ImHistNet: Learnable Image Histogram Based DNN with Application to Noninvasive Determination of Carcinoma Grades in CT Scans

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. Clear cell RCC (ccRCC) is the major subtype of RCC and its biological aggressiveness affects prognosis and treatment planning. An important ccRCC prognostic predictor is its 'grade' for which the 4-tiered Fuhrman grading system is used. Although the Fuhrman grade can be identified by percutaneous renal biopsy, recent studies suggested that such grades may be non-invasively identified by studying image texture features of the ccRCC from computed tomography (CT) data. Such image feature based identification currently mostly relies on laborious manual processes based on visual inspection of 2D image slices that are time-consuming and subjective. In this paper, we propose a learnable image histogram based deep neural network approach that can perform the Fuhrman low (I/II) and high (III/IV) grade classification for ccRCC in CT scans. Validated on a clinical CT dataset of 159 patients from the TCIA database, our method classified ccRCC low and high grades with 80% accuracy and 85% AUC.

1 Introduction

Renal cell carcinoma (RCC) is the seventh most common cancer accounting for an estimated 140,000 global deaths annually [1]. Clear cell RCC (ccRCC) accounts for approximately 80% of RCC [2] and its biological aggressiveness affects the prognosis and treatment planning [3]. The 'grade' of a ccRCC is one of the important prognostic predictors of 5-year survival where higher grade tumors have an elevated risk of postoperative recurrence [2]. Although the 4-tiered Fuhrman grading system (FGS) [4] is used for ccRCC grading, in current clinical practice, a simplified 2-tiered FGS that reduces variability and improves reproducibility of the tumor grade is preferred by pathologists [1,2,3]. The 2-tier FGS, which divides grades to low grade (Fuhrman I/II) and high grade (Fuhrman III/IV), was shown to be as effective as 4-tiered FGS in predicting cancer-specific mortality in a study population of 2,415 ccRCC patients [5].

Clinically, invasive percutaneous renal biopsy is currently used for ccRCC FGS [1]. However, inter-observer reproducibility of grades assigned by pathologists ranges from 31.3% to 97% [1]. Oh et al. [6] tried to assess the correlation

between the CT features and Fuhrman grade of ccRCC, where ccRCCs were retrospectively reviewed in consensus by two radiologists. Using logistic regression, they showed a threshold tumor size of 36 mm to predict (AUC: 70%) the high Fuhrman grade. Recently, Sasaguri et al. [7] suggested that RCCs can be characterized and graded based on CT textural features. Ding et al. [1] employed logistic regression on both non-textural features, e.g. pseudocapsule, round mass, as well as textural ones, e.g. 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. [2] also employed logistic regression on CT textural features, e.g. GLCM, GLRLM, gray level size zone matrix (GLSZM), and achieved an FGS accuracy of 77%. Huhdanpaa et al. [8] used histogram analysis of the peak tumor enhancement, tumor heterogeneity and percent contrast washout in CT, and reported these parameters to be statistically different between low and high grade ccRCC.

Current textural feature identification and quantification nonetheless faces two main challenges: it requires (1) ccRCC segmentation in CT, and (2) manual feature engineering. To our knowledge, there is no automatic ccRCC segmentation method present for CT. On the other hand, manual tumor segmentation relying on human visual inspection for feature identification is laborious, time consuming, and suffers from high intra/inter-observer variability [9].

Avoiding complex manual feature engineering, supervised deep learning using convolutional neural networks (CNN) have exploded in popularity for automatic feature learning, classification, as well as localization and dense labelling. In a classical CNN, the learned features in the first layer 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 [10]. These features of a classical CNN tend to ignore diffuse textural features [11] that are often important for medical imaging applications, e.g. tumor characterization and analysis. In an attempt to learn textural features via CNNs, Andrearczyk et al. [11] 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 variation in the input image. In a recent study [12], Wang et al. 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 resulted in a slightly better (1.9%) prediction accuracy.

