Deep Learning to Distinguish Benign from Malignant Renal Lesions Based on Routine MR Imaging

Ianto Lin Xi1, Yijun Zhao2, Robin Wang1, Marcello Chang5, Subhanik Purkayastha4, Ken Chang5, Raymond Y. Huang6, Alvin C. Silva7, Martin Vallières6, Peiman Habibollahi5, Yong Fan10, Beiji Zou11, Terence P. Gade12, Paul J. Zhang13, Michael C. Soulen12, Zishu Zhang2, Harrison X. Bai4, and S. William Stavropoulos12

ABSTRACT

Purpose: With increasing incidence of renal mass, it is important to make a pretreatment differentiation between benign renal mass and malignant tumor. We aimed to develop a deep learning model that distinguishes benign renal tumors from renal cell carcinoma (RCC) by applying a residual convolutional neural network (ResNet) on routine MR imaging.

Experimental Design: Preoperative MR images (T2-weighted and T1-postcontrast sequences) of 1,162 renal lesions definitely diagnosed on pathology or imaging in a multicenter cohort were divided into training, validation, and test sets (70:20:10 split). An ensemble model based on ResNet was built combining clinical variables and T1C and T2WI MR images using a bagging classifier to predict renal tumor pathology. Final model performance was compared with expert interpretation and the most optimized radiomics model.

Results: Among the 1,162 renal lesions, 655 were malignant and 507 were benign. Compared with a baseline zero rule algorithm, the ensemble deep learning model had a statistically significant higher test accuracy (0.70 vs. 0.56, P = 0.004). Compared with all experts averaged, the ensemble deep learning model had higher test accuracy (0.70 vs. 0.60, P = 0.053), sensitivity (0.92 vs. 0.81, P = 0.017), and specificity (0.41 vs. 0.35, P = 0.450). Compared with the radiomics model, the ensemble deep learning model had higher test accuracy (0.70 vs. 0.62, P = 0.081), sensitivity (0.92 vs. 0.79, P = 0.012), and specificity (0.41 vs. 0.39, P = 0.770).

Conclusions: Deep learning can noninvasively distinguish benign renal tumors from RCC using conventional MR imaging in a multi-institutional dataset with good accuracy, sensitivity, and specificity comparable with experts and radiomics.

Introduction

Renal cell carcinomas (RCC) account for 85% of renal malignancies, affecting approximately 65,000 new patients each year (1). Resecting renal masses radiographically suspicious for RCC without a tissue diagnosis is within the standard management (2, 3), leading to overtreatment of benign renal tumors. About 20% of surgically removed renal masses are reported to be benign (4), which challenges the necessity of surgery for all suspicious lesions because of concurrent risk and morbidity exposed to patients (5).

Noninvasive preoperative imaging techniques, such as ultrasound, CT, and MRI, are widely used for characterization of renal masses. CT and MRI have limited specificity and sensitivity in the discrimination of benign, indolent masses, such as oncocytoma and angiomylipoma (AML), from aggressive malignant renal tumors (6–8). Percutaneous biopsy is nondiagnostic in 20% and provides an erroneous diagnosis in another 10% of resected case (9). There is an unmet need for more accurate imaging diagnosis of small renal masses.

Radiomics features have been used for renal tumor evaluation (10–16). In these studies, predefined features such as shape, intensity and texture were selected and fed into a machine learning algorithm. However, these manually formulated or “handcrafted” features may not capture the full range of the information contained within the images and are limited by low reproducibility (17–24). Deep learning extracts features directly from the raw images and generates features adaptive to a given problem. Although machine learning and deep learning have been successful in predicting tumor molecular features, treatment response, and prognosis in oncology, there are very few studies focusing on renal tumors in the literature (13, 14, 25, 26).

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

I.L. Xi and Y. Zhao are co-authors of this article.

