow-up. PloS one **13**(5) (2018) e0196427
5. AAlAbdulsalam, A.K., Garvin, J.H., Redd, A., Carter, M.E., Sweeny, C., Meystre, S.M.: Automated extraction and classification of cancer stage mentions from unstructured text fields in a central cancer registry. AMIA Summits on Translational Science Proceedings **2018** (2018) 16
6. Bradley, A., MacDonald, L., Whiteside, S., Johnson, R., Ramani, V.: Accuracy of preoperative CT T staging of renal cell carcinoma: which features predict advanced stage? Clinical radiology **70**(8) (2015) 822–829
7. Tan, H.J., Norton, E.C., Ye, Z., Hafez, K.S., Gore, J.L., Miller, D.C.: Long-term survival following partial vs radical nephrectomy among older patients with early-stage kidney cancer. Jama **307**(15) (2012) 1629–1635
8. Shah, P.H., Moreira, D.M., Patel, V.R., Gaunay, G., George, A.K., Alom, M., Kozel, Z., Yaskiv, O., Hall, S.J., Schwartz, M.J., et al.: Partial nephrectomy is associated with higher risk of relapse compared with radical nephrectomy for clinical stage T1 renal cell carcinoma pathologically up staged to T3a. The Journal of urology **198**(2) (2017) 289–296

9. Hussain, M.A., Hamarneh, G., Garbi, R.: ImHistNet: Learnable image histogram based DNN with application to noninvasive determination of carcinoma grades in CT scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2019) 1–8
10. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., et al.: The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of digital imaging **26**(6) (2013) 1045–1057
11. Aerts, H.J.: The potential of radiomic-based phenotyping in precision medicine: a review. JAMA oncology **2**(12) (2016) 1636–1642
12. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision, Springer (2014) 818–833
13. Hussain, M.A., Hamarneh, G., Garbi, R.: Noninvasive determination of gene mutations in clear cell renal cell carcinoma using multiple instance decisions aggregated cnn. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2018) 657–665