Kidney Cancer Staging using Deep Learning Neural Network: Comparing Models Trained on Whole Kidney with Cancer and Only the Cancer

Nathan Hadjiyski

Ann Arbor Pioneer High School, Ann Arbor, MI, US, natehadjr@gmail.com
University of Rochester, Rochester, NY, US, nhadjiys@u.rochester.edu

ABSTRACT

In kidney cancer, Stage 1 is an important threshold for the decision of organ preservation surgery versus chemotherapy and organ removal for higher stages. The purpose of this study is to compare the classification accuracy of two different Inception V3 deep learning neural networks, trained on cropped computer tomography (CT) images of either whole kidney containing cancer (DLNN-WK) or only the kidney cancer (DLNN-OC). The National Cancer Institute TCIA database provided anonymized 3D CT scans and clinical data from 227 patients for the training and testing of the DLNN-WK and DLNN-OC. The dataset was split into 48% training, 10% validation, and 42% testing sets. The area under the ROC curve (AUC) for the DLNN-WK was 0.96 for training, 0.88 for validation and 0.87 for test sets. The AUC for the DLNN-OC was of 0.97 for training, 0.91 for validation, and 0.90 for test sets. Both AI systems show promise for potentially assisting physicians in kidney cancer staging.

Keywords: kidney cancer staging, AI, deep learning, CT, whole kidney with cancer, only kidney cancer

1 INTRODUCTION

The American Cancer Society's most recent estimates for kidney cancer in the United States for 2021 are that about 76,080 new cases of kidney cancer (48,780 in men and 27,300 in women) will occur, and about 13,780 people (8,790 men and 4,990 women) will die from this disease [1]. It is critical for an oncologist to accurately determine the stage of the kidney cancer in order to provide necessary treatment for the affected patients. Decision of treatment is dependent on the stage of the cancer [2], which may result in organ preservation surgery for Stage 1 or chemotherapy and organ removal for Stages 2, 3, 4 [3, 4]. Incorrect staging could lead to under- or over-treatment, causing complications for the patient.

Previous studies used convolutional neural networks (CNN) for automatic localization and segmentation of kidneys on contrast enhanced CT scans [5] or to classify the subtypes of renal cell carcinoma (RCC), using drawn region of interest as inputs and biopsy results as labels [6]. In [7] CNNs were used to extract morphological features from histopathology images to predict RCC patient survival outcome. In [8] RCC staging was performed with a Learnable Image Histogram-Based deep learning neural network model to classify the likelihood of RCC stages 1 & 2 versus stages 3 & 4 in CT scans using 159 cases.

Recently a deep learning neural network (DLNN) was developed to predict Stage 1 versus higher stages of kidney cancer using cropped 3D computer tomography (CT) scans of the kidney cancer [9].

The current study will compare the classification accuracy of DLNN trained on cropped CT images of whole kidneys containing cancer (DLNN-WK) to the classification accuracy of DLNN trained on cropped CT images of only kidney cancers (DLNN-OC) [9] to predict Stage 1 versus higher stages of kidney cancer.

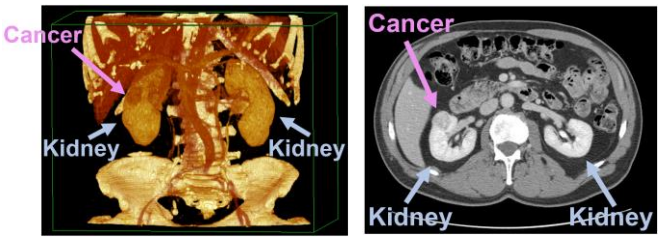
2 PURPOSE

The purpose of this study is (1) to develop a DLNN-WK system to predict kidney cancer Stage 1 versus higher stages using cropped CT scans containing the whole kidney with cancer; and (2) to compare the DLNN-WK to the DLNN-OC trained on cropped CT scans containing only the cancer.

3 MATERIALS AND METHODS

3.1 Data Sets

In order to keep continuity with the previous study [9], the same dataset, TCGA-KIRC, was used for the training, validation, and testing of the DLNNs. The dataset was obtained through the publicly available kidney cancer research database from the National Cancer Institute TCIA [10]. This dataset provided anonymized CT scans (Figure 1) and clinical data from 227 patients containing 231 kidney cancers of different stages.