Corresponding Authors: S. William Stavropoulos, Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania 19104. Phone: 215-662-6839; Fax: 215-615-3545; E-mail: S.Stavropoulos@uphs.upenn.edu; and Harrison X. Bai, Department of Diagnostic Imaging, Warren Alpert Medical School of Brown University, Providence, RI 02912. Phone: 248-633-4792; Fax: 248-633-4792; E-mail: Harrison_Bai@Brown.edu; and Zishu Zhang, Department of Radiology, The Second Xiangya Hospital, Central South University, No. 159 Middle Renmin Road, Changsha, Hunan 410011, P.R. China. Phone: 86-735-9731-6017; Fax: 8607-31852-9216; E-mail: zishuzhang@csu.edu.cn

Deep Learning for Characterization of Renal Lesions

Translational Relevance
With the wide use of imaging modalities, the detection of incidental renal tumors increases rapidly. There is a substantial number of patients with benign renal tumors who undergo unnecessary surgery, with its concurrent risk and morbidity. Ultrasound, CT, and MRI have limited sensitivity and specificity in the discrimination of benign, indolent masses from aggressive malignant renal tumors. Percutaneous biopsy is nondiagnostic in 20% and provides an erroneous diagnosis in another 10% of resected cases. More accurate imaging diagnosis of renal masses is an urgent need. We trained a deep learning model to distinguish benign renal lesions from renal cell carcinoma with good accuracy, sensitivity, and specificity when compared with experts and radiomics. Our algorithm offers broad applicability by using conventional MR imaging sequences. Furthermore, it has the potential to save patients from unnecessary invasive biopsies/surgeries and to help guide clinical management.

Materials and Methods

Patient cohort
Patients with renal lesions confirmed by histology or imaging were retrospectively identified from two large academic centers in the United States (HUP and MAY), two hospitals in People’s Republic of China (SXH and PHH) and The Cancer Imaging Archive (TCIA) from 1984 to 2019. The study was conducted in accordance with Declaration of Helsinki and approved by the Institutional Review Boards of HUP, MAY, SXH, and PHH. With the agreement to use TCIA data, the IRB approval of our study was waived for TCIA. The informed consent of patients was waived. For histologic confirmed renal lesions, the inclusion criteria were (i) pathologically confirmed renal lesions, (ii) available preoperative MRI including T2-weighted (T2) and T1-contrast (T1C) enhanced sequences, (iii) quality of the images was adequate for analysis, without motion or artifacts. For a benign renal lesion to have a definite diagnosis on imaging, it must have typical imaging features and be stable on imaging follow-up. Our final cohort consisted of 1,162 renal lesions (913 lesions from HUP, 118 lesions from MAY, 56 lesions from TCIA, 25 lesions from SXH, and 50 lesions from PHH; Supplementary Fig. S1).

Among the 1,162 lesions, 655 were malignant based on histopathology and 507 were benign (162 lesions were confirmed by histopathology, other 345 lesions were diagnosed radiographically). The images were randomly divided into a training set of 816 lesions with 408,000 augmented images, a validation set of 234 lesions, and a test set of 112 lesions. All 347 radiographically diagnosed lesions were kept within the training or validation set. The detailed clinical characteristics of the total cohort and among train, validation and test sets are shown in Table 1 and Supplementary Table S1.

Image segmentation
MR images of all patients were loaded into 3D Slicer software (v4.6), regions of interest were manually drawn slice-by-slice on the T2 and T1C sequences by an abdominal radiologist (Y. Zhao) with 5 years of experience reading abdominal MRI (27). Two additional radiologists (Y.C. and L.Z.), with 4 and 3 years of reading abdominal MRI respectively, segmented all the images in the test set, creating a total of three sets of segmentations for the images in the test set. A Dice similarity coefficient (DSC; ref. 28) was calculated between these three sets of segmentations.

Image processing
All images were downloaded in DICOM format at their original dimensions and resolution. Images were then processed with N4 bias correction (29) using ANTS (30) and intensity normalization of T2 images (31) using SimpleITK (32), using the T1 image of the same lesion as the reference image (33). Segmentations were used to crop the images. To create two dimensional images suitable for ResNet input, we refer to this method as the 2.5D Model (34). This process is visualized in Fig. 1A.

Model building
A 7:2:1 training/validation/testing split was performed as detailed above. Model building was performed on the segmented images. Two

Table 1. Histologic types and clinical characteristics of our cohort.

