ColorNephroNet: Kidney tumor malignancy prediction using medical image colorization

Aleksander Obuchowski[1,2], Barbara Klaudel[1,3], Roman Karski[1,2,3], Bartosz Rydziński[1,2], Mateusz Glembin[1,5], Paweł Syty[1,2,4], Patryk Jasik[1,2,4]

[1]Radiato.ai Gdańsk University of Technology: [2]Faculty of Applied Physics and Mathematics,
[3]Faculty of Electronics, Telecommunications and Informatics, [4]BioTechMed Center
[5]COPERNICUS, St. Adalbert's Hospital Gdańsk, Department of Urology
{aleksander.obuchowski, barbara.klaudel, roman.karski, bartosz.rydziński}@student.pg.edu.pl, mglembin@copernicus.gda.pl,
{ pawel.syty,partyk.jasik}@pg.edu.pl,

Abstract

Renal tumor malignancy classification is one of the crucial tasks in urology, being a primary factor included in the decision of whether to perform kidney removal surgery (nephrectomy) or not. Currently, tumor malignancy prediction is determined by the radiological diagnosis based on computed tomography (CT) images. However, it is estimated that up to 16% of nephrectomies could have been avoided because the tumor that had been diagnosed as malignant, was found to be benign in the postoperative histopathological examination. The excess of false-positive diagnoses results in unnecessarily performed nephrectomies that carry the risk of periprocedural complications. In this paper, we present a machine-aided diagnosis system that predicts the tumor malignancy based on a CT image. The prediction is performed after radiological diagnosis and is used to capture false-positive diagnoses. Our solution is able to achieve a 0.84 F1-score in this task. We also propose a novel approach to knowledge transfer in the medical domain in terms of colorization based pre-processing that is able to increase the F1-score by up to 1.8pp.

Introduction

Renal cancer is a serious disease with a predicted incidence of more than 73,000 new cases in America in 2020 and its number increasing every year (Cokkinides 2020). Currently, the decision on whether to perform kidney removal surgery (nephrectomy) or segmentectomy (surgery to remove part of a kidney or tumor) is primarily based on radiological diagnosis of the tumor. In most cases, the decision is reduced to identifying whether the tumor is malignant or benign, based on, among others, its density and type of attenuation observed in CT images. However, 13%-16% of removed tumors can in fact be benign despite having been marked as malignant during radiological diagnosis (Kay and Pedrosa 2018). With the increased false-positive malignancy prediction comes the lack of the ability to accurately assess the benefit-risk ratio between conducting nephrectomy or segmentectomy and leaving the tumor under observation. If a tumor turns out to be malignant, it increases the risk of further surgeries and can even cause a patient's death. If, however, it is indeed benign, it is often safer not to perform the operation and leave the tumor intact. This is especially the case with elderly patients, as the risk of the surgery increases and potential time for the tumor to grow decreases with age. A study has shown that for patients over the age of 65, radical nephrectomy was significantly associated with death from any cause (Thompson et al. 2008). Those patients are in fact the majority of renal tumor cases; therefore, there is an especially high need for limiting the false-positive malignant tumor diagnoses.

In our work, we show how reduction of false-positive predictions can be achieved using a deep learning-based solution, where the model's intent is to serve as a second opinion system employed in addition to the radiological diagnosis. Such a model, providing its high specificity, should function as a stimulus to raise doctors' awareness of cases that might be misclassified as malignant and, therefore, limit the number of false-positive examples. Its role is to reassure doctors in their diagnosis or warn them of a possible mistake. A proposed implementation of the system is shown in Figure 1. Physicians can subsequently perform additional tests, such as biopsy, or seek other experts' opinion on the case to confirm or contradict the diagnosis. Biopsy, although considered a gold standard in renal tumor classification, is also associated with an additional risk and extra costs (Yu et al. 2017).

In the paper, we present a deep learning model trained to distinguish malignant and benign tumors based on the CT image. The model is able to achieve the accuracy of 86%, while also having high recall. We also test the performance of popular pre-trained neural networks in the tumor malignancy prediction task. Furthermore, we show that image colorization results in better knowledge transfer between pretraining and fine-tuning phases, improving accuracy in the medical image classification task.

Related work

Deep learning has shown a promising performance in the field of biomedical imaging. With the release of an open-source dataset KiTS19 (Heller et al. 2019), we have observed a growing interest in kidney tumor classification and segmentation. The majority of approaches involved deep learning with convolutional neural networks pretrained on ImageNet dataset (Deng et al. 2009), however, some studies used methods such as adaptive neuro fuzzy inference system (ANFIS) (Nikita, Sadawarti, and Singla 2020) or random

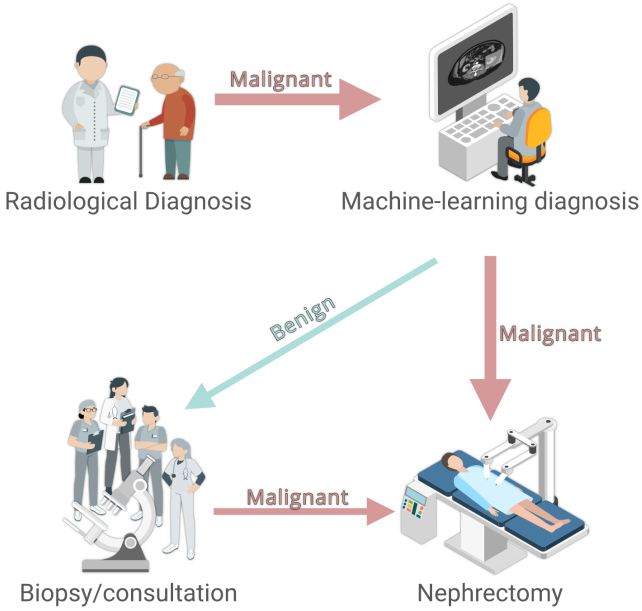


Figure 1: Proposed implementation of the system.

forest (Erdim et al. 2020). The task of kidney tumor malignancy identification has been narