Noninvasive Determination of Gene Mutations in Clear Cell Renal Cell Carcinoma using Multiple Instance Decisions Aggregated CNN

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. Kidney clear cell renal cell carcinoma (ccRCC) is the major sub-type of RCC, constituting one the most common cancers worldwide accounting for a steadily increasing mortality rate with 350,000 new cases recorded in 2012. Understanding the underlying genetic mutations in ccRCC provides crucial information enabling malignancy staging and patient survival estimation thus plays a vital role in accurate ccRCC diagnosis, prognosis, treatment planning, and response assessment. Although the underlying gene mutations can be identified by whole genome sequencing of the ccRCC following invasive nephrectomy or kidney biopsy procedures, recent studies have suggested that such mutations may be noninvasively identified by studying image features of the ccRCC from Computed Tomography (CT) data. Such image feature identification currently 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 convolutional neural network approach for automatic detection of underlying ccRCC gene mutations from 3D CT volumes. We aggregate the mutation-presence/absence decisions for all the ccRCC slices in a kidney into a robust singular decision that determines whether the interrogated kidney bears a specific mutation or not. When validated on clinical CT datasets of 267 patients from the TCIA database, our method detected gene mutations with 94% accuracy.

1 Introduction

Kidney cancer, or renal cell carcinomas (RCC) are a common group of chemotherapy resistant diseases that accounted for an estimated 62,000 new patients and 14,000 deaths in the United States in 2015 alone [1]. North America and Europe have recently reported the highest numbers of new cases of RCC in the world [2]. The most common histologic sub-type of RCC is clear cell RCC (ccRCC) [3], which is known to be a genetically heterogeneous disease [4]. Recent studies [5,6] have identified several mutations in genes associated with ccRCC. For example, the von Hippel-Lindau (VHL) tumor suppressor gene, BRCA1-associated protein 1 (BAP1) gene, polybromo 1 (PBRM1) gene, and SET domain containing 2 (SETD2) gene have been identified as the most commonly mutated genes in ccRCC [5].

Traditionally, ccRCC underlying gene mutations are identified by genome sequencing of the ccRCC of the kidney samples after invasive nephrectomy or kidney biopsy [5]. This identification of genetic mutations is clinically important because advanced stages of ccRCC and poor patient survival have been found to be associated with the VHL, PBRM1, BAP1, SETD2, and lysine (K)-specific demethylase 5C (KDM5C) gene mutations [5,7]. Therefore, knowledge of the genetic make-up of a patient's kidney ccRCC has great prognostic value that is helpful for treatment planning [5,7]. Correlations between mutations in genes and different ccRCC features seen in patient CT images has been shown in recent work [4,8]. For example, an association between well-defined tumor margin, nodular enhancement and intratumoral vascularity with the VHL mutation has been reported [8]. Ill-defined tumor margin and renal vein invasion were also reported to be associated with the BAP1 mutation [4] whereas PBRM1 and SETD2 mutations are mostly seen in solid (non-cystic) ccRCC cases [8]. Such use of radiological imaging data as a noninvasive determinant of the mutational status and a complement to genomic analysis in characterizing disease biology is refereed to as 'Radiogenomics' [4,8]. Radiogenomics requires robust image feature identification, which is typically performed by expert radiologists. However, relying on human visual inspection is laborious, time consuming, and suffers from high intra/inter-observer variability.

A number of machine learning (ML) tools have been used to facilitate the processes of high-throughput quantitative feature extraction from volumetric medical images, with some subsequently used in treatment decision support [9]. This practice is generally known as 'Radiomics'. Radiomics uses higher order statistics (on the medical images) combined with clinical and radiogenomic data to develop models that may potentially improve cancer diagnostic, prognostic, and predictive accuracy [9,10]. The typical tumor radiomic analysis workflow has 4 distinct steps: (1) 3D imaging, (2) manual or automatic tumor detection and segmentation, (3) tumor phenotype quantification, and (4) data integration (i.e. phenotype+genotype+clinical+proteomic) and analysis [10]. Discriminant features needed for proper mutation detection are often not seen in the marginal ccRCC cross-sections (e.g. axial slices in the top and bottom regions of a ccRCC). This scenario makes 'single-instance' ML approaches, especially convolutional neural network (CNN) very difficult to train, as some of the input slices do not contain discriminating features, thus do not correspond to the assigned mutation label. Another solution is to use the full 3D volume as a single instance. However, 3D CNNs are considerably more difficult to train as they contain significantly more parameters and consequently require many more training samples, while the use of the 3D volume itself severely reduces the available number of training samples than its 2D counterpart. An alternative approach is multiple-instance learning (MIL) [11], where the learner receives a set of labeled bags (e.g. mutation present/absent), each containing multiple instances (e.g. all the ccRCC slices in a kidney). A MIL model labels a bag with a class even if some or most of the instances within it are not members of that class.