<table>
<thead>
<tr>
<th>Tumor subtype</th>
<th>Benign (n = 507)</th>
<th>Malignant (n = 655)</th>
<th>P values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 1</td>
<td>160 (31.8%)</td>
<td>354 (53.9%)</td>
<td>&lt;0.001a</td>
</tr>
<tr>
<td>Grade 2</td>
<td>32 (6.3%)</td>
<td>92 (13.9%)</td>
<td>&lt;0.01a</td>
</tr>
<tr>
<td>Grade 3</td>
<td>152 (29.6%)</td>
<td>154 (23.5%)</td>
<td>0.223</td>
</tr>
<tr>
<td>Papillary RCC</td>
<td>406 (79.7%)</td>
<td>36 (5.5%)</td>
<td>&lt;0.001a</td>
</tr>
<tr>
<td>Multilocular cystic RCC</td>
<td>3 (0.6%)</td>
<td>7 (1.1%)</td>
<td></td>
</tr>
<tr>
<td>Unclassified RCC</td>
<td>92 (18.1%)</td>
<td>406 (60.1%)</td>
<td></td>
</tr>
<tr>
<td>Oncocytoma</td>
<td>50 (9.8%)</td>
<td>406 (60.1%)</td>
<td></td>
</tr>
<tr>
<td>Metanephric adenoma</td>
<td>3 (0.6%)</td>
<td>3 (0.6%)</td>
<td></td>
</tr>
<tr>
<td>Renal adenoma</td>
<td>3 (0.6%)</td>
<td>3 (0.6%)</td>
<td></td>
</tr>
<tr>
<td>Fuhrman/ISUP grade</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 1</td>
<td>66 (10.1%)</td>
<td>66 (10.1%)</td>
<td></td>
</tr>
<tr>
<td>Grade 2</td>
<td>304 (46.4%)</td>
<td>304 (46.4%)</td>
<td></td>
</tr>
<tr>
<td>Grade 3</td>
<td>160 (24.4%)</td>
<td>160 (24.4%)</td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td>35 (5.4%)</td>
<td>35 (5.4%)</td>
<td></td>
</tr>
<tr>
<td>Unavailable</td>
<td>90 (13.7%)</td>
<td>90 (13.7%)</td>
<td></td>
</tr>
</tbody>
</table>

*aStatistically significant.

Abbreviation: ISUP, International Society of Urological Pathology.
Figure 1.
Deep Learning for Characterization of Renal Lesions

separate models were independently trained on T1C and T2W images. During training, images were scaled up or down to 200 by 200 pixel squares using bilinear interpolation, then augmented live with horizontal flip, vertical flip, shear, and zoom transformations to add variability to the training set. Models were trained with a batch size of 16. Early stopping was used with the patience parameter set to 50; finally, each model was given a maximum of 500 epochs of training if it was never stopped early. After 100 training trials, the model with the best validation accuracy was selected. During training, a predetermined probability of 0.5 was assigned to the final sigmoid activation neuron as a threshold for classification of malignancy. The models were trained to maximize accuracy of the model prediction.

A 5-fold cross-validation was used to evaluate the model building pipeline. In addition, we ran our pipeline with all 118 lesions from one institution (MAY) placed in a separate test set, with the remaining lesions split into training (811 lesions) and validation (233 lesions) cohorts.

Model architecture

Models trained on T1C and T2WI image input were based on the ResNet50 architecture (35) with the following modifications: the 1000-class softmax fully connected layer was replaced with five fully-connected layers of decreasing width (256, 128, 64, 32, 16) with ReLU activations, and a single sigmoid output neuron for probability output and binary classification (benign or malignant); to address class imbalance, the loss was weighted by the reciprocal of the frequency. Pretrained convolutional neural network weights from ImageNet were used to initialize our model’s weights and biases and were left unfrozen during training (36). No pretrained weights from the T1C model, and the T2WI model.

An ensemble model was created using a bagging classifier (37) combining the output of the clinical variable logistic regression model, the T1C model, and the T2WI model. Figure 1B demonstrates the architecture of the final ensemble model. This model was trained on the same training data and evaluated on both the validation and test set data. One hundred ensemble models were trained with the highest validation accuracy model selected as the final ensemble model. The ensemble model was calibrated using the Pratt method using validation data (38).

