

# Learnable Image Histograms-based Deep Radiomics for Renal Cell Carcinoma Grading and Staging

Mohammad Arafat Hussain<sup>a</sup>, Ghassan Hamarneh<sup>b</sup>, Rafeef Garbi<sup>a</sup>

<sup>a</sup>*BiSICL, University of British Columbia, Vancouver, BC V6T 1Z4, Canada*

<sup>b</sup>*Medical Image Analysis Lab, Simon Fraser University, Burnaby, BC V5A 1S6, Canada*

---

## Abstract

Fuhrman cancer grading and tumor-node-metastasis (TNM) cancer staging systems are typically used by clinicians in the treatment planning of renal cell carcinoma (RCC), a common cancer in men and women worldwide. Pathologists typically use percutaneous renal biopsy for RCC grading, while staging is performed by volumetric medical image analysis before renal surgery. Recent studies suggest that clinicians can effectively perform these classification tasks non-invasively by analysing image texture features of RCC from computed tomography (CT) data. However, image feature identification for RCC grading and staging often relies on laborious manual processes, which is error prone and time-intensive. To address this challenge, this paper proposes a learnable image histogram in the deep neural network framework that can learn task-specific image histograms with variable bin centers and widths. The proposed approach enables learning statistical context features from raw medical data, which cannot be performed by a conventional convolutional neural network (CNN). The linear basis function of our learnable image histogram is piece-wise differentiable, enabling back-propagating errors to update the variable bin centers and widths during training. This novel approach can segregate the CT textures of an RCC in different intensity spectra, which enables efficient Fuhrman low (I/II) and high (III/IV) grading as well as RCC low (I/II) and high (III/IV) staging. The proposed method is validated on a clinical CT dataset of 159 patients from The Cancer Imaging Archive (TCIA) database, and it demonstrates 80% and 83% accuracy in RCC grading and staging, respectively.

---

*Email addresses:* arafat@ece.ubc.ca (Mohammad Arafat Hussain), hamarneh@sfu.ca (Ghassan Hamarneh), rafeef@ece.ubc.ca (Rafeef Garbi)

*Keywords:* Learnable image histogram, deep neural network, renal cell carcinoma, cancer grade, cancer stage

---

## 1. Introduction

Renal cell carcinoma (RCC) is the seventh most common in men and tenth most common in women, accounting for an estimated 140,000 global deaths annually (Ding et al., 2018). The biological aggressiveness of RCC affects the prognosis and treatment planning (Ishigami et al., 2014). The natural growth pattern varies across RCC, which has led to the development of different prognostic models to assess patient-wise risk (Escudier et al., 2016). The ‘grade’ and ‘stage’ of an RCC are the critical prognostic predictors of cancer-specific survival (Janssen et al., 2018), where higher-grade and higher-stage tumors have an elevated risk of postoperative recurrence (van der Mijn et al., 2019). Typically, radiologists rely on the expertise of pathologists, who use the 4-tiered Fuhrman grading system (FGS) (Fuhrman et al., 1982) for RCC grading by examining the histopathologic images of RCC samples. Although the International Society of Urological Pathology (ISUP) introduced a new grading system for clear cell and papillary RCCs, which is also incorporated in the World Health Organization (WHO) renal tumor classification system (Delahunt et al., 2013), Fuhrman grading is still widely used for RCC grading in the clinical treatment planning.

Anatomical Stages	T Stages	N Stages	M Stages
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

Table 1: The American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (UICC) specified criteria for RCC staging.

In contrast, radiologists perform RCC staging by examining the physical extent, characteristics, and aggressiveness of a tumor in the volumetric medical image. The American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (UICC) specified the criteria for tumor-node-metastasis (TNM) staging of each cancer as shown in Table 1. Tumor stage depends on the primary tumor size (T0-4), number and location of lymph node involvement (N1-2), and metastatic nature, i.e., tumor

spreading to other organs (M0-1) (Escudier et al., 2016; AAlAbdulsalam et al., 2018). Clinical guidelines require clinicians to assign TNM stages before initiating any treatment (AAlAbdulsalam et al., 2018).

### *1.1. RCC Grading*

Accurate grading of RCC is essential in treatment planning as cancer-specific survival (CSS) correlates with grades. Kuthi et al. (2017) reported that the rate of 5-year CSS for Fuhrman grade I and II of RCC is significantly higher than that of RCC grade III and IV. Besides, low-grade RCCs are managed with minimally invasive techniques, while high-grade RCCs are treated with radical operation (Lin et al., 2019). Conventionally, pathologists use the 4-tiered FGS (Fuhrman et al., 1982) for RCC grading. However, to reduce the variability and improve the reproducibility of the tumor grade, a simplified 2-tiered FGS is preferred by pathologists in current clinical practice (Ding et al., 2018; Shu et al., 2018; Ishigami et al., 2014). The 2-tier FGS, which divides grades to low grade (Fuhrman I/II) and high grade (Fuhrman III/IV), is shown to be as effective as 4-tiered FGS in predicting cancer-specific mortality in a study population of 2,415 clear cell RCC (ccRCC) patients (Becker et al., 2016). Nevertheless, recent studies (Ding et al., 2018; Jeon et al., 2016) reported that the inter-observer reproducibility of grades assigned by pathologists ranges from 31.3% to 97%. In addition, renal biopsy often causes complications, such as hemorrhage and infection (Lin et al., 2019). This scenario motivates the development of a medical image-based automatic, noninvasive, and reproducible system of RCC grading. Several machine learning approaches (Tian et al., 2019; Chen et al., 2020) have been proposed for Fuhrman grading using histopathological images. A few predictive models for FGS have also been proposed (Lane et al., 2007; Jeldres et al., 2009) using clinical variables such as patient age, gender, symptoms and tumor size; however, they showed an accuracy close to that of flipping a coin ( $\sim 50\%$ ) (Ding et al., 2018).

Recently, Oh et al. (2017) assessed the correlation between computed tomography (CT) features and Fuhrman grade of ccRCC, where ccRCCs were retrospectively reviewed in consensus by two radiologists. Using logistic regression (LR), they showed a threshold tumor size of 36 mm to predict (AUC: 70%) the high Fuhrman grade. Sasaguri and Takahashi (2018) suggested that RCCs can be characterized and graded based on CT textural features. Deng et al. (2019) argued that CT-based filtration-histogram parameters are correlated to biological RCC characteristics like glucose metabolism, hypoxia, and

tumor angiogenesis. Ding et al. (2018) employed LR on both non-textural features like the pseudo capsule, round mass, as well as textural ones like histogram, gray-level co-occurrence matrices (GLCM), gray level run length matrix (GLRLM), and reported that textural features better discriminated high from low-grade ccRCC. Shu et al. (2018) also employed LR on CT textural features, e.g., GLCM, GLRLM, gray level size zone matrix (GLSZM), and achieved an FGS accuracy of 77%. In a subsequent study, Shu et al. (2019) used the similar textural features in k-nearest neighbor (KNN), LR, multi-layer perceptron (MLP), random forest (RF), and support vector machine (SVM) to classify ccRCCs in low (WHO/ISUP I-II) and high (WHO/ISUP III-IV) grades. Huhdanpaa et al. (2015) used histogram analysis of the peak tumor enhancement, tumor heterogeneity, and percent contrast washout in CT. They reported these parameters to be statistically different between low and high-grade ccRCC. In recent years, several studies (Yu et al., 2017; Bektas et al., 2019; Feng et al., 2019; Lin et al., 2019; Sun et al., 2019; Haji-Momenian et al., 2020; Yan et al., 2020; Nazari et al., 2020) used textural features in terms of histogram, gradient, run-length matrix and co-occurrence matrix in different conventional machine learning methods like SVM, MLP, naive Bayes, KNN, RF, etc. for RCC Fuhrman grading, and showed AUC in the range of 0.73~0.87. A few artificial neural network-based approaches (Kocak et al., 2019; He et al., 2020) have also been proposed for Fuhrman low and high, and ISUP low and high grading of RCC. These methods also use hand-engineered textural features. These histogram, GLCM, GLRLM, GLSZM, and other tumor intensity-based features (e.g., peak tumor enhancement, tumor heterogeneity, etc.) are known as statistical context features (Wang et al., 2016) and are found to be very useful for RCC grade classification.