We propose a deep CNN approach that addresses the challenge of automatic mutation detection in kidney ccRCC. Our method is a variant of the conventional

MIL approach, where we use multiple instances for robust binary classification, while using single instances for training the CNN to facilitate higher number and variation of training data. The CNN automatically learns the ccRCC image features and the binary decisions (i.e. presence/absence of a mutation) for all the ccRCC slices in a particular kidney sample are aggregated into a robust singular decision that ultimately determines whether an interrogated kidney sample has undergone a certain mutation or not. Our method can be incorporated in the Radiomics step-3 given that the tumor boundary is already known in step-2. The estimated mutation data can subsequently be integrated in the step-4. The frequency of occurrence of various mutations in ccRCC varies significantly, e.g. VHL, PBRM1, BAP1, SETD2 and KDM5C were found in 76%, 43%, 14%, 14% and 8% of kidney samples of our dataset, respectively. In this study we consider the four most prevalent gene mutations (i.e. VHL, PBRM1, BAP1 and SETD2). We achieve this via four multiple instance decisions aggregation CNNs, however, our approach is directly extendable to more mutation types depending on the availability of sufficient training data.

2 Materials and Methods

2.1 Data

We obtained access to 267 patients' CT scans from The Cancer Imaging Archive (TCIA) database [12]. In this dataset, 138 scans contained at least one mutated gene because of ccRCC. For example, 105 patients had VHL, 60 patients had PBRM1, 60 patients had SETD2, and 20 patients had BAP1 mutations. In addition, some of the patients had multiple types of mutations. However, 9 patients had CT scans acquired after nephrectomy and, therefore, those patients' data were not usable for this study. The images in our database included 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. Ground truth mutation labels were collected from the *cBioPortal for Cancer Genomics* [7].

Table 1. Number of kidney samples used in training, validation and testing per mutation case. Acronym used: M: mutation.

Genes	# Training Samples		# Validation Samples		# Test Samples	
	M-Present	M-Absent	M-Present	M-Absent	M-Present	M-Absent
VHL	74	74	10	10	15	15
PBRM1	35	35	6	6	10	10
SETD2	11	11	3	3	5	5
BAP1	10	10	3	3	4	4

We show the number of kidney samples used in the training, validation and testing stages in Table 1. During training, validation and testing, we use only those slices of the kidney that contain ccRCC as our CNNs aim to learn ccRCC features. We form a 3-channel image from each scalar-valued CT slice by generating channel intensities [I, I-50, I-100] HU, where I represents the original intensities in a CT image slice tightly encompassing a kidney+ccRCC cross-section (Fig. 1), whereas I-50 and I-100 represent two variants of I with different

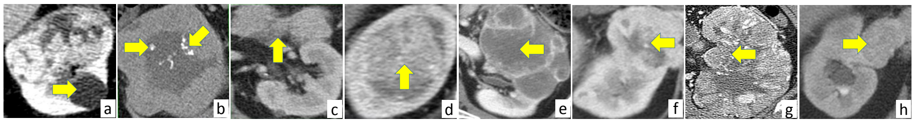

Fig. 1. Illustration of CT features of ccRCC seen in the data of this study. (a) Cystic tumor architecture, (b) calcification, (c) exophytic tumor, (d) endophytic tumor, (e) necrosis, (f) ill-defined tumor margin, (g) nodular enhancement, and (h) renal vein invasion. Arrow indicates feature of interest in each image.

HU values. We add these variations in channel intensity values as similar ccRCC features may have different X-ray attenuation properties across patients [4]. We resized all the image data by resampling into a size of 227×227×3 pixels. We augmented the number of training samples by a factor of 24 by flipping and rotating the 3-ch image slices as well as by re-ordering the 3 channels in each image. We normalized the training and validation data before training by subtracting the image mean and dividing by the image standard deviation.