Expert evaluation

Four expert radiologists (H.L., T.H., L.S., and D.P.), with 23, 12, 13, and 10 years of experience reading abdominal MRI respectively, blind to histopathologic data, evaluated unsegmented MRI images of the renal lesions in the test set for malignancy. They were given clinical information of each patient. The model’s results were compared to these expert evaluations to assess model performance.

Radiomics model building and evaluation

Radiomics features were extracted from each patient’s MRI for both T1C and T2WI sequences. For each image space, 79 non-texture (morphology and intensity-based) and 94 texture features were extracted according to the guidelines defined by the Image Biomarker Standardization Initiative (39). Each of the 94 texture features were computed 32 times using all possible combinations of the following extraction parameters, a process known as “texture optimization” (REF): (i) isotropic voxels of size 1, 2, 3, and 4 mm, (ii) fixed bin number (FBN) discretization algorithm, with and without equalization, and (iii) the number of gray levels of 8, 16, 32, and 64 for FBN. A total of (79 + 32*94), or 3087, radiomics features were thus computed in this study. All the features were normalized using unity based normalization and features from T1C and T2WI were combined into one dataset. In order to reduce dimensionality of the datasets, radiomics features were selected for training using thirteen different feature selection methods. Ten machine learning classifiers were trained and tested on features from the same splits of patients used in the deep learning methods. Names of the classifiers and feature selection methods can be found in Supplementary Figure S2. Each classifier was trained on the training set thirteen times using thirteen different feature selection methods and validated through 10-fold cross-validation. Classifiers were trained on 10, 20, 30, 40, 50, and 100 selected features and performances were compared on the validation set. In addition to performance, the stability of both classifiers and feature selection methods was recorded. We calculated relative standard deviation (RSD%) for classifier stability and used a stability measure proposed by Nogueira et al for feature selection stability (40). The performances of the top-performing classifiers were then compared to the performance of the automated optimized machine learning pipeline computed by TPOT, a Tree-Based Pipeline Optimization Tool that chooses the most optimal machine learning pipeline for an inputted dataset through genetic programming (41). The best-performing models were then tested on the final testing set.

Model assessment and statistical analysis

Each trained model was finalized with weights and biases frozen, and then assessed for its performance on validation and test sets based on its F1 score, accuracy, sensitivity, specificity, and AUC of its precision recall curve. In addition, models were evaluated on two additional sets of the same test data images that were segmented by two other radiologists. Finally, the model pipeline was run on a train/validation/test split where the test set was comprised of lesions from a separate institution that was explicitly not included in the training or validation sets.

Activations from the last convolutional layer of the best performing models were visualized by t-distributed Stochastic Neighbor Embedding (42).

Confidence interval was calculated using the adjusted Wald method (43). P values were calculated using a binomial test. A P value of 0.05 was considered as the threshold for significance. To evaluate the calibration of our final models, calibration scores were calculated using the expected calibration error (44).

Code availability

The implementation of our deep learning model was based on the Keras package (45) with the Tensorflow library as our backend (46). Our models were trained on a computer with two NVidia V100 GPUs. To allow other researchers to develop their models, the code is publicly available on Github at https://github.com/intrepi dilemon/renal-mri. The implementation of the radiomics feature extraction was based on “radiomics-develop” package from the Naga lab in McGill University (47, 48). This code is available for public use on Github at https://github.com/mwallieres/radiomics-develop. The radiomics pipeline was developed using Python’s sklearn package. This code is publicly available at https://github.com/subhanik1999/Radiomics-ML.
Results

Segmentation similarity

The average DSC between all of our segmenters was 0.81. Segmenters 1 and 2 had an average DSC of 0.82. Segmenters 2 and 3 had an average DSC of 0.81. Segmenters 1 and 3 had an average DSC of 0.81. Benign lesions had an average DSC of 0.76 across all segmenters while malignant lesions had an average DSC of 0.82 across all segmenters.

Model performance

Performance of T1C, T2, clinical features and ensemble models in training, validation, and test sets compared with expert evaluation and radiomics model.

Table 2. Performance of TIC, T2, clinical features and ensemble models in training, validation, and test sets compared with expert evaluation and radiomics model.