(a) (b)

Figure 1: (a) 3D rendered volume of the renal area from a CT scan. (b) A slice of the CT scan of kidneys. The blue arrows point to the kidneys. The pink arrows point to the kidney cancer.

The kidney cancers were of Stages 1 to 4, determined by pathology. Examples of different stage kidney cancers on CT scans are presented in Figure 2. The CT scans where the cancer was best enhanced and visible were selected for processing. The slice thickness of the CT scans varied in the range of 1.25mm to 5mm. The dataset was used for training, validation and testing of the system.

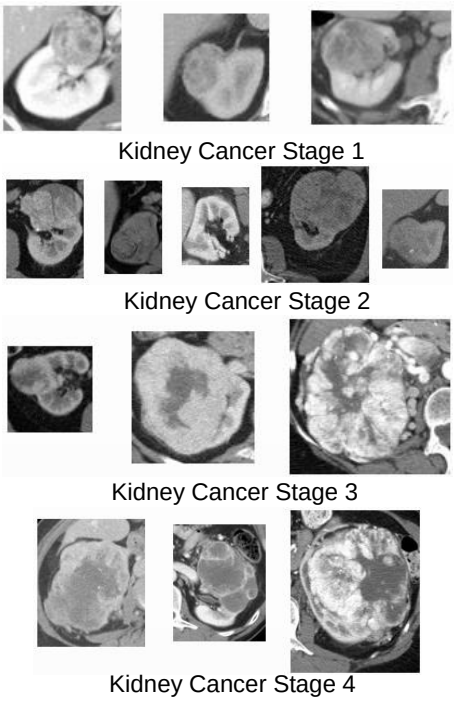


Figure 2: Comparison of kidney cancer CT scans from Stages 1 to 4 .

After downloading the anonymized 3D CT scans, the kidneys containing cancer were cropped from the 3D CT scans for all 227 patients using ImageJ, an image editing software for medical images (Figure 3). Approximately 7800 cropped CT kidney images were obtained (Table 1). The use of the cropped images had focused the DLNN training only on information related to kidney cancer (Figure 3).

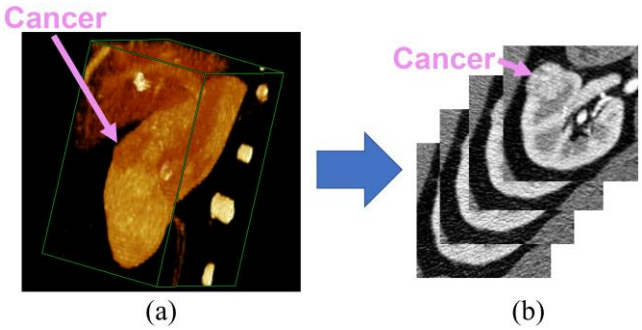


Figure 3: (a) Cropped kidney in a 3D rendered volume. (b) CT slices from the cropped kidney from (a). The pink arrows point to the kidney cancer.

The dataset was split on training, validation and test sets to ensure accurate and reliable evaluation results. The DLNN was trained using the training and validation sets. The trained DLNN was then tested on the test set.

It is difficult to distinguish different stages by the size of the cancer (Figure 2).

3.2 Network Structures

The Inception V3 deep learning network structure (IV3-DLNN), within the TensorFlow platform [11], was implemented to remain consistent with the study [9] used as a comparison.

The cropped CT image of the whole kidney with cancer was the input to the IV3-DLNN, where the likelihood of the kidney cancer being Stage 1 was the output (Figure 4).

The IV3-DLNN was first trained using the ImageNet dataset, consisting of more than 1,000,000 natural scene images. Transfer learning technique [12] was used to train the IV3-DLNN for the task of staging kidney cancer. The transfer learning is especially useful when the dataset for the target task is not very large. It retrains only part of the network structure with the target task at hand.

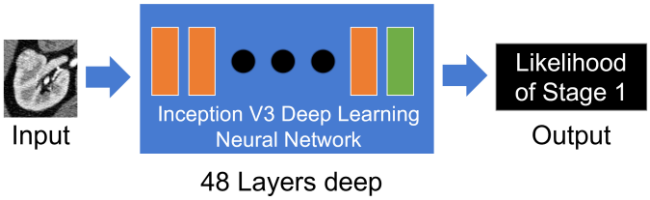


Figure 4: Flowchart of Inception V3 deep learning neural network (IV3-DLNN). The orange blocks represent the 48 layers of the IV3-DLNN. The green block represents the bottleneck layer (the final layer) of the neural network used to retrain IV3-DLNN.

The IV3-DLNN is retrained within the TensorFlow platform by updating the bottleneck layer (the final layer) of the neural network using the whole kidney with cancer cropped CT images to obtain DLNN-WK. The reason the final layer retraining can work for this new task is because the type of information incorporated in the network after training to distinguish between all the 1,000 classes in ImageNet is often also useful for distinguishing between the classes of the new task – in this case the different stages of kidney cancer.

The PC used contained an Intel Core-i5 3.8 GHz processor, 32GB of RAM, 2TB disk, and graphics processing unit (GPU) Nvidia 1080ti.

3.3 Training and Deployment

While training the DLNN-WK multiple learning rates were attempted, in the range of 0.001 to 0.05, until the training error (cross entropy) converged. After the neural network was trained it was later deployed on the validation set to select the best performing model, in relation to learning rate and number of iterations. Once the best model

was selected it was then deployed on the test set to obtain the final independent test results for analysis.

3.4 Evaluation

The classification performance is estimated using the area under the ROC curve (AUC) [13], where value of 1 represents perfect classification and value of 0.5 represents random choice or no classification at all. Accuracy of the prediction of Stage 1 versus Stages 2, 3, and 4 was also estimated. The accuracy was calculated by estimating the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) after applying a threshold to the DLNN likelihood scores of Stage 1.

4 RESULTS

The CT scan case images were split into 48% Training, 10% Validation, and 42% Testing, similar to the split in [9] for the DLNN-OC.

The CT scan cases were balanced so that there is a similar proportion to the cases in Stage 1 and Stage 2,3,4 for the training, validation and test sets (Table 1).

Datasets	Stage 1	Stages 2, 3, 4	*Total*
Training	56 (1938)	55 (1674)	*111 (3612)*
Validation	11 (532)	11 (229)	*22 (761)*
Test	51 (1899)	47 (1581)	*98 (3480)*
Total	*118 (4369)*	*113 (3484)*	***231 (7853)***

Table 1: Distribution of the Cancers (Images) in the datasets.

The DLNN-WK was trained up to 5000 iterations until the training error converged (Figure 5). It took 6 min and 35 sec to train the model for 5000 iterations.

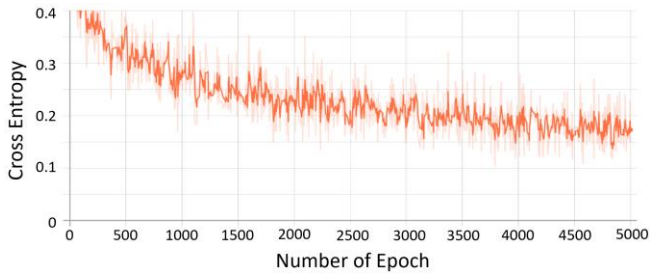


Figure 5: Convergence curve of the DLNN-WK training error (the cross entropy). The DLNN-WK was trained up to 5000 iterations with the kidney cancer training set.

The best result on the Validation set was obtained for the DLNN-WK trained for 4000 iterations with learning rate of 0.01. It took 5 min and 19 sec to train the model for 4000 iterations.

The AUC was 0.96 for the Training set, 0.88 for the Validation set and 0.87 for the Test set (Figure 6). The threshold of 0.63 was selected on the validation set, as it provided the most favorable results, and was applied on the DLNN-WK likelihood scores to calculate the accuracy. The accuracy was 93% for the Training set and 86% for the Validation set. The Test set provided an accuracy of 81% with TP=41, TN=38, FP=10, FN=9.