Inspired by Wang's approach, which was designed to learn histograms of likelihood-maps (for segmentation) or class-likelihoods (for classification) generated by a conventional CNN, we propose ImHistNet, a deep neural network (DNN) for end to end texture-based image classification. Our proposed work makes the following contributions: (1) we modify the learnable histogram approach by Wang et al. [12] into a learnable image histogram (LIH) layer within a DNN framework capable of learning complex and subtle task-specific textural features from raw images directly, adhering to the classical input-output mapping of a CNN; (2) we remove the requirement for fine pre-segmentation of the ccRCC as the proposed learnable image histogram can stratify tumor and background textures well thus enabling the model to focus specifically on the tumor texture; (3) we demonstrate ImHistNet's capabilities by performing automatic ccRCC grade classification for the 2-tiered FGS on an extended clinical dataset from real patients.

2 Materials and Methods

2.1 Data

We used CT scans of 159 patients from The Cancer Imaging Archive (TCIA) database [13]. These patients were diagnosed with ccRCC, of which 64 were graded Fuhrman low (I/II) and 95 were graded Fuhrman 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 normalized the intensity of all the datasets between [-1000, 3000] Hounsfield Units. We divided the dataset for training/validation/testing as 44/5/15 and 75/5/15 for Fuhrman low and and Fuhrman high, respectively. Note that typical tumor radiomic analysis comprises [14]: (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+ccRCC (i.e. both mutually inclusively present) bounding boxes. We do not require any fine pre-segmentation of the ccRCC rather only assume a kidney+ccRCC bounding box, generated in step-ii. For this study, kidney+ccRCC bounding boxes are manually generated. We also do not require any voxel spacing normalization among the datasets. Given data imbalance where samples for Fuhrman low are fewer than for Fuhrman high, we allowed more overlap among adjacent patches for the Fuhrman low dataset. The amount of overlap is calculated to balance the samples from both cohorts.

2.2 Learnable Image Histogram for Classification

Learnable Image Histogram: Our proposed learnable image histogram (LIH) stratifies the pixel values in an image x in different learnable and possibly overlapping intervals (bins of width w_b) with arbitrary 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 fall 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 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^{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. Similar to [12], the linear basis function $H_b(x)$ of the LIH is also piece-wise differentiable and can back-propagate (BP) errors to update β_b and $\widetilde{w}_b$ during training. The gradients of β_b and $\widetilde{w}_b$ for a loss $\mathcal{L}$ are estimated as:

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

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

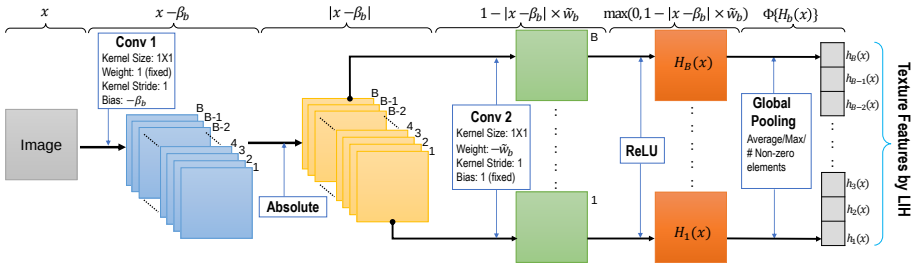

Fig. 1. The architecture of our learnable image histogram using CNN layers.

Design of LIH using CNN Layers: The proposed LIH is implemented using 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 β_b is equivalent to convolving the input by a 1×1 kernel with fixed weight of 1 (i.e. with no updating by back-propagation [BP]) and a learnable bias term β_b ('Conv 1' in Fig. 1). A total of $B = |\mathcal{B}|$ number of

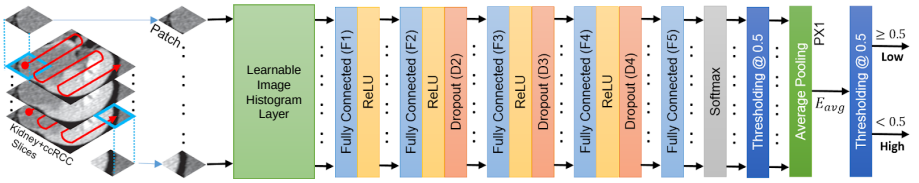

Fig. 2. Multiple instance decisions aggregated ImHistNet for grade classification.