<table>
<thead>
<tr>
<th>Training modality</th>
<th>F1 Score</th>
<th>ROC AUC</th>
<th>PR AUC</th>
<th>Acc (95% CI)</th>
<th>TPR (95% CI)</th>
<th>TNR (95% CI)</th>
<th>PPV</th>
<th>NPV</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical</td>
<td>0.76</td>
<td>0.73</td>
<td>0.72</td>
<td>0.69 (0.65–0.72)</td>
<td>0.87 (0.83–0.89)</td>
<td>0.46 (0.41–0.51)</td>
<td>0.67</td>
<td>0.72</td>
<td>0.33</td>
</tr>
<tr>
<td>TIC</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99 (0.99–1.00)</td>
<td>1.00 (0.98–1.00)</td>
<td>0.99 (0.97–1.00)</td>
<td>0.99</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>T2</td>
<td>0.95</td>
<td>0.99</td>
<td>0.99</td>
<td>0.94 (0.92–0.96)</td>
<td>0.97 (0.95–0.99)</td>
<td>0.90 (0.86–0.95)</td>
<td>0.95</td>
<td>0.96</td>
<td>0.07</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.96</td>
<td>0.99</td>
<td>0.99</td>
<td>0.95 (0.94–0.97)</td>
<td>0.99 (0.98–1.00)</td>
<td>0.91 (0.87–0.93)</td>
<td>0.93</td>
<td>0.99</td>
<td>0.07</td>
</tr>
<tr>
<td>Validation modality</td>
<td>F1 Score</td>
<td>ROC AUC</td>
<td>PR AUC</td>
<td>Acc (95% CI)</td>
<td>TPR (95% CI)</td>
<td>TNR (95% CI)</td>
<td>PPV</td>
<td>NPV</td>
<td>FDR</td>
</tr>
<tr>
<td>Clinical</td>
<td>0.77</td>
<td>0.64</td>
<td>0.64</td>
<td>0.68 (0.62–0.74)</td>
<td>0.92 (0.86–0.96)</td>
<td>0.56 (0.28–0.46)</td>
<td>0.65</td>
<td>0.79</td>
<td>0.35</td>
</tr>
<tr>
<td>TIC</td>
<td>0.72</td>
<td>0.69</td>
<td>0.73</td>
<td>0.66 (0.60–0.72)</td>
<td>0.77 (0.69–0.84)</td>
<td>0.52 (0.42–0.61)</td>
<td>0.68</td>
<td>0.64</td>
<td>0.33</td>
</tr>
<tr>
<td>T2</td>
<td>0.77</td>
<td>0.73</td>
<td>0.71</td>
<td>0.71 (0.65–0.77)</td>
<td>0.83 (0.75–0.88)</td>
<td>0.57 (0.47–0.66)</td>
<td>0.71</td>
<td>0.72</td>
<td>0.29</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.80</td>
<td>0.77</td>
<td>0.73</td>
<td>0.75 (0.69–0.80)</td>
<td>0.87 (0.80–0.92)</td>
<td>0.60 (0.50–0.69)</td>
<td>0.74</td>
<td>0.78</td>
<td>0.26</td>
</tr>
<tr>
<td>Test modality</td>
<td>F1 Score</td>
<td>ROC AUC</td>
<td>PR AUC</td>
<td>Acc (95% CI)</td>
<td>TPR (95% CI)</td>
<td>TNR (95% CI)</td>
<td>PPV</td>
<td>NPV</td>
<td>FDR</td>
</tr>
<tr>
<td>Clinical</td>
<td>0.66</td>
<td>0.43</td>
<td>0.54</td>
<td>0.52 (0.43–0.61)</td>
<td>0.83 (0.71–0.90)</td>
<td>0.12 (0.05–0.25)</td>
<td>0.55</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>TIC</td>
<td>0.73</td>
<td>0.62</td>
<td>0.59</td>
<td>0.64 (0.55–0.73)</td>
<td>0.87 (0.77–0.94)</td>
<td>0.35 (0.23–0.49)</td>
<td>0.63</td>
<td>0.68</td>
<td>0.37</td>
</tr>
<tr>
<td>T2</td>
<td>0.77</td>
<td>0.71</td>
<td>0.70</td>
<td>0.69 (0.60–0.77)</td>
<td>0.90 (0.80–0.96)</td>
<td>0.41 (0.28–0.55)</td>
<td>0.66</td>
<td>0.77</td>
<td>0.34</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.77</td>
<td>0.73</td>
<td>0.76</td>
<td>0.70 (0.61–0.77)</td>
<td>0.92 (0.82–0.97)</td>
<td>0.41 (0.28–0.55)</td>
<td>0.67</td>
<td>0.80</td>
<td>0.33</td>
</tr>
<tr>
<td>Expert 1</td>
<td>0.73</td>
<td>N/A</td>
<td>N/A</td>
<td>0.65 (0.56–0.73)</td>
<td>0.84 (0.73–0.91)</td>
<td>0.41 (0.28–0.55)</td>
<td>0.65</td>
<td>0.67</td>
<td>0.35</td>
</tr>
<tr>
<td>Expert 2</td>
<td>0.70</td>
<td>N/A</td>
<td>N/A</td>
<td>0.59 (0.50–0.68)</td>
<td>0.84 (0.73–0.91)</td>
<td>0.27 (0.16–0.40)</td>
<td>0.60</td>
<td>0.57</td>
<td>0.40</td>
</tr>
<tr>
<td>Expert 3</td>
<td>0.68</td>
<td>N/A</td>
<td>N/A</td>
<td>0.60 (0.51–0.68)</td>
<td>0.75 (0.63–0.84)</td>
<td>0.41 (0.28–0.55)</td>
<td>0.62</td>
<td>0.56</td>
<td>0.38</td>
</tr>
<tr>
<td>Expert 4</td>
<td>0.68</td>
<td>N/A</td>
<td>N/A</td>
<td>0.58 (0.49–0.67)</td>
<td>0.78 (0.66–0.86)</td>
<td>0.33 (0.21–0.47)</td>
<td>0.60</td>
<td>0.53</td>
<td>0.40</td>
</tr>
<tr>
<td>Radiomics</td>
<td>0.70</td>
<td>0.59</td>
<td>0.77</td>
<td>0.62 (0.52–0.70)</td>
<td>0.79 (0.68–0.88)</td>
<td>0.39 (0.26–0.53)</td>
<td>0.62</td>
<td>0.59</td>
<td>0.38</td>
</tr>
</tbody>
</table>