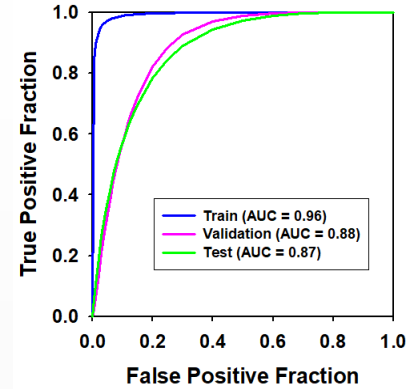
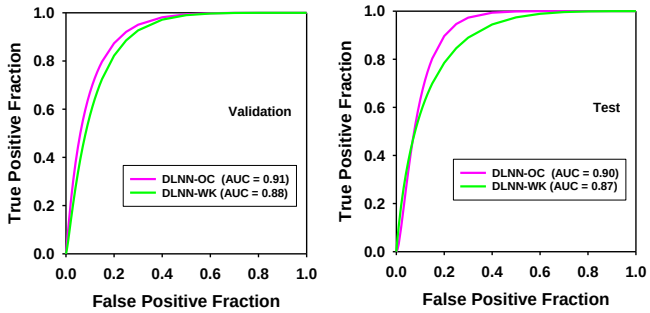


Figure 6: ROC curves for the Training, Validation and Test sets. The AUC was 0.96 for Training set, 0.88 for the Validation set and 0.87 for the Test set

5 COMPARISON OF DLNN MODELS

The performance of the DLNN-WK trained on the whole kidney with cancer provided very promising results as shown in Figure 6. When comparing these to the results obtained by the DLNN-OC trained on only cancer [9], we can notice that DLNN-OC has a higher AUC (0.90) and Accuracy (86%) on the test set than the DLNN-WK (AUC: 0.87, Accuracy: 81%) as seen in Figure 7, Table 2, and Table 3. However, the difference is relatively small. Examples of correct classification of kidney cancer of different stages are presented in Figure 8.



(a) (b)

Figure 7: Comparison of the ROC curves for the DLNN-WK and DLNN-OC (a) Validation set and (b) Test set.

DLNN Model	Training	Validation	Test
DLNN-WK (whole kidney)	0.96	0.88	0.87
DLNN-OC (only cancer)	0.97	0.91	0.90

Table 2: AUC Comparison of DLNN-WK and DLNN-OC

CT Scan Type	Training	Validation	Test
DLNN-WK (whole kidney)	93%	86%	81%
DLNN-OC (only cancer)	92%	86%	86%

Table 3: Accuracy Comparison, DLNN-WK and DLNN-OC

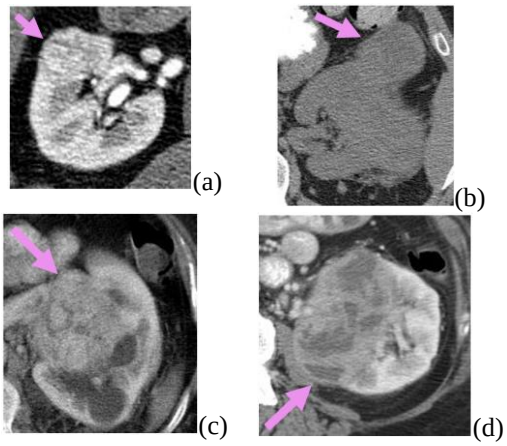


Figure 8: DLNN-WK and DLNN-OC correct classification of kidney cancer stages on test CT scans. A score of 1.0 corresponds to perfect classification of cancer Stage 1. A score of 0 corresponds to perfect classification of cancer Stages 2, 3, and 4. (a) Stage 1, DLNN-WK: 0.906, DLNN-OC: 0.934. (b) Stage 2, DLNN-WK: 0.044, DLNN-OC: 0.025. (c) Stage 3, DLNN-WK: 0.063, DLNN-OC: 0.036. (d) Stage 4, DLNN-WK: 0.024, DLNN-OC: 0.015. The pink arrows point to the kidney cancer.

It is important to also highlight that the DLNN sometimes incorrectly classifies the kidney cancer stages. An incorrect classification of Stage 1 kidney cancer test case is shown in Figure 9a. A Stage 2 test case of kidney cancer that was incorrectly classified is shown in Figure 9b.