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. 2). The first seven layers contain trainable weights. The input is a 64×64 pixel image patch extracted from the kidney+ccRCC 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 [15]. 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: Fuhrman low, and 0: Fuhrman high). We employed a Dropout unit (Dx) that drops 20%, 30%, and 40% of units in F2, F3 and F4 layers, respectively (Fig. 2) 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. We did not use any batch normalization. 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.

ccRCC Grade Classification: After training ImHistNet (layers 1 to 7) by estimating errors at the 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 ccRCC-containing kidney slices during training and validation, not all the

ccRCC 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 similar multiple instance decision aggregation procedure proposed by Hussain et al. [9]. 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 Fuhrman low, and Fuhrman high otherwise (see Fig. 2).

3 Results and Discussion

Table 1. Automatic ccRCC Fuhrman grade classification performance comparison. NTS: Number of test samples, HE: hand engineered, SVM: support vector machines, xFCV: x-fold cross-validation, LxOCV: leave-x-out cross-validation, '-': Not reported.

Row	Method Types	Methods	NTS	Accuracy	AUC
1	Conventional CNNs	Full image+ResNet-50	30	53%	0.4302
2		Full image+AlexNet	30	56%	0.4756
3		Patch+ResNet-50	30	50%	0.6680
4		Patch+AlexNet	30	56%	0.4505
5	HE Features + Conventional Machine Learning (ConML)	Patch+Histogram (128 bins)+SVM	30	56%	0.5046
6		Patch+Histogram (256 bins)+SVM	30	63%	0.5140
7		Ding et al. [1]	92	-	0.6700
8		Shu et al. [2] (5FCV on 260 samples)	-	77%	0.8220
9		Fei et al. [16] (L1OCV on 90 samples)	-	70%	-
10		Oh et al. [6]	173	-	0.7000
11	HE Features + Deep Learning	Patch+Histogram (128 bins)+5 FCL	30	50%	0.5664
12		Patch+Histogram (256 bins)+5 FCL	30	50%	0.6449
13	LIH + ConML	Patch+LIH (128 bins)+AP+SVM	30	60%	0.5885
14	LIH + Different Number of FCL/bins + Different Pooling Types	Patch+LIH (128 bins)+NZEC+5 FCL	30	50%	0.5502
15		Patch+LIH (128 bins)+AP+4 FCL	30	50%	0.6388
16		Patch+LIH (128 bins)+AP+6 FCL	30	50%	0.6379
17		Patch+LIH (64 bins)+AP+5 FCL	30	50%	0.6386
18		Patch+LIH (256 bins)+AP+ 5 FCL	30	43%	0.6378
19	Combined LIH & Conventional CNN	Patch+LIH (128 bins)+AP+5 FCL	30	53%	0.6501
20		Full Image+LIH (128 bins)+AP+ 5 FCL	30	50%	0.4883
21	**Proposed**	ImHistNet [LIH (128 bins)+AP+5 FCL]	30	**80%**	**0.8495**

We compared our ccRCC grade classification performance in terms of accuracy (%) and area under the curve (AUC) to a wide range of methods in Table 1. 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 64×64 pixels across all contrasted methods.

First, we use ResNet-50 and AlexNet (rows 1-4) with transfer learning in order to test the performance of conventional CNNs. Here, we used the full kidney+ccRCC slices as well as patches as inputs. As we mentioned in Sect. 1 that a classical CNN typically fails to capture textural features, it has become evident from the results where such CNNs performed poorly in learning the textural features of ccRCC. Next, in order to evaluate the performance of hand-engineered (HE) features-based conventional machine learning (ConML) approaches, we tested SVM (rows 5-6) employing the conventional image histogram of 128 and 256 bins. We also compared four state-of-the-art methods in rows 7-10. Since we do not have access to their codes and datasets, we conservatively quote authors' best self-reported performances. These methods mostly relied on the ccRCC textural features, and used classical predictive models, e.g. logistic regression. Here, the method by Shu et al. [2] performed the best with 77% classification accuracy. Then, to contrast the performance of a SVM against a DNN, we fed the conventional histogram (128 and 256 bins) features to a DNN of 5 FCL with weight sizes (4096×1)-(4096×1)-(4096×1)-(4096×1)-(2×1) (rows 11-12). We choose this FCL configuration as our ImHistNet contains the same. The better AUC score by the FCL approaches suggest that it better classify tumor grade than that by the SVM (rows 5-6). Next, to evaluate the HE features against LIH features, we used LIH features to train a SVM (row 13). We see that the SVM with LIH features outperformed the SVM with conventional histogram features (row 5). We also varied the number of bins (64/128/256) and FCLs of size 4096×1 (4/5/6), and the pooling types (AP/NZEC) with the LIH layer (rows 14-18). However, the classification performance in terms of AUC by any of these combinations did not exceed ∼65%. After that, in order to evaluate the performance of a DNN, combining a CNN and the ImHistNet, we added a CNN of AlexNet equivalent configuration in parallel to the ImHistNet. The last FCLs of size 4096×1 in both networks were concatenated and the total network was trained end-to-end. We implemented two such approaches using the full kidney+ccRCC images, as well as the patches as inputs (rows 19-20). We observed that the classical CNN affect the performance of the proposed ImHistNet negatively, i.e. results were worse than those by ImHistNet (row 21). In conclusion, our proposed ImHistNet achieved the highest accuracy and AUC performance among all contrasted methods as can be seen in row 21.