ROC AUC, area under ROC curve; PR AUC, area under precision-recall curve; TPR, true positive rate or sensitivity; TNR, true negative rate or specificity; Acc, accuracy; PPV, positive predictive value; NPV, negative predictive value; N/A, not applicable; 95% CI, 95% confidence interval.
0.28–0.55); expert 4 had a test accuracy of 0.58 (95% CI, 0.49–0.67), F1 score of 0.68, and sensitivity of 0.78 (95% CI, 0.66–0.86), and specificity of 0.33 (95% CI, 0.21–0.47).

Compared with a baseline zero rule algorithm, the ensemble deep learning model had a statistically significant higher test accuracy (0.70 vs. 0.56, $P = 0.004$). Compared with all experts averaged, the ensemble deep learning model had higher test accuracy (0.70 vs. 0.60, $P = 0.053$), sensitivity (0.92 vs. 0.80, $P = 0.017$), and specificity (0.41 vs. 0.35, $P = 0.450$). Compared with the radiomics model, the ensemble deep learning model had higher test accuracy (0.70 vs. 0.62, $P = 0.081$), sensitivity (0.92 vs. 0.79, $P = 0.012$), and specificity (0.41 vs. 0.39, $P = 0.770$). Figure 2 shows the precision recall curves of all models overlaid with expert performance. Expected calibration error of our ensemble deep learning model was 0.19 on the training set, 0.065 on the validation set, and 0.094 on the test set. Reliability diagrams of the ensemble model are plotted in Supplementary Fig. S6. t-SNE representation of the final dense layer of ResNet demonstrates good separation of malignant and benign lesions by the deep learning model when compared with histopathologic diagnosis (Fig. 3).