### *1.2. RCC Staging*

Similar to RCC grade, information on RCC stages significantly helps clinicians in treatment planning and outcome prediction. Studies (Bradley et al., 2015; Janssen et al., 2018) suggested that nephron-sparing surgery in patients having lower-stage tumors significantly improves cancer-specific survival. In contrast, complete removal of a kidney with/without removing the adrenal gland and neighboring lymph nodes in patients having higher-stage tumors improves the survival time. Typically, clinicians perform kidney tumor staging based on the tumor size and its extent. However, Bradley et al. (2015) argued that the correlation trend between tumor size and stage

often deviates for higher tumor stages, and thus, suggested using CT image features to improve tumor staging.

TNM staging of RCC is currently a manual process, which radiologists perform twice for the same patient in the clinical workflow (Escudier et al., 2016). The first evaluation of the tumor stage in the workflow is called ‘clinical’ staging, which radiologists perform before treatment via physical examination and CT image measurements of a tumor. Clinicians designate the determined TNM stages with the prefix ‘c’ (i.e., cT and cM). The final evaluation of the tumor stage is called ‘pathological’ staging and is based on the resected tumor pathology results either during or after surgery (AAIAbdulsalam et al., 2018). Pathologists designate this estimated stage with the prefix ‘p’ (i.e., pT and pM). Clinical staging (i.e., cT, cM) of RCC is primarily used for treatment management decisions (Bradley et al., 2015). For example, partial nephrectomy (PN), also known as nephron-sparing surgery, is typically preferred for cT1 and cT2 tumors (Escudier et al., 2016). After studying 7,138 patients with pT1 kidney cancer, Tan et al. (2012) suggested that treatment with PN was associated with improved survival. In a similar study on pT2 tumor patients, Janssen et al. (2018) showed that patients having PN had a significantly longer overall survival. Radical nephrectomy (RN), which refers to complete removal of a kidney with/without removing the adrenal gland and neighboring lymph node, is generally reserved for cT3 and cT4 tumors (Bradley et al., 2015).

The presurgery clinical tumor staging often suffers from miss-classification errors. For example, in a recent study, Bradley et al. (2015) reported 23 disagreement cases between cT and pT stages of 90 patients. The study further indicated that five patients were miss-classified with cT3 but later downstaged to pT2, while six patients were miss-classified with cT2 but later upstaged to pT3 for the same patient cohort ( $\sim 12\%$ ). In another study on 1,250 patients who underwent nephrectomy, Shah et al. (2017) reported 11% (140 patients) upstaging of tumors from cT1 to pT3. Besides, there was tumor recurrence in 44 patients (31.4% of the pT3 promoted 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 reduced recurrence in high stage (T3 and T4) tumors. However, high stage tumors (T3-4) are often miss-classified as the low stage (T1-2) in the clinical staging phase. Additionally, we see in rows 1-3 of Table 1 that the tumor classifying criterion is not well defined for stages T1, T2, and T3. Therefore, radiologists often use the TNM description to assign

an overall ‘Anatomical stage’ from 1 to 4 using the Roman numerals I, II, III, and IV (Escudier et al., 2016), see Table 1.

For accurate staging of RCC before treatment planning, contrast-enhanced abdominal CT is considered essential (Escudier et al., 2016). Typically, clinicians recognize the tumor size for tumor staging. Although several machine learning approaches (Coy et al., 2019; Schieda et al., 2020; Yap et al., 2020) have been proposed for classifying solid renal mass between benign and malignant cases, by studying the pT stages of 94 kidney samples, Bradley et al. (2015) argued that the correlation trend between the tumor size and stage often deviates for stages beyond T3. Thus, they suggested using CT image-based textural features to improve tumor staging, like in the FGS system. Recently, Ökmen et al. (2019) used CT textural features in KNN for TNM staging of RCC.

### *1.3. Learning Textural Features*

Despite the importance of textural features for image classification tasks, identifying such features from images relies on human visual inspection, which is difficult, time-consuming, and suffers from a lack of quantification. To overcome the limitations of manual feature engineering, supervised deep learning using convolutional neural networks (CNN) has exploded in popularity. In a classical CNN, the first layer’s learned features typically capture low-level features such as edges. The second layer detects motifs by spotting particular arrangements of edges. The third layer assembles motifs into larger combinations representing parts of objects, and subsequent layers detect objects as combinations of these parts (LeCun et al., 2015). These features are nonstatistical context features (Wang et al., 2016) and the classical CNN tends to put less emphasis on the diffused statistical textural features that are often important, especially for medical imaging applications like tumor analysis. In an attempt to learn statistical textural features via CNNs in computer vision tasks, Andrearczyk and Whelan (2016) proposed deploying a global average pooling over each feature map of the last convolution layer of a conventional CNN to make the model object shape unaware. However, the pooling still operates on the learned object-edge/motifs that do not capture complex and subtle textural variations in the input image. In a recent study, Wang et al. (2016) proposed an approach to learn histograms that back-propagates errors to learn optimal bin centers and widths during training. Wang’s approach has 2 stages: in stage-1, a conventional CNN learns the appearance feature maps followed by producing a class likelihood (for

classification) or likelihood map (for segmentation). A learnable histogram is subsequently trained on the stage-1 likelihood estimates, and the resultant features of this histogram are concatenated with the appearance features learned in stage-1. The combined appearance plus histogram features are then used to produce a fine-tuned stage-2 likelihood-map/class-likelihood, which results in a slightly better (1.9%) prediction accuracy. We emphasize that this method does not directly learn histogram features from the image, instead works on the CNN-produced appearance features.

Learning statistical textural features directly from images using CNN is vital for medical imaging applications, e.g., tumor characterization and analysis. It is also evident from earlier works (Ding et al., 2018; Shu et al., 2018; Huhdanpaa et al., 2015) that CT intensity-based statistical features can be used for RCC grading, and suggested being used for improved RCC staging (Bradley et al., 2015). Therefore, it is necessary to develop a deep neural network (DNN)-based texture learning approach for automatic tumor characterization.

#### 1.4. Contribution

We propose ImHistNet, a DNN for an end-to-end texture-based image classification approach. The preliminary version of this work appeared in Husain et al. (2019a,b). ImHistNet has the following contributions:

1. **Learnable Image Histogram (LIH):** We propose an LIH layer within a DNN framework, capable of learning complex and subtle task-specific textural features from raw images. Different from the work of Wang et al. (2016), our ImHistNet learns the global statistical features directly from the image intensity.
2. **No Tumor Segmentation:** We remove the requirement for the fine pre-segmentation of the RCC. The proposed learnable image histogram can stratify tumor and background textures well, thus enabling the model to focus specifically on the tumor texture.
3. **RCC Grading:** We demonstrate ImHistNet’s capabilities by performing automatic RCC grade classification for the 2-tiered FGS on an extended clinical dataset from real patients.
4. **RCC Staging:** We also demonstrate ImHistNet’s capability in the automatic categorization of RCC into anatomical stage low (I/II) and high (III/IV) on an extended clinical dataset from real patients; the proposed method is the first and only work that performs CT-based

RCC staging using deep learning. It is an important finding of our experiments that we observe a correlation between the RCC stages and the deep-learned CT textural features, which to our knowledge, no one thoroughly investigated to date. As TNM staging criteria have overlaps among the classes, we aim to perform automatic anatomical staging instead of TNM staging (see Table 1).

In this paper, we present additional experimental findings, results, and discussions over our previous work (Hussain et al., 2019a,b) as follows:

1. **Intensity Stratification by LIH:** We demonstrate our new experimental findings on how the learnable bins stratify the CT intensities to facilitate task-specific textural feature learning.
2. **Ability of LIH to Pick Task-specific Intensity Spectra:** We perform additional experiments to find on which intensity spectra the learnable bins of our ImHistNet put emphasis on for RCC grading and staging.
3. **Efficacy of LIH in ImHistNet:** To show the efficacy of our proposed LIH layer further, we perform additional experiments by switching off the LIH layer in the ImHistNet, and report the performance on RCC grading and staging.
4. **Comprehensive Discussion on ImHistNet Implementation:** We discuss the specific architecture and implementation of the ImHistNet in detail.