4 Conclusions

We proposed a learnable image histogram based DNN framework for end to end image classification. We demonstrated our approach on a cancer grade prediction task providing automatic 2-tiered FGS (Fuhrman low and Fuhrman high) grade classification of ccRCC 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, as well address class imbalances in the training data via overlap control. We also used multiple instance decision

aggregation to further robustify binary classification. 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 using the learnable image histogram may aid in better diagnosis.

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

References

1. 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
2. Shu, J., Tang, Y., Cui, J., Yang, R., Meng, X., Cai, Z., Zhang, J., Xu, W., Wen, D., Yin, H.: Clear cell renal cell carcinoma: CT-based radiomics features for the prediction of fuhrman grade. European journal of radiology **109** (2018) 8–12
3. Ishigami, K., Leite, L.V., Pakalniskis, M.G., Lee, D.K., Holanda, D.G., Kuehn, D.M.: Tumor grade of clear cell renal cell carcinoma assessed by contrast-enhanced computed tomography. SpringerPlus **3**(1) (2014) 694
4. Fuhrman, S.A., Lasky, L.C., Limas, C.: Prognostic significance of morphologic parameters in renal cell carcinoma. The American journal of surgical pathology **6**(7) (1982) 655–663
5. Becker, A., Hickmann, D., Hansen, J., Meyer, C., Rink, M., Schmid, M., Eichelberg, C., Strini, K., Chromecki, T., Jesche, J., et al.: 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**(3) (2016) 419–425
6. Oh, S., Sung, D.J., Yang, K.S., Sim, K.C., Han, N.Y., Park, B.J., Kim, M.J., Cho, S.B.: Correlation of ct imaging features and tumor size with fuhrman grade of clear cell renal cell carcinoma. Acta Radiologica **58**(3) (2017) 376–384
7. Sasaguri, K., Takahashi, N.: CT and MR imaging for solid renal mass characterization. European journal of radiology **99** (2018) 40–54
8. Huhdanpaa, H., Hwang, D., Cen, S., Quinn, B., Nayyar, M., Zhang, X., Chen, F., Desai, B., Liang, G., Gill, I., et al.: Ct prediction of the fuhrman grade of clear cell renal cell carcinoma (rcc): towards the development of computer-assisted diagnostic method. Abdominal imaging **40**(8) (2015) 3168–3174
9. 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
10. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553) (2015) 436
11. Andrearczyk, V., Whelan, P.F.: Using filter banks in convolutional neural networks for texture classification. Pattern Recognition Letters **84** (2016) 63–69
12. Wang, Z., Li, H., Ouyang, W., Wang, X.: Learnable histogram: Statistical context features for deep neural networks. In: European Conference on Computer Vision, Springer (2016) 246–262

13. 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
14. Aerts, H.J.: The potential of radiomic-based phenotyping in precision medicine: a review. JAMA oncology **2**(12) (2016) 1636–1642
15. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision, Springer (2014) 818–833
16. Meng, F., Li, X., Zhou, G., Wang, Y.: Fuhrman grade classification of clear-cell renal cell carcinoma using computed tomography image analysis. Journal of Medical Imaging and Health Informatics **7**(7) (2017) 1671–1676