**Discussion**

Today, most patients who have a newfound kidney lesion either undergo biopsy or surgery. In this study, we developed a deep learning approach combining conventional MR images and clinical variables that demonstrate good accuracy in differentiating benign from malignant renal lesion. Our model was based on the ResNet architecture, which utilizes the concept of residuals connections between the
convolulational layers and allows models to be trained to much deeper depths while maintaining a low complexity (35). Our final ensemble model achieved higher accuracy in distinguishing benign from malignant renal lesions compared to experts and radiomics. Ideally, a deep learning model would be implemented as a tool for risk stratification, in helping clinicians and patients understand the risk of malignancy for a lesion. To this end, specificity is the most important statistic as with higher specificity, fewer people will receive unnecessary biopsies; our deep learning model had the highest specificity between experts and models. Sensitivity is also an important metric as an increase in sensitivity would lead to fewer false negatives which may provide false relief. To evaluate inter-segmenter variation, two additional radiologists segmented the renal lesions in the test set, achieving high Dice similarity score among the 3 segmenters. Consistently good accuracy, sensitivity and specificity of the model when evaluated on any of the 3 test set segmentations prove stability against intersegmenter variation and support potential for clinical applicability. When we evaluated our pipeline using data that the model had never seen before from a separate institution, we found that our model continued to be able to discriminate between benign and malignant lesions, suggesting that our model is able to generalize beyond a single institution. Our final ensemble test model is adequately calibrated and its expected calibration within expectation for a neural network–based model (49).

Preoperative differentiation between benign and malignant renal lesions using noninvasive imaging modalities is important for treatment planning but challenging. Several radiologic studies have used traditional machine-learning techniques, such as support vector machine and random forest, to distinguish benign from malignant renal lesions basing on CT radiomics (11–16). However, these radiomic features were low throughput and predefined by radiologists’ expert knowledge. Deep learning represents an improvement over radiomic as it can automatically extract high-throughput features and increase the accuracy needed for clinical utilization (50). Recently, Lee and colleagues combined deep features with hand-crafted features in a random forest classifier to differentiate 39 AML without visible fat from 41 clear cell RCC on abdominal CT. The best model combining hand-crafted features with deep features from AlexNet using texture image patches achieved an accuracy of 76.6% (25). The study is limited by the small cohort size. Zhou and colleagues performed a transfer learning model on CT with the Inception V3 model pretrained on the ImageNet dataset to differentiate 58 benign from 134 malignant renal lesions (26). Although the validation accuracy reported for the best model was 97%, the small cohort size and the lack of a true test set makes generalization questionable.

Compared with previous machine-learning studies, our study has several differences. First, we chose MRI instead of CT. MRI provides multi-parametric sequence, which theoretically provide information than simple attenuation differences measured in Hounsfield units on CT. Second, our cohort included all available benign and malignant renal lesions that were pathologically or radiographically diagnosed from five institutions with MRI, which is more representative of the real clinical scenario. In contrast, most of previous studies only included selected tumor types. Third, we compared the model accuracy to expert radiologist interpretation to assess the model’s clinical usefulness, which none of the previous studies has attempted.

In our study, the deep learning model outperformed the most optimized radiomic model in distinguishing benign and malignant renal tumors, with higher test accuracy, sensitivity, and specificity on the test set. Two main reasons likely account for the worse performance and lower reproducibility of radiomics when compared with deep learning: acquisition parameters and implementation. Radiomics can vary significantly based on acquisition parameters such as segmentation, size of lesion, and entropy of lesion and surrounding tissues. Given that our dataset contains lesions from multiple institutions, there is naturally variation in the model of MRI scanner used. Models of CT scanners have been found to influence reproducibility of radiomics features (19). For implementation, multi-user studies and studies for different radiomics software packages have shown inconsistency in results for the same kinds of analysis. Implementation