## 2. Methods

In this section, we first describe the learnable image histogram layer of the proposed ImHistNet. Then we describe the implementation details of the learnable image histogram layer via traditional CNN filters/operations. After that, we outline our classification network (i.e., ImHistNet) that leverages the LIH layer to classify RCC grades and stages from CT. We comprehensively discuss the technical insights of LIH and its implementation details in this section, which were not covered in our preliminary work (Hussain et al., 2019a,b).

### 2.1. Learnable Image Histogram

Our proposed learnable image histogram (LIH) stratifies the pixel values in an image  $x$  into different learnable and possibly overlapping intervals (bins



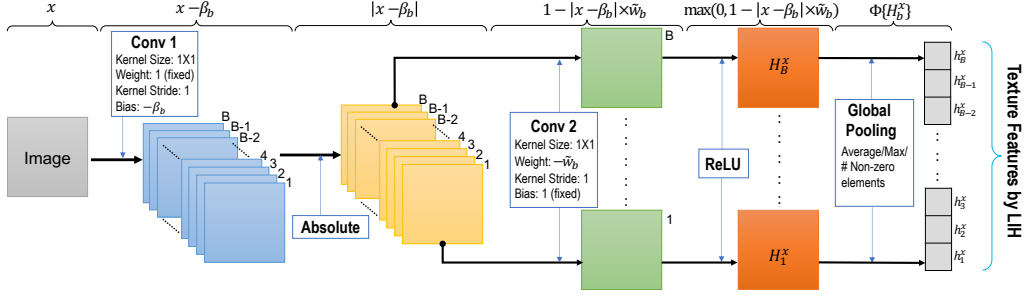


Figure 1: The graphical representation of the architecture of our learnable image histogram using CNN layers. We also break down our piece-wise linear basis function  $H_b^x$  on top of the figure in relation to different parts of the learnable image histogram architecture.

of width  $w_b$ ) with arbitrary learnable means (bin centers  $\beta_b$ ). Given a 2D image (or a 2D region of interest or patch)  $x : \mathcal{R}^2 \rightarrow \mathcal{R}$ , the feature value  $h_b^x : b \in \mathcal{B} \rightarrow \mathcal{R}$ , corresponding to the number of pixels in  $x$  whose values fall within the  $b^{\text{th}}$  bin, is estimated as:

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

where  $\mathcal{B}$  is the set of all bins,  $\Phi$  is the global pooling operator,  $H_b^x$  is the piece-wise linear basis function that accumulates positive votes from the pixels in  $x$  that fall in the  $b^{\text{th}}$  bin of interval  $[\beta_b - w_b/2, \beta_b + w_b/2]$ , and  $\tilde{w}_b$  is the learnable weight related to the width  $w_b$  of the  $b^{\text{th}}$  bin:  $\tilde{w}_b = 2/w_b$ . 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  $\Phi$  over each  $H_b^x$  separately. Depending on the task-specific requirement, the pooling can be the nonzero element count:  $\Phi\{H_b^x\} = \sum_p^P \sum_q^Q 1$  for  $H_b^x(p, q) > 0$ , max-pooling:  $\Phi\{H_b^x\} = \max(H_b^x)$ , or average-pooling:  $\Phi\{H_b^x\} = \text{mean}(H_b^x)$ , where  $P \times Q$  is the total number of pixels in  $x$ . The linear basis function  $H_b^x$  of the LIH is piece-wise differentiable and can back-propagate (BP) errors to update  $\beta_b$  and  $\tilde{w}_b$  during training. The gradients of  $\beta_b$  and  $\tilde{w}_b$  for a loss  $\mathcal{L}$  are calculated from Eq. 1 as:

$$\frac{\partial \mathcal{L}}{\partial \beta_b} = \begin{cases} \tilde{w}_b & \text{if } H_b^x > 0 \text{ and } x - \beta_b > 0, \\ -\tilde{w}_b & \text{if } H_b^x > 0 \text{ and } x - \beta_b < 0, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

$$\frac{\partial \mathcal{L}}{\partial \tilde{w}_b} = \begin{cases} -|x - \beta_b| & \text{if } H_b^x > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

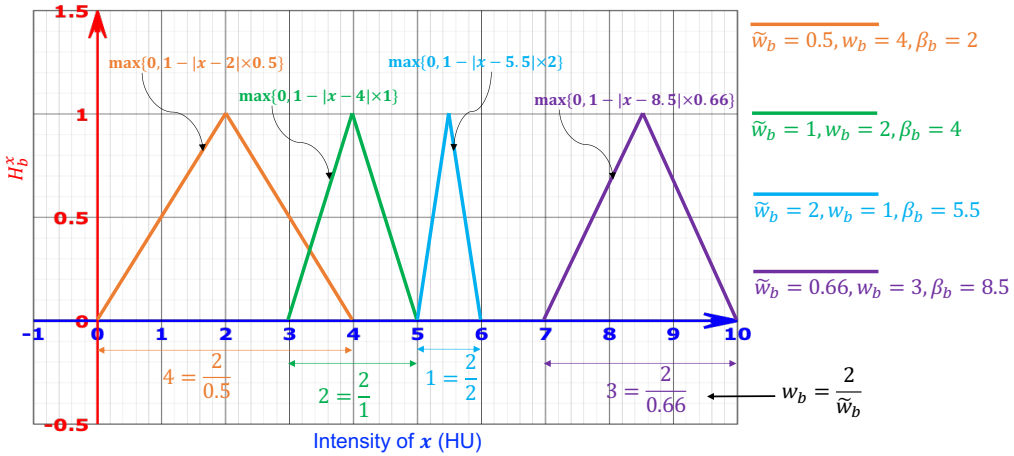


Figure 2: Schematic representation of our LIH bins. We show four arbitrary bins with different bin centers  $\beta_b$  and bin widths  $w_b$ .

In Fig. 2, we show a visual representation of four arbitrary LIH bins with different bin centers  $\beta_b$  and bin widths  $w_b$ . Also note that our LIH layer can be used at different depths of the network for multiple times, as well as can be applied directly on an input image or on a feature map.

## 2.2. Design of LIH using CNN Layers

We implement the LIH using conventional CNN layers, as illustrated in Fig. 1. The input of LIH can be a 2D or vectorized 1D image, and the output is a  $|\mathcal{B}| \times 1$  histogram feature vector. The operation  $x - \beta_b$  for a bin centered at  $\beta_b$  is equivalent to convolving the input by a  $1 \times 1$  kernel with fixed weight of 1 (i.e., with no updating by BP) and a learnable bias term  $\beta_b$  (‘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  $1 \times 1$  kernels and a fixed bias of 1 (i.e., no updating by BP) to model the operation of  $1 - |x - \beta_b| \times \tilde{w}_b$ . 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.

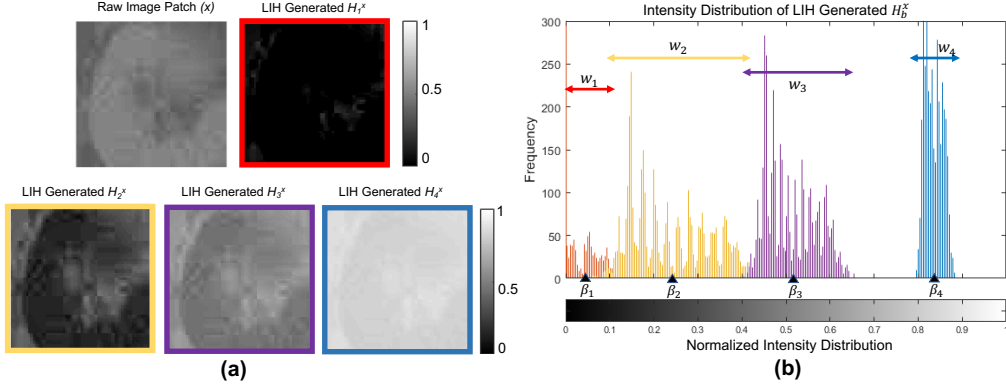


Figure 3: Illustration of LIH generated  $H_b^x$  with variable intensity distribution. (a) Raw CT image patch ( $x$ ) of size  $64 \times 64$  pixels and four randomly selected image patches ( $H_B^x$ ) before the global pooling in Fig. 1. (b) Corresponding intensity distributions of patches 1-4 in (a) are shown with Histogram of variable bin centers  $\beta_b$  and widths  $w_b$ .