2.2 Multiple Instance Decision Aggregation for Mutation Detection

Typically, ccRCC grows in different regions of the kidney and is clinically scored on the basis of their CT slice-based image features, such as size, margin (well- or ill-defined), composition (solid or cystic), necrosis, growth pattern (endophytic or exophytic), calcification etc. [4]. Some of these features seen in our dataset are shown in Fig. 1. We propose to learn corresponding features from the CT images using four different CNNs: VHL-CNN, PBRM1-CNN, SETD2-CNN and BAP1-CNN, each for one of the four mutations (VHL, PBRM1, SETD2 and BAP1). Using a separate CNN per mutation alleviates the problem of data imbalance among mutation types, given that the mutations are not mutually exclusive.

CNN Architecture: All the CNNs in this study (i.e. VHL-CNN, PBRM1-CNN, SETD2-CNN and BAP1-CNN) have similar configuration but are trained separately (Fig. 2). Each CNN has twelve layers excluding the input: five convolutional (Conv) layers; three fully connected (FC) layers; one softmax layer; one average pooling layer; and two thresholding layers. All but the last three layers contain trainable weights. The input is the 227×227×3 pixel image slice containing the kidney+ccRCC. We train these CNNs (layers 1–9) using a balanced dataset for each mutation case separately (i.e. a particular mutation-present and absent). During training, images are fed to the CNNs in a randomly shuffled single instance fashion. Typically, Conv layers are known for sequentially learning the high-level non-linear spatial image features (e.g. ccRCC size, orientation, edge variation, etc). We used five Conv layers as the 5th Conv layer typically grabs an entire object (e.g. ccRCC shape) in an image even if there is a significant pose variation [13]. Subsequent FC layers prepare those features for optimal classification of an interrogated image. In our case, three FC layers are deployed to make the decision on the learned features from the 3-ch images to decide if a particular gene mutation is probable or not. The number of FC layers plays a vital role as the overall depth of the model is important for obtaining good performance [13], and we achieve optimal performance with three FC layers. Layers

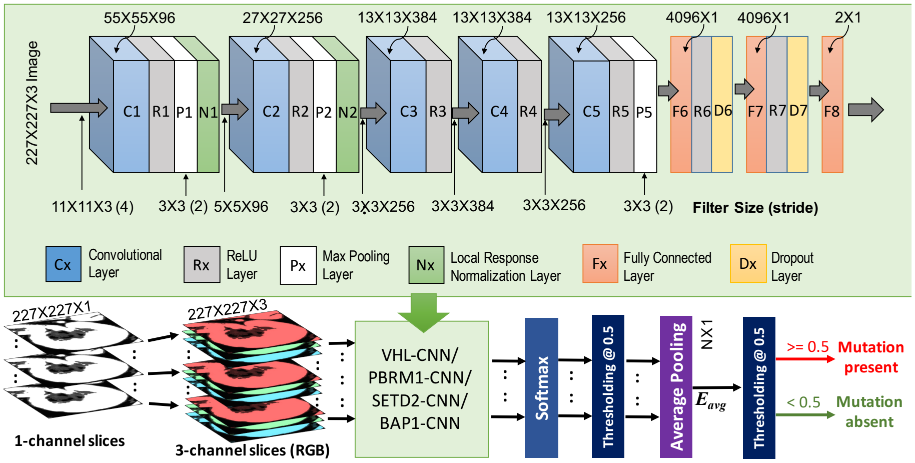

Fig. 2. Multiple instance decisions aggregated CNN for gene mutation detection.

10, 11 and 12 (i.e. two thresholding and one average pooling layers) of the CNNs are used during the testing phase and do not contain any trainable weights.

Solver: These networks were trained by minimizing the softmax loss between the expected and detected labels (1: mutation present and 0: mutation absent). We used the *Adam* optimization method [14]. All the parameters for this solver were set to the suggested (by [14]) default values, i.e. $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. We also employed a Dropout unit (Dx) that drops 50% of units in both F6 and F7 layers (Fig. 2) and used a weight decay of 0.005. The base learning rate was set to 0.01 and was decreased by a factor of 0.1 to 0.0001 over 250,000 iterations with a batch of 256 images processed at each iteration. 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.