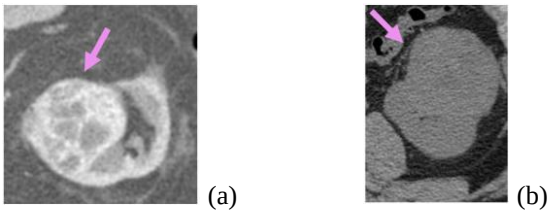


Figure 9: Kidney cancer CT scans, in Test set, incorrectly classified by both DLNNs. (a) Stage 1, DLNN-WK: 0.441, DLNN-OC: 0.209. (b) Stage 2, DLNN-WK: 0.611, DLNN-OC: 0.626. The pink arrows point to the kidney cancer.

6 CONCLUSION

Both DLNN systems, trained on either the whole kidney with the cancer or only the cancer, were able to accurately determine the kidney cancer Stage. The DLNN system trained with CT images containing only cancer was slightly more accurate than the DLNN system trained with the CT images of whole kidney with cancer. However, cropping the whole kidney requires less time and instruction, which may make it easier for the physician to use the DLNN system. It would be interesting in the future to see which method a physician would prefer and use consistently.

ACKNOWLEDGEMENTS

I would like to thank Dr. Alon Weizer for his help supervising this research continuation.

REFERENCES

[1] "Information and Resources about for Cancer: Breast, Colon, Lung, Prostate, Skin," American Cancer Society. Available: http://www.cancer.org/.

[2] L. Flegar , C. Groeben , R. Koch , M. Baunacke , A. Borkowetz , K. Kraywinkel , C. Thomas, and J. Huber, "Trends in Renal Tumor Surgery in the United States and Germany Between 2006 and 2014: Organ Preservation Rate Is Improving," Annals of Sugical Oncology, 2020; 27:1920–1928.

[3] S. C. Campbell, A. C. Novick, A. Belldegrun, et al. "Guideline for management of the clinical T1 renal mass," The Journal of Urology, 2009; 182:1271–1279.

[4] B. Ljungberg, K. Bensalah, S. Canfield, et al. "EAU guidelines on renal cell carcinoma: 2014 update," European Urology, 2015; 67:913–924.

[5] K. Yin, C. Liu, M. Bardis, J. Martin, H. Liu, A. Ushinsky, J. Glavis-Bloom, C. Chantaduly, D. S. Chow, R. Houshyar, and P. Chang, "Deep learning segmentation of kidneys with renal cell carcinoma," Journal of Clinical Oncology, 2019. Available: https://ascopubs.org/doi/abs/10.1200/JCO.2019.37.15_suppl.e16098.

[6] S. Han, S. Il Hwang, and H. J. Lee, "The Classification of Renal Cancer in 3-Phase CT Images Using a Deep Learning Method," Journal of Digital Imaging, 2019; 32:638–643.

[7] S. Tabibu, P. K. Vinod, and C. V. Jawah, "Pan-Renal Cell Carcinoma classifcation and survival prediction from histopathology images using deep learning," Scientific Reports, 2019; 9:10509.

[8] M. A. Hussain , G. Hamarneh, and R. Garbi, "Renal Cell Carcinoma Staging with Learnable Image Histogram-Based Deep Neural Network," Machine Learning in Medical Imaging, 2019; 11861:533-540.

[9] N. Hadjiyski, "Kidney Cancer Staging: Deep Learning Neural Network Based Approach," Proc. 8th IEEE International Conference on e-Health and Bioengineering (EHB), 2020, pp. 1-4, doi: 10.1109/ EHB50910.2020.9280188.

[10] "Cancer Imaging Archive Wiki," TCGA-KIRC - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki. Available: https://wiki. cancerimagingarchive.net/display/Public/TCGA-KIRC.

[11] "Advanced Guide to Inception v3 on Cloud TPU | Google Cloud," Google. Available: https:// cloud.google .com/tpu/docs/inception-v3-advanced.

[12] "Retraining an Image Classifier : TensorFlow Hub," TensorFlow. Available: https://www.tensorflow.org/hub/ tutorials/image_retraining.

[13] Dogan, John C. "Computing a ROC Curve with Python." *Medium.com*, Medium, 17 Sept. 2018, <https://medium.com/.../computing-an-roc-graph-with-python-a3aa20b9a3fb>.