We further clarify the specific way of using ‘Conv 2’ in our proposed LIH module. This convolution operator works differently than a typical convolution layer. Typically, a convolution kernel shares the learnable filter parameters among the channels. In contrast, the filter parameters of ‘Conv 2’ in our LIH are not shared among channels. Since the channels generated from ‘Conv 1’ correspond to different learnable bins  $B$  with different bin centers  $\beta_b$ , ‘Conv 2’ needs to operate on each of the channels  $B$  separately to associate different bin widths parameter  $\tilde{w}_b$ . That is why ‘Conv 2’ performs convolution on each channel separately.

In Fig. 3(a), we show an example raw CT image patch  $x$  and corresponding LIH generated image patches randomly selected from the feature maps of  $H_b(x)$  (Fig. 1). We also show the intensity distributions of the selected patches in Fig. 3(a) in terms of a histogram in Fig. 3(b), where we can observe the learned histogram of variable bin centers  $\beta_b$  and bin widths  $w_b$ . We also observe in Fig. 3(b) that the learned  $w_b$  for different feature maps in  $H_b(x)$  have overlaps among those.

### 2.3. 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 (Fig. 4). The first seven layers contain trainable weights. The input is a  $64 \times 64$  pixel image patch extracted

from the kidney+RCC slices. During training, we fed randomly shuffled image patches individually to the network. The LIH layer learns the variables  $\beta_b$  and  $\tilde{w}_b$  to extract characteristic 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  $4096 \times 1$ . The number of FCLs plays a vital role as the model’s overall depth is important for good performance (Zeiler and Fergus, 2014). 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.

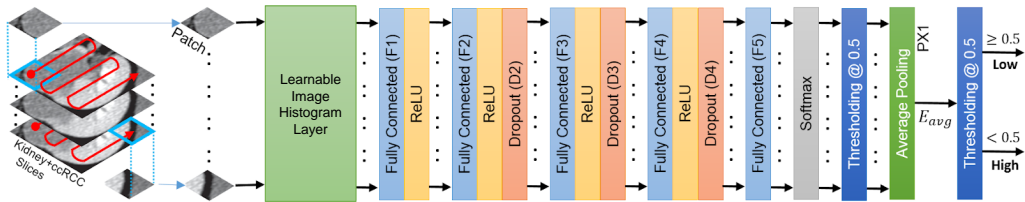


Figure 4: Multiple instance decisions aggregated ImHistNet for RCC grade and stage classification. The light green block represents the proposed LIH layer shown in Fig. 1.

#### 2.4. Training

We trained two separate ImHistNets for Fuhrman grading and anatomical staging of RCC. We implemented our networks in Caffe (Jia et al., 2014) command line environment and trained by minimizing the binary cross-entropy loss between the ground truth and predicted labels (1: Fuhrman low/stage low, and 0: Fuhrman high/stage high). As we mentioned in Section 2.2 that the ‘Conv 2’ operator in our LIH module needs to be applied on each channel separately, in our Caffe implementation, we used the ‘group’ strategy to facilitate separate and parallel convolution operations on each channel in the end-to-end learning. We set  $group = B$ . We used stochastic gradient descent for updating the parameters. We employed a Dropout unit (Dx) that drops 20%, 30%, and 40% of the units in F2, F3, and F4 layers, respectively (Fig. 4) 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. We ran the training for 250,000 iterations with a batch of 128 patches. The training ran on a workstation with an Intel 4.0 GHz Core-i7 processor, an Nvidia GeForce Titan Xp GPU with 12 GB of VRAM and 32 GB of RAM. Our workstation runs on Ubuntu 16.04 operating system with cuDNN version 9.0. Training the

ImHistNet took about 5 hours to reach an error saturation. The inference time per kidney sample is about 1 sec.

### 2.5. RCC Grade and 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 RCC cross-sections contained discriminant features for proper grade identification. Thus, our trained network may miss-classify the interrogated image patch. To reduce such miss classification, we adopt a similar multiple instance decision aggregation procedure to our earlier work (Hussain et al., 2018). In this approach, we feed randomly shuffled single image patches as inputs to the model during training. We feed all candidate image patches of a particular kidney to the trained network during inference 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 Fuhrman low or stage low, and  $E_{avg} < 0.5$  indicates Fuhrman high or stage high (see Fig. 4).

## 3. Data

We used CT scans of 159 patients from The Cancer Imaging Archive (TCIA) database (Clark et al., 2013). These patients' diagnosis was clear cell RCC, of which 64 belonged to Fuhrman low (I/II), and 95 belonged to Fuhrman high (III/IV). Also, 99 patients were staged low (I-II), and 60 were staged high (III-IV) in the same cohort. The images in this database have variations in CT scanner models and spatial resolution. We divided the dataset for training/validation/testing as 44/5/15 and 75/5/15 for Fuhrman low and Fuhrman high, respectively. For anatomical staging, we divided the dataset for training/validation/testing as 81/3/15 and 42/3/15 for stage low and stage high, respectively. This database does not specify the time delay between the contrast media administration and acquisition of the image. Therefore, we cannot distinguish a CT volume in terms of the cortico-medullary and nephrographic phase. We show a summary of the data we used in this work in Table 2:

Items	Descriptions
Modality	CT
Pixel Dimensions	Axial: 1.5 ~ 7.5 mm Coronal: 0.29 ~ 1.87 mm Sagittal: 0.29 ~ 1.87 mm
Total Patients	159
Number of Males	106
Number of Females	53
Age	Mean: 61.18±12.07 Y Minimum Age: 34 Y Maximum Age: 89 Y
Race	White: 155 Black or African American: 4

Table 2: Summary of relevant and available information about the CT data used in this work.

Our method’s input data are 2D image patches of size  $64 \times 64$  pixels, taken from a region-of-interest that contains both the kidney and RCC. We do not require any fine pre-segmentation of the RCC. Our LIH layer can effectively stratify the overall image intensity range into a predefined number of bins (i.e.,  $B$ ). It facilitates ignoring background/unwanted intensity values from the intensity of interest (i.e., RCC and kidney intensity). Therefore, our method can effectively learn RCC textural features from a non-delineated kidney and RCC. It is also essential to check if an RCC extends up to Gerota’s fascia, as it indicates higher stages of RCC. As our approach does not use pre-segmentation of the kidney, rather uses a loose and wide enough ROI around the kidney and RCC, it facilitates the inclusion of surrounding areas, including parts of Gerota’s fascia, into the analysis. Given the data imbalance where samples for Fuhrman low are fewer than for Fuhrman high and stage high are fewer than for stage low, we allowed more overlap among adjacent patches for the Fuhrman low and staged high datasets. We calculate the amount of overlap to balance the samples from both cohorts.

#### 4. Results and Discussion

First, we present the visual representation of learned bin centers and

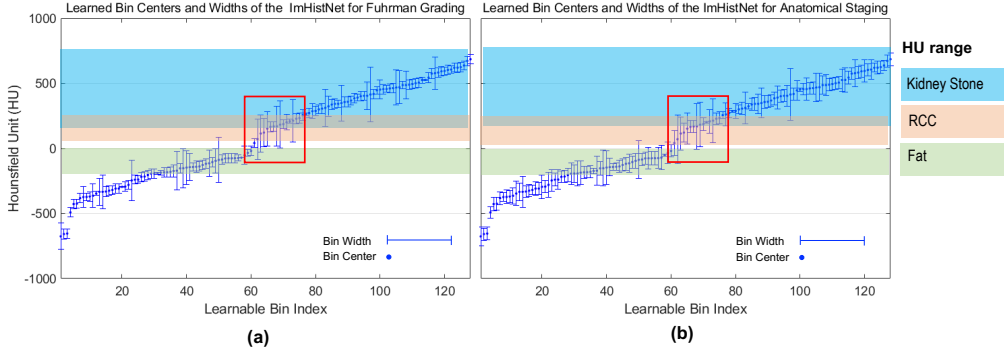


Figure 5: Illustration of the learned bin centers and widths by the proposed ImHistNet for (a) RCC grading and (b) RCC staging. The bin centers are plotted after sorting from low to high values of the CT Hounsfield Unit. The red boxes in both figures indicate the Hounsfield Unit spectrum of RCC, where high variation in bin widths among the neighboring bins are observed.

widths by the proposed ImHistNet for RCC grading and staging in Fig. 5. Our CT data was normalized between  $[-1000, 3000]$  Hounsfield Unit (HU), and we did not use any windowing of CT data in terms of HU values. From the plots in Fig. 5, we see that our ImHistNet concentrates roughly between  $[-500, 760]$  HU, the range that covers the fatty tissue  $[-20, -150]$  HU (Kim et al., 1999), kidney  $[20, 45]$  HU (Lepor, 2000), RCC  $[30, 150]$  HU (Ching et al., 2017), and kidney calcification  $[110, 760]$  HU (Shahnani et al., 2014). It indicates that the ImHistNet was successful in identifying the HU range related to kidney diseases. These figures also indicate that ImHistNet does not require any fine pre-segmentation of the RCC. It is clear from Fig. 5(a) and (b) that our LIH layer can effectively focus on a certain image intensity range. If we further observe the pattern of learned bin width variations in the close neighborhood (i.e.,  $\sim 20$  neighboring bins) in both Fig. 5(a) and (b), we see that the learned bin widths are more irregular among neighboring bins in the region that falls within the RCC HU spectrum (shown with red boxes in Fig. 5(a) and (b)). It indicates that the ImHistNet concentrated more on analyzing the textural variations of the RCC for the grading and staging tasks.