Mutation Detection: After all the CNNs are trained (from layer 1 to 9), we use the full configuration (from layer 1 to 12) in the testing phase. Although we use only ccRCC containing kidney slices during training and validation, often not all the ccRCC cross-sections contains the discriminating features for proper mutation detection. Therefore, our trained CNN (from layer 1 to 9) often miss-classifies the interrogated image slice based on the probability estimated at the layer 9 (i.e. softmax layer). In order to address this misclassification by our CNNs, we adopt a multiple instance decision aggregation procedure. In this procedure, we feed all the candidate image slices of a particular kidney to the trained CNN and accumulate the slice-wise binary classification labels (0 or 1) at layer 10 (the thresholding layer). These labels are then fed into a $N \times 1$ average pooling layer, where N is the total number of 3-channel axial slices of an interrogated kidney. Finally, the estimated average (E_{avg}) from layer 11 is fed to the second thresholding layer (layer 12), where $E_{avg} \geq 0.5$ indicates the presence of the mutation in that kidney, and no-mutation otherwise (see Fig. 2).

Table 2. Automatic gene mutation detection performance of different methods. We use acronyms as: M: mutation, x: one of VHL/PBRM1/SETD2/BAP1, Aug: augmentation, SI: single instance, MI: multiple instance, 3ch: 3-channel data with augmentation by channel re-ordering, F: augmentation by flipping, and R: augmentation by rotation.

Methods	Genes	# Test Samples		# Correct Detection		Overall	Mean
	x	M-present	M-absent	M-present	M-absent	Error (%)	Error (%)
Random Forest (SI+1ch) No Aug	VHL	15	15	5	7	60	53.75
	PBRM1	10	10	4	5	55	
	SETD2	5	5	2	3	50	
	BAP1	4	4	2	2	50	
x-CNN (SI+1ch) No Aug	VHL	15	15	7	8	50	41.88
	PBRM1	10	10	6	6	40	
	SETD2	5	5	3	3	40	
	BAP1	4	4	2	3	37.50	
x-CNN (SI+3ch)	VHL	15	15	12	9	30	29.38
	PBRM1	10	10	4	7	45	
	SETD2	5	3	4	4	30	
	BAP1	4	4	3	4	12.5	
x-CNN (SI+1ch +F+R)	VHL	15	15	11	13	20	21.88
	PBRM1	10	10	8	7	25	
	SETD2	5	5	3	4	30	
	BAP1	4	4	4	3	12.50	
x-CNN (SI+3ch +F+R)	VHL	15	15	15	11	13.33	13.96
	PBRM1	10	10	9	9	10	
	SETD2	5	5	5	3	20	
	BAP1	4	4	3	3	12.50	
Proposed (MI+3ch +F+R +Average)	VHL	15	15	14	13	**10**	**6.25**
	PBRM1	10	10	9	10	**5**	
	SETD2	5	5	5	4	**10**	
	BAP1	4	4	4	4	**0**	

3 Results

We compare the mutation detection performance by a wide range of methods. At first, we tested the performance using a single instance (SI)-based random forest (RF) approach, where hand-engineered image features were used. In a typical SI-based classification approach, the class-label is decided from the maximum among the predicted class-probabilities [15]. Similarly in our SI-based approaches, presence or absence of a certain mutation is decided from the maximum among the estimated probabilities associated with all the ccRCC image slices in a particular kidney. Then we demonstrate the effectiveness of automatic feature learning compared to the hand-engineered features generation using the CNN approach. Afterwards, we show the effect of incorporating augmented data in the training dataset and compared the mutation detection performance for three different types of augmentation (i.e. image flipping+rotation, 3-ch re-ordering and those combined). Finally, we demonstrated the effectiveness of using multiple instance decisions aggregation in our proposed method.

In row 1 of the comparison Table 2, we show results of a traditional RF approach with hand-engineered image features shown to be effective in anatomy classification task [15]: histogram of oriented gradient, Haar features, and local binary patterns. Here, we did not augment any manually transformed data to the training samples. We trained four RFs for the four different mutation cases and as we see in Table 2, the resulting mean detection error was the highest (∼54%) among all contrasted methods. Row 2 shows the results of a deep CNN (namely, x-CNN, where x: VHL/PBRM1/SETD2/BAP1 (see Fig. 2)) approach with no data augmentation. Since the CNN learns the image features automatically, it may have helped this CNN method perform better (mean error ∼42%) than that of the hand-engineered features-based RF approach. Row 3 shows results for x-CNN, where we used data augmentation by deploying 3-ch data and re-ordering of channels (see Sect. 2.1). These data were fed to x-CNN and it can be seen how the SI-based mutation detection performance by this approach (mean error ∼29%) outperformed that with no data augmentation. Thus, including channels with different intensity ranges, mimicking the tumor intensity variation across patients, have shown positive impact on the mutation detection task. Row 4 shows results for x-CNN with a different augmentation process, which deploys the flipping and rotating of the 1-ch training samples. This approach (mean error ∼22%) outperformed that with 3-ch augmentation. So it is clear that the flipping+rotation-based augmentation introduced more variation in the training data than that by the 3-ch augmentation, resulting in better generalization of the model. In the method shown in row 5, we combined the flipping+rotation augmentation with the 3-ch re-ordering augmentation. The performance of the x-CNN with these data was better in mutation detection (mean error ∼14%) than that of flipping+rotation or 3-ch augmentation alone (see Table 2). Finally, row 6 demonstrates results of our proposed method, where flipping, rotation and 3-ch re-ordering augmentations were used. In addition, binary classification was performed based on the multiple instance decisions aggregation. We see in the Table 2 that the mean mutation detection error by our method is ∼6%, which is the lowest tested. In addition, detection errors for individual mutation cases were also low and in the range of 10%. Thus, our multiple instance decisions aggregation procedure made our CNN models more robust on SI-based miss-classification.

4 Conclusions

In this paper, we proposed a multiple instance decision aggregation-based deep CNN approach for automatic mutation detection in kidney ccRCC. We have shown how our approach automatically learned discriminating ccRCC features from CT images and aggregated the binary decisions on the mutation-presence/absence for all the ccRCC slices in a particular kidney sample. This aggregation produced a robust decision on the presence of a certain mutation in an interrogated kidney sample. In addition, our multiple instance decision aggregation approach achieved better accuracy in mutation detection than that of a typical single instance-based approach. On the other hand, better performance by conventional MIL approaches is subject to the availability of sufficient number of data, while in applications such as ours, there are usually very few data samples for some of the mutation cases. Therefore, an end-to-end MIL approach will most likely fail for those mutation cases with few data samples. However, this paper included a number of meaningful comparisons to highlight the effects of different augmentation, pooling schemes etc within the context of insufficient data, which we believe provide more interesting findings and appears to be suitable for ccRCC Radiomics, where the learned mutations would aid in better ccRCC diagnosis, prognosis and treatment response assessment. Our experimental results demonstrated an approximately 94% accuracy in kidney-wise mutation detection.

References

1. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2015. CA: a cancer journal for clinicians **65**(1) (2015) 5–29
2. Ridge, C.A., Pua, B.B., Madoff, D.C.: Epidemiology and staging of renal cell carcinoma. In: Seminars in interventional radiology. Volume 31., Thieme Medical Publishers (2014) 003–008
3. Lam, J.S., Shvarts, O., Leppert, J.T., Figlin, R.A., Belldegrun, A.S.: Renal cell carcinoma 2005: New frontiers in staging, prognostication and targeted molecular therapy. The Journal of urology **173**(6) (2005) 1853–1862
4. Shinagare, A.B., Vikram, R., Jaffe, C., Akin, O., Kirby, J., et al.: Radiogenomics of clear cell renal cell carcinoma: preliminary findings of The Cancer Genome Atlas–Renal Cell Carcinoma (TCGA–RCC) Imaging Research Group. Abdominal imaging **40**(6) (2015) 1684–1692
5. Network, C.G.A.R., et al.: Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature **499**(7456) (2013) 43
6. Guo, G., Gui, Y., Gao, S., Tang, A., Hu, X., Huang, Y., Jia, W., et al.: Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. Nature genetics **44**(1) (2012) 17
7. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., et al.: Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. **6**(269) (2013) pl1–pl1
8. Karlo, C.A., Di Paolo, P.L., Chaim, J., Hakimi, A.A., et al.: Radiogenomics of clear cell renal cell carcinoma: associations between ct imaging features and mutations. Radiology **270**(2) (2014) 464–471
9. Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: images are more than pictures, they are data. Radiology **278**(2) (2015) 563–577
10. Aerts, H.J.: The potential of radiomic-based phenotyping in precision medicine: a review. JAMA oncology **2**(12) (2016) 1636–1642
11. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence **89**(1-2) (1997) 31–71
12. 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
13. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision, Springer (2014) 818–833
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Wang, H., Moradi, M., Gur, Y., Prasanna, P., Syeda-Mahmood, T.: A multi-atlas approach to region of interest detection for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2017) 168–176