Figure 3.
TSNE transformed representation of the final layer of the T1C (A) and T2WI (B) neural network before the classification node for every image in the validation dataset color-coded as benign or malignant.
networks model is much easier to overfit, especially when trained with small numbers of data. Although we used data augmentation and early stopping to prevent overfitting of our model, a large amount of data is essential for training of deep neural networks. Although we included 1,162 renal lesions on MRI in our cohort, more than any previous machine learning study on the topic, algorithm development can benefit further from a larger patient cohort, especially considering the heterogeneity of different image acquisition parameters at different institutions. Second, only conventional MR sequences were used. Inclusion of advanced MR sequences such as diffusion-weighted imaging and arterial spin labeling may further increase model accuracy (52, 53). Third, manual segmentations were performed in this study by a radiologist with 5 years of experience reading abdominal MR. Automatic segmentation based on deep learning has been successfully performed in other organs (28, 54–56) and kidney volume on CT with comparable accuracy to expert segmentation (57). However, deep learning segmentation of renal lesion on MRI is challenging. The algorithm needs to distinguish the lesion of interest from other targets in the kidney (e.g., cyst), from other organs in the body which may have similar intensity and finally from background in small renal lesions. We have not identified any study in the literature that reported automatic segmentation of kidney lesion on MRI. Finally, the earliest phase of the postcontrast phases was selected for analysis. Compared with single phase, multiple phases can provide more information about the enhancement pattern of the lesion and further improve model accuracy. Automatic renal tumor segmentation on MRI and multi-phase analysis will be incorporated in future work.

In this study, we developed a deep learning-based model using conventional MR imaging to noninvasively classify benign and malignant renal tumors with good accuracy, sensitivity, and specificity comparable with experts and radiomics. If further validated, the model can spare patients unnecessary biopsy/surgery and help guide management in a clinical setting.

Disclosure of Potential Conflicts of Interest

T.P. Gade is an employee/paid consultant for Trisalus Life Sciences. S.W. Stavropoulos is an employee/paid consultant for Becton Dickinson, and reports receiving commercial research grants from Sillajen and Cook Medical. No potential conflicts of interest were disclosed by the other authors.

Authors’ Contributions

Conception and design: I.L. Xi, R.Y. Huang, P.J. Zhang, M.C. Soulen, H.X. Bai, S.W. Stavropoulos

Development of methodology: I.L. Xi, R. Wang, S. Purkayastha, K. Chang, R.Y. Huang, Y. Fan, H.X. Bai, S.W. Stavropoulos

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): I.L. Xi, Y. Zhao, R. Wang, M. Chang, A.C. Silva, P.J. Zhang, Z. Zhang, H.X. Bai, S.W. Stavropoulos

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): I.L. Xi, Y. Zhao, R. Wang, S. Purkayastha, K. Chang, R.Y. Huang, A.C. Silva, M. Vallières, Y. Fan, H.X. Bai

Writing, review, and/or revision of the manuscript: I.L. Xi, Y. Zhao, R. Wang, M. Chang, S. Purkayastha, K. Chang, R.Y. Huang, A.C. Silva, M. Vallières, P. Habibollahi, Y. Fan, T.P. Gade, M.C. Soulen, H.X. Bai, S.W. Stavropoulos

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): I.L. Xi, Z. Zhang, H.X. Bai, S.W. Stavropoulos

Study supervision: B. Zou, T.P. Gade, Z. Zhang, H.X. Bai, S.W. Stavropoulos

Other (code development): R. Wang

Acknowledgments

This study was supported by RSNA fellow research grant (RF1802), National Natural Science Foundation of China grant (81811011287), SIR Foundation Radiology Resident Research Grant, and National Cancer Institute of the National Institutes of Health under Award Number R03CA249534 to H.X. Bai. This project was supported by a training grant from the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under award number ST32EB11860 and by the National Cancer Institute (NCI) of the National Institutes of Health under Award Number F30CA239407 to K. Chang. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The study was supported by funding from the Nicole Foundation for Kidney Cancer Research to S.W. Stavropoulos. The authors acknowledge the help of Hui Liu (H.L.), Ting Huang (T.H.), Dehong Peng (D.P.), and Quandiang Shang (Q.S.) in evaluating all renal lesions in the test set; Lin Zhu (L.Z.) and Yeyu Cai (Y.C.) in segmenting all renal lesions in the test set; and Sukhdeep Khurana, Aidan McGirr, An Xie, and Jianbin Liu in data collection.

Received January 30, 2019, revised April 30, 2019; accepted January 10, 2020; published first January 14, 2020.

References

16. Lee HS, Hong H, Jung DC, Park S, Kim J. Differentiation of fat-poor angio-


44. Deriche A, Ammari S, Bateste M, Durand PB, Hasping T, Massard C, et al. Limi-