#### 4.1. RCC Fuhrman Grade Classification

We compared our RCC grade classification performance in terms of accuracy to a wide range of methods. Note that we trained models with shuffled

single image patches for all our implementations and used multiple instance decision aggregation per kidney during inference. We fixed our patch size to  $64 \times 64$  pixels across all contrasting methods.

To provide a rigorous justification for choosing different hyper-parameters of our proposed model, we evaluated the effects of our model’s different components incrementally in Table 3. First, we use ResNet-50 (He et al., 2016) with transfer learning to test the performance of conventional CNN (Table 3). Here, we used the full kidney+RCC slices as input in experiment 1 and patches as input in experiment 2. As we mentioned in Sect. 1 that a classical CNN typically gives less emphasis in capturing textural features, it has become evident from the results in Table 3 where such CNNs performed poorly in learning the textural features of RCC.

Next, to evaluate the performance of hand-engineered feature-based conventional machine learning approaches, we tested a support vector machine (SVM) employing the conventional image histogram of 128 and 256 bins, as shown in Table 3 experiments 3 and 4, respectively. We also compared two state-of-the-art methods (Shu et al., 2018; Meng et al., 2017), denoted as experiments 5 and 6, where we quote authors’ best self-reported performances. These methods mostly relied on the RCC textural features and used classical predictive models, e.g., logistic regression. Here, the method by Shu et al. (2018) performed the best with 77% classification accuracy (Table 3 experiment 5).



Exp.	Aspects Checked	Methods	NTS	Acc.
<i>Conventional CNN</i>				
1	Full image <i>vs.</i> Patch	Full image+ResNet-50 (He et al., 2016)	30	53.33%
2		Patch+ResNet-50 (He et al., 2016)	30	50.00%
<i>Hand-engineered Features-based Conventional Machine Learning</i>				
3	Number of	Patch+Histogram (128 bins)+SVM	30	56.66%
4	Histogram-bins	Patch+Histogram (256 bins)+SVM	30	63.33%
5	State-of-the-art	Shu et al. (2018) (5FCV on 260 samples)	-	77.00%
6		Meng et al. (2017) (L1OCV on 90 samples)	-	70.00%
<i>Hand-engineered Features-based Deep Neural Networks</i>				
7	Number of	Patch+Histogram (128 bins)+5 FCL	30	50.00%
8	Histogram-bins	Patch+Histogram (256 bins)+5 FCL	30	50.00%
<i>LIH Features-based Conventional Machine Learning</i>				
9	-	Patch+LIH (128 bins)+AP+SVM	30	60.00%
<i>Combined LIH Features-based DNN and Conventional CNN</i>				
10	Full image <i>vs.</i> Patch	Patch+LIH+AP+5 FCL    AlexNet	30	53.33%
11		Full Image+LIH+AP+5 FCL    AlexNet	30	50.00%
<i>Deep Neural Network</i>				
12	Absence of Histogram	Patch+5 FCL	30	36.67%
<i>LIH Features-based Deep Neural Networks</i>				
13	Number of FCLs, Number of Bins, Pooling Types	Patch+LIH (128 bins)+NVEC+5 FCL	30	50.00%
14		Patch+LIH (128 bins)+AP+4 FCL	30	53.33%
15		Patch+LIH (128 bins)+AP+6 FCL	30	50.00%
16		Patch+LIH (64 bins)+AP+5 FCL	30	50.00%
17		Patch+LIH (256 bins)+AP+ 5 FCL	30	43.33%
18	<b>Proposed</b>	ImHistNet [LIH (128 bins)+AP+5 FCL]	30	<b>80.00%</b>

Table 3: Comparison of automatic RCC Fuhrman grade classification performance by different methods. Acronyms used- Exp.: Experiment, NTS: Number of test samples, Acc.: Accuracy, SVM: support vector machines, xFCV: x-fold cross-validation, LxOCV: leave-x-out cross-validation, ‘-’: Not reported.

Then, we examine the performance of hand-engineered features with deep neural network (DNN) and the LIH features with SVM in experiments 7-9 in Table 3. To contrast the performance of an SVM against a DNN, in experiments 7 and 8, we fed the conventional histogram (128 and 256 bins) features to a DNN of 5 FCL with weight sizes  $(4096 \times 1)$ - $(4096 \times 1)$ - $(4096 \times 1)$ - $(4096 \times 1)$ - $(2 \times 1)$ . We choose this FCL configuration as our ImHistNet contains the same. Next, to evaluate the hand-engineered features against LIH features, we used LIH features to train an SVM in experiment 9. We see in Table 3 that the SVM with LIH features outperformed the SVM with conventional histogram features.

To evaluate the performance of a DNN, combining a conventional CNN and ImHistNet, we added AlexNet (Krizhevsky et al., 2012) in parallel to the ImHistNet in experiments 10 and 11. We concatenated the last FCLs of size  $4096 \times 1$  in both networks, and the whole network was trained end-to-end. We implemented two such approaches using the full kidney+RCC images and patches as inputs in experiments 10 and 11, respectively. To use patches as input to the AlexNet, we upsampled those to a size of  $227 \times 227$  pixels. We observed in Table 3 that the classical CNN affect the performance of the proposed ImHistNet negatively, i.e., results were worse than those by ImHistNet.

In experiment 12, we also did an ablation study to check the performance of ImHistNet without the LIH layer. The resulting DNN consists of only 5 FCLs. Then, we fed image patches directly to FCLs, without any hand-engineered or LIH-generated intermediate features. We see from Table 3 that the accuracy of this approach is the worst among all comparing techniques. Thus, it is clear that our LIH layer learns discriminatory textural features.

Finally, to achieve the optimum results from LIH, we varied the number of bins (64/128/256) and FCLs of size  $4096 \times 1$  (4/5/6), and the pooling types (AP/NZEC) with the LIH layer and present the results as experiments 13-18 in Table 3. We see that ImHistNet with 128 bins, average pooling, and 5 FCL achieved the highest accuracy (80%) among all contrasting methods shown in Table 3. The corresponding confusion matrix is shown in Fig. 6. The method of Shu et al. (2018) showed the closest performance to ImHistNet with 77% accuracy (see experiment 5 in Table 3). The estimated RCC grading accuracy of the ImHistNet on the training data is 82%.

		Actual Grade	
		Fuhrman I-II	Fuhrman III-IV
Predicted Grade	Fuhrman I-II	12	3
	Fuhrman III-IV	3	12

Figure 6: Confusion matrix showing the actual and predicted Fuhrman grades for the test data.

#### 4.2. RCC Stage Classification

We also compared our RCC stage classification performance in terms of accuracy to a wide range of methods in Table 4. To our knowledge, there is no automatic and machine learning-based approach for RCC stage classification. Therefore, we compare the RCC staging performance of different methods by implementing those in our capacity. Similar to RCC grade classification, we trained models with shuffled single image patches and used multiple instance decision aggregation per kidney during inference. We fixed our patch size to  $64 \times 64$  pixels across all contrasting methods except for ResNet-50.

Exp.	Aspects Checked	Methods	NTS	Acc.
<i>Hand-engineered Features-based Conventional Machine Learning</i>				
19	Number of	Patch+Histogram (16 bins)+SVM	30	53.33%
20	Histogram-bins	Patch+Histogram (64 bins)+SVM	30	53.33%
<i>Hand-engineered Features-based Conventional Machine Learning</i>				
21	Number of	Patch+Histogram (16 bins)+5 FCL	30	50.00%
22	Histogram-bins	Patch+Histogram (64 bins)+5 FCL	30	50.00%
<i>Conventional CNN</i>				
23	-	Full Image+ResNet-50 (He et al., 2016)	30	60.00%
<i>Deep Neural Network</i>				
24	Absence of Histogram	Patch+5 FCL	30	43.33%
<i>LIH Features-based Deep Neural Networks</i>				
25	<b>Proposed</b>	ImHistNet [LIH (128 bins)+AP+5 FCL]	30	<b>83.33%</b>

Table 4: Comparison of automatic RCC stage classification performance by different methods. Acronyms used- Exp.: Experiment, NTS: Number of test samples, Acc.: Accuracy, SVM: support vector machines.

First, to compare the performance of ImHistNet to that of traditional hand-engineered feature-based machine learning approaches, we evaluated an SVM as experiments 19 and 20 employing a conventional image histogram of 16 and 64 bins, respectively, and Table 4 shows a resulting poor performance at 53% accuracy for both the cases.

Next, to contrast the performance of SVM against DNN, in experiments 21 and 22, we fed the conventional histogram (16 and 64 bins, respectively) features to a DNN of 5 FCL with weight sizes  $(4096 \times 1)$ - $(4096 \times 1)$ - $(4096 \times 1)$ - $(4096 \times 1)$ - $(2 \times 1)$ . We chose this FCL configuration for fairer comparisons

since our ImHistNet contains the same. Table 4 shows that the FCL with conventional histogram performed worse, achieving 50% accuracy.

Then we used ResNet-50 (He et al., 2016) in experiment 23 with transfer learning to test the performance of high performing modern CNN (see Table 4). We used full kidney+RCC slices of size  $224 \times 224$  pixels as input. As mentioned in Sect. 1, a classical CNN typically gives less emphasis 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.

We also did an ablation study for RCC staging similar to experiment 12 to check the performance of ImHistNet without the LIH layer. In experiment 24, we fed image patches directly to FCLs, without any hand-engineered or LIH-generated intermediate features. We see from Table 4 that the accuracy of this approach is the worst among all comparing techniques. Thus, it is clear from here too that our LIH layer learns discriminatory textural features.

Finally, we show our proposed method’s performance in Table 4 experiment 25, where ImHistNet achieved the highest accuracy (83%) among all contrasted methods. The corresponding confusion matrix is shown in Fig. 7. The estimated RCC staging accuracy of the ImHistNet on the training data is 87%.

		Actual Stage	
		Anatomical Stage I-II	Anatomical Stage III-IV
Predicted Stage	Anatomical Stage I-II	13	3
	Anatomical Stage III-IV	2	12

Figure 7: Confusion matrix showing the actual and predicted anatomical stages for the test data.

### 4.3. Discussion

In this work, we achieved over 80% accuracy in RCC grade/stage low and high classification using a straightforward but robust DNN architecture. The proposed learnable image histogram layer in conjugation with five fully

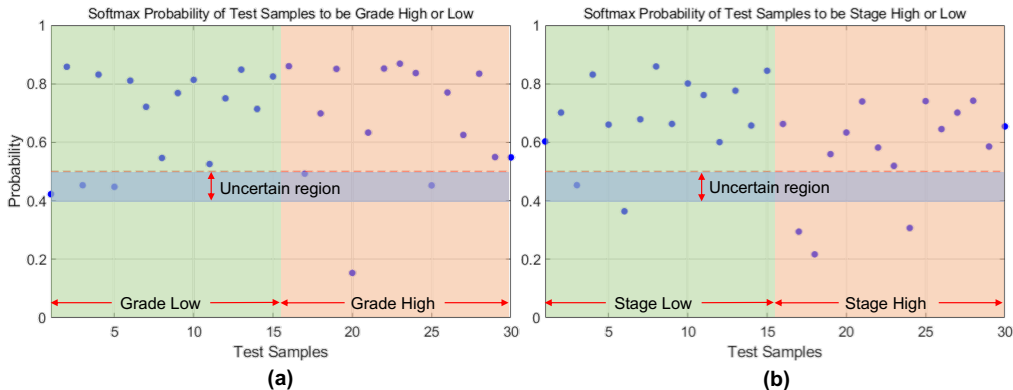


Figure 8: Illustration of the Softmax probability of each test case used in RCC (a) grading and (b) staging. The green-shaded regions represent grade/stage low and red-shaded regions represent grade/stage high. We also show blue-shaded regions, representing a probability range 0.41–0.50, as an uncertain decision region.

connected layers outperformed the state-of-the-art computer vision image classifier, i.e., ResNet (He et al., 2016). The proposed ImHistNet is also efficient in segregating different textures in a CT image that it easily stratify the tumor/cancer textures from the background. Thus our method does not require any segmentation of the RCC. Besides, to our knowledge, for the first time, we have shown that the RCC stages can also be estimated automatically by analyzing the CT textural features of RCC.

We faced a big challenge in this work because of the small size of our experimental dataset, especially in the RCC staging experiment. In our 159 patient cohort, the number of RCC stage I, II, III, and IV patients are 84, 15, 41, and 19, respectively. These numbers are highly imbalanced and insufficient for some cases to train a deep neural network properly. Thus, we choose to group them into stage low (I/II) and stage high (III/IV) as the nature of treatment is often similar between stages I and II, and between stages III and IV (Escudier et al., 2016; Bradley et al., 2015). Furthermore, we used small image patches to increase the number and variability of training samples and address class imbalances in the training data via controlling the overlap between adjacent patches.

Although we compared RCC grading performance of the ImHistNet with conventional machine learning approaches by Shu et al. (2018) and Meng et al. (2017), there are two neural network-based RCC grading approaches (Kocak et al., 2019; He et al., 2020) present in the literature. However, these

methods used hand-engineered RCC features and subsequently used simple neural networks to reduce the feature dimensionality. In addition, the method by Kocak et al. (2019) is designed for unenhanced CT, while the method by He et al. (2020) is designed for the WHO renal tumor classification system. Therefore, the performance of these methods cannot be compared directly to our approach.

On the other hand, we found only one CT feature-based RCC staging approach (Ökmen et al., 2019) that used K-nearest neighbors. However, this method is designed for TNM I-IV staging, while our approach is designed for low and high anatomical staging. Thus, our performance cannot be compared directly to that by Ökmen et al. (2019).

The proposed ImHistNet also produced several miss classifications in both grading and staging tasks. To examine those miss classification cases, we plot the Softmax probability of each test case used in RCC grading and staging experiments in Fig. 8(a) and (b), respectively. We defined the probability range 0.41–0.50 as an uncertain decision region, where our method fails to decide with certainty. We see in Fig. 8(a) that the probabilities associated with all three miss-classification of actual grade low cases fall within the uncertain region. Two of the three probabilities associated with the three miss classification of actual grade high cases fall within this uncertain region. Thus, our method lacked confidence in a total of five out of six miss classification test cases in RCC grading. In the RCC staging task, we see in Fig. 8(b) that only one of the two probabilities associated with the miss classification of actual stage low cases fall within the uncertain region.

Image-based RCC grading and staging has a promising clinical implication. Although biopsy-based RCC grading is an inseparable part of the clinical workflow, it often requires considerable time in the process of performing the biopsy and subsequent radiological analysis. It is also reported that percutaneous imaging-guided renal fine-needle aspiration suffered from low sensitivity and frequent nondiagnostic results (Maturen et al., 2007). A recent study (Patel et al., 2016) of core biopsy on 2,979 patients found a notable Fuhrman upgrading (16%) from low to high grade after surgical resection of the renal mass. Since our image-based noninvasive approach covers the full tumor region, it is less susceptible to misdiagnosis. Therefore, our approach can be useful in clinical decision support. Besides, while a patient waits for the biopsy conduction and results, an image-based approach can help physicians diagnose and prepare the treatment plan. The biopsy results can confirm the decision.

## 5. Conclusion

We proposed a learnable image histogram-based DNN framework for end-to-end image classification. We demonstrated our approach on a cancer grade and stage prediction task providing automatic 2-tiered FGS (Fuhrman low and Fuhrman high) grade classification and stage low and stage high classification of RCC from CT scans. Our approach learns a histogram directly from the image data and deploys it to extract representative discriminant textural image features. We increased efficacy by using small image patches to increase the number and variability of training samples and address class imbalances in the training data via overlap control. We also used multiple instance decision aggregation to robustify binary classification further. Our proposed ImHistNet outperformed current competing approaches for this task, including conventional ML, deep learning, as well as manual human radiology experts. ImHistNet appears well suited for radiomic studies, where learned textural features from the learnable image histogram may aid in better diagnosis.

Our proposed ImHistNet efficiently stratifies the intensity spectrum into learnable bins. We plan to investigate a process to incorporate the spatial texture context via learning the co-occurrence statistics within the DNN framework. It is also essential to find an optimal architecture of ImHistNet that can lead to better RCC analysis performance. That being said, the problem of comprehensive exploration of possible network architectures is beyond the scope of this work. Our future work envisions to include neural architecture search (NAS) to find the optimal network for this work.

Furthermore, to learn domain-invariant RCC features, we plan to use the classification and contrastive semantic alignment (CCSA) loss Yoon et al. (2019). This approach would facilitate learning RCC-specific features irrespective of the data source. Thus, it would avoid a deep model being negatively affected by the domain-specific features. It is also a widely accepted fact that pathological data are scarce compared to healthy data. Another possible future direction is to address this imbalance in training data by synthesizing pathological RCC cases, as previously reported for lung nodules in Han et al. (2019) and brain tumour synthesis in Yu et al. (2018). Synthesized pathological cases may enhance the result of cross-domain training testing.

## Acknowledgment

We thank Nvidia Corporation for supporting our research through their GPU Grant Program by donating the GeForce Titan Xp. We also thank anonymous reviewers for their feedback that helped in improving the paper.

## References

- AAIAbdulsalam, A.K., Garvin, J.H., Redd, A., Carter, M.E., Sweeny, C., Meystre, S.M., 2018. Automated extraction and classification of cancer stage mentions from unstructured text fields in a central cancer registry. *AMIA Summits on Translational Science Proceedings* 2018, 16.
- Andrearczyk, V., Whelan, P.F., 2016. Using filter banks in convolutional neural networks for texture classification. *Pattern Recognition Letters* 84, 63–69.
- Becker, A., Hickmann, D., Hansen, J., Meyer, C., Rink, M., Schmid, M., Eichelberg, C., Strini, K., Chromecki, T., Jesche, J., et al., 2016. Critical analysis of a simplified fuhrman grading scheme for prediction of cancer specific mortality in patients with clear cell renal cell carcinoma—impact on prognosis. *European Journal of Surgical Oncology (EJSO)* 42, 419–425.
- Bektas, C.T., Kocak, B., Yardimci, A.H., Turkcanoglu, M.H., Yucetas, U., Koca, S.B., Erdim, C., Kilickesmez, O., 2019. Clear cell renal cell carcinoma: machine learning-based quantitative computed tomography texture analysis for prediction of Fuhrman nuclear grade. *European radiology* 29, 1153–1163.
- Bradley, A., MacDonald, L., Whiteside, S., Johnson, R., Ramani, V., 2015. Accuracy of preoperative CT T staging of renal cell carcinoma: which features predict advanced stage? *Clinical Radiology* 70, 822–829.
- Chen, S., Zhang, N., Jiang, L., Gao, F., Shao, J., Wang, T., Zhang, E., Yu, H., Wang, X., Zheng, J., 2020. Clinical use of a machine learning histopathological image signature in diagnosis and survival prediction of clear cell renal cell carcinoma. *International Journal of Cancer* .
- Ching, B.C., Tan, H.S., Tan, P.H., Toh, C.K., Kanesvaran, R., Ng, Q.S., Tan, M.H., 2017. Differential radiologic characteristics of renal tumours on multiphasic computed tomography. *Singapore medical journal* 58, 262.



- Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., , et al., 2013. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of Digital Imaging* 26, 1045–1057.
- Coy, H., Hsieh, K., Wu, W., Nagarajan, M.B., Young, J.R., Douek, M.L., Brown, M.S., Scalzo, F., Raman, S.S., 2019. Deep learning and radiomics: the utility of Google TensorFlow™ inception in classifying clear cell renal cell carcinoma and oncocytoma on multiphasic CT. *Abdominal Radiology* 44, 2009–2020.
- Delahunt, B., Cheville, J.C., Martignoni, G., Humphrey, P.A., Magi-Galluzzi, C., McKenney, J., Egevad, L., Algaba, F., Moch, H., Grignon, D.J., et al., 2013. The international society of urological pathology (ISUP) grading system for renal cell carcinoma and other prognostic parameters. *The American journal of surgical pathology* 37, 1490–1504.
- Deng, Y., Soule, E., Samuel, A., Shah, S., Cui, E., Asare-Sawiri, M., Sundaram, C., Lall, C., Sandrasegaran, K., 2019. CT texture analysis in the differentiation of major renal cell carcinoma subtypes and correlation with Fuhrman grade. *European Radiology* 29, 6922–6929.
- Ding, J., Xing, Z., Jiang, Z., Chen, J., Pan, L., Qiu, J., Xing, W., 2018. CT-based radiomic model predicts high grade of clear cell renal cell carcinoma. *European Journal of Radiology* 103, 51–56.
- Escudier, B., Porta, C., Schmidinger, M., Rioux-Leclercq, N., Bex, A., Khoo, V., Gruenvald, V., Horwich, A., 2016. Renal cell carcinoma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology* 27, v58–v68.
- Feng, Z., Shen, Q., Li, Y., Hu, Z., 2019. CT texture analysis: a potential tool for predicting the fuhrman grade of clear-cell renal carcinoma. *Cancer Imaging* 19, 6.
- Fuhrman, S.A., Lasky, L.C., Limas, C., 1982. Prognostic significance of morphologic parameters in renal cell carcinoma. *The American Journal of Surgical Pathology* 6, 655–663.

- Haji-Momenian, S., Lin, Z., Patel, B., Law, N., Michalak, A., Nayak, A., Earls, J., Loew, M., 2020. Texture analysis and machine learning algorithms accurately predict histologic grade in small ( $< 4$  cm) clear cell renal cell carcinomas: a pilot study. *Abdominal Radiology* 45, 789–798.
- Han, C., Kitamura, Y., Kudo, A., Ichinose, A., Rundo, L., Furukawa, Y., Umemoto, K., Li, Y., Nakayama, H., 2019. Synthesizing diverse lung nodules wherever massively: 3D multi-conditional GAN-based CT image augmentation for object detection, in: 2019 International Conference on 3D Vision (3DV), IEEE. pp. 729–737.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
- He, X., Wei, Y., Zhang, H., Zhang, T., Yuan, F., Huang, Z., Han, F., Song, B., 2020. Grading of clear cell renal cell carcinomas by using machine learning based on artificial neural networks and radiomic signatures extracted from multidetector computed tomography images. *Academic Radiology* 27, 157–168.
- Huhdanpaa, H., Hwang, D., Cen, S., Quinn, B., Nayyar, M., Zhang, X., Chen, F., Desai, B., Liang, G., Gill, I., et al., 2015. CT prediction of the Fuhrman grade of clear cell renal cell carcinoma (RCC): towards the development of computer-assisted diagnostic method. *Abdominal Imaging* 40, 3168–3174.
- Hussain, M.A., Hamarneh, G., Garbi, R., 2018. 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. pp. 657–665.
- Hussain, M.A., Hamarneh, G., Garbi, R., 2019a. 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. pp. 130–138.
- Hussain, M.A., Hamarneh, G., Garbi, R., 2019b. Renal cell carcinoma staging with learnable image histogram-based deep neural network, in: Interna-

- tional Workshop on Machine Learning in Medical Imaging, Springer. pp. 533–540.
- Ishigami, K., Leite, L.V., Pakalniskis, M.G., Lee, D.K., Holanda, D.G., Kuehn, D.M., 2014. Tumor grade of clear cell renal cell carcinoma assessed by contrast-enhanced computed tomography. *SpringerPlus* 3, 694.
- Janssen, M., Linxweiler, J., Terwey, S., Ruge, S., Ohlmann, C.H., Becker, F., Thomas, C., Neisius, A., Thüroff, J., Siemer, S., et al., 2018. Survival outcomes in patients with large ( $\geq 7$ cm) 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, e0196427.
- Jeldres, C., Sun, M., Liberman, D., Lughezzani, G., De La Taille, A., Tostain, J., Valeri, A., Cindolo, L., Ficarra, V., Artibani, W., et al., 2009. Can renal mass biopsy assessment of tumor grade be safely substituted for by a predictive model? *The Journal of urology* 182, 2585–2589.
- Jeon, H.G., Seo, S.I., Jeong, B.C., Jeon, S.S., Lee, H.M., Choi, H.Y., Song, C., Hong, J.H., Kim, C.S., Ahn, H., et al., 2016. Percutaneous kidney biopsy for a small renal mass: a critical appraisal of results. *The Journal of urology* 195, 568–573.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093* .
- Kim, S., Lee, G., Lee, S., Park, S., Pyo, H., Cho, J., 1999. Body fat measurement in computed tomography image. *Biomedical sciences instrumentation* 35, 303–308.
- Kocak, B., Durmaz, E.S., Ates, E., Kaya, O.K., Kilickesmez, O., 2019. Unenhanced CT texture analysis of clear cell renal cell carcinomas: a machine learning-based study for predicting histopathologic nuclear grade. *American Journal of Roentgenology* 212, W132–W139.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: *Advances in Neural Information Processing Systems*, pp. 1097–1105.

- Kuthi, L., Jenei, A., Hajdu, A., Németh, I., Varga, Z., Bajory, Z., Pajor, L., Iványi, B., 2017. Prognostic factors for renal cell carcinoma subtypes diagnosed according to the 2016 WHO renal tumor classification: a study involving 928 patients. *Pathology & Oncology Research* 23, 689–698.
- Lane, B.R., Babineau, D., Kattan, M.W., Novick, A.C., Gill, I.S., Zhou, M., Weight, C.J., Campbell, S.C., 2007. A preoperative prognostic nomogram for solid enhancing renal tumors 7 cm or less amenable to partial nephrectomy. *The Journal of urology* 178, 429–434.
- LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521, 436.
- Lepor, H., 2000. Prostatic diseases. volume 2000. WB Saunders Company.
- Lin, F., Cui, E.M., Lei, Y., Luo, L.p., 2019. CT-based machine learning model to predict the fuhrman nuclear grade of clear cell renal cell carcinoma. *Abdominal Radiology* 44, 2528–2534.
- Maturen, K.E., Nghiem, H.V., Caoili, E.M., Higgins, E.G., Wolf Jr, J.S., Wood Jr, D.P., 2007. Renal mass core biopsy: accuracy and impact on clinical management. *American Journal of Roentgenology* 188, 563–570.
- Meng, F., Li, X., Zhou, G., Wang, Y., 2017. Fuhrman grade classification of clear-cell renal cell carcinoma using computed tomography image analysis. *Journal of Medical Imaging and Health Informatics* 7, 1671–1676.
- van der Mijn, J.C., Al Awamlh, B.A.H., Khan, A.I., Posada-Calderon, L., Oromendia, C., Fainberg, J., Alshak, M., Elahjji, R., Pierce, H., Taylor, B., et al., 2019. Validation of risk factors for recurrence of renal cell carcinoma: Results from a large single-institution series. *PloS one* 14.
- Nazari, M., Shiri, I., Hajianfar, G., Oveisi, N., Abdollahi, H., Deevband, M.R., Oveisi, M., Zaidi, H., 2020. Noninvasive Fuhrman grading of clear cell renal cell carcinoma using computed tomography radiomic features and machine learning. *La radiologia medica* , 1–9.
- Oh, S., Sung, D.J., Yang, K.S., Sim, K.C., Han, N.Y., Park, B.J., Kim, M.J., Cho, S.B., 2017. Correlation of CT imaging features and tumor size with fuhrman grade of clear cell renal cell carcinoma. *Acta Radiologica* 58, 376–384.

- Ökmen, H.B., Uysal, H., Guvenis, A., 2019. Automated prediction of TNM stage for clear cell renal cell carcinoma disease by analyzing CT images of primary tumors, in: 2019 11th International Conference on Electrical and Electronics Engineering (ELECO), IEEE. pp. 456–459.
- Patel, H.D., Johnson, M.H., Pierorazio, P.M., Sozio, S.M., Sharma, R., Iyoha, E., Bass, E.B., Allaf, M.E., 2016. Diagnostic accuracy and risks of biopsy in the diagnosis of a renal mass suspicious for localized renal cell carcinoma: systematic review of the literature. *The Journal of urology* 195, 1340–1347.
- Sasaguri, K., Takahashi, N., 2018. CT and MR imaging for solid renal mass characterization. *European Journal of Radiology* 99, 40–54.
- Schieda, N., Nguyen, K., Thornhill, R.E., McInnes, M.D., Wu, M., James, N., 2020. Importance of phase enhancement for machine learning classification of solid renal masses using texture analysis features at multi-phasic CT. *Abdominal Radiology* 45, 2786–2796.
- 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., 2017. 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, 289–296.
- Shahnani, P.S., Karami, M., Astane, B., Janghorbani, M., 2014. The comparative survey of hounsfield units of stone composition in urolithiasis patients. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences* 19, 650.
- Shu, J., Tang, Y., Cui, J., Yang, R., Meng, X., Cai, Z., Zhang, J., Xu, W., Wen, D., Yin, H., 2018. Clear cell renal cell carcinoma: CT-based radiomics features for the prediction of fuhrman grade. *European Journal of Radiology* 109, 8–12.
- Shu, J., Wen, D., Xi, Y., Xia, Y., Cai, Z., Xu, W., Meng, X., Liu, B., Yin, H., 2019. Clear cell renal cell carcinoma: Machine learning-based computed tomography radiomics analysis for the prediction of WHO/ISUP grade. *European journal of radiology* 121, 108738.

- Sun, X., Liu, L., Xu, K., Li, W., Huo, Z., Liu, H., Shen, T., Pan, F., Jiang, Y., Zhang, M., 2019. Prediction of ISUP grading of clear cell renal cell carcinoma using support vector machine model based on CT images. *Medicine* 98.
- Tan, H.J., Norton, E.C., Ye, Z., Hafez, K.S., Gore, J.L., Miller, D.C., 2012. Long-term survival following partial vs radical nephrectomy among older patients with early-stage kidney cancer. *Jama* 307, 1629–1635.
- Tian, K., Rubadue, C.A., Lin, D.I., Veta, M., Pyle, M.E., Irshad, H., Heng, Y.J., 2019. Automated clear cell renal carcinoma grade classification with prognostic significance. *PloS one* 14, e0222641.
- Wang, Z., Li, H., Ouyang, W., Wang, X., 2016. Learnable histogram: Statistical context features for deep neural networks, in: *European Conference on Computer Vision*, Springer. pp. 246–262.
- Yan, L., Chai, N., Bao, Y., Ge, Y., Cheng, Q., 2020. Enhanced computed tomography-based radiomics signature combined with clinical features in evaluating nuclear grading of renal clear cell carcinoma. *Journal of computer assisted tomography* 44, 730–736.
- Yap, F.Y., Varghese, B.A., Cen, S.Y., Hwang, D.H., Lei, X., Desai, B., Lau, C., Yang, L.L., Fullenkamp, A.J., Hajian, S., et al., 2020. Shape and texture-based radiomics signature on CT effectively discriminates benign from malignant renal masses. *European Radiology* , 1–11.
- Yoon, C., Hamarneh, G., Garbi, R., 2019. Generalizable feature learning in the presence of data bias and domain class imbalance with application to skin lesion classification, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 365–373.
- Yu, B., Zhou, L., Wang, L., Fripp, J., Bourgeat, P., 2018. 3D cGAN based cross-modality MR image synthesis for brain tumor segmentation, in: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE. pp. 626–630.
- Yu, H., Scalera, J., Khalid, M., Touret, A.S., Bloch, N., Li, B., Qureshi, M.M., Soto, J.A., Anderson, S.W., 2017. Texture analysis as a radiomic marker for differentiating renal tumors. *Abdominal Radiology* 42, 2470–2478.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer. pp. 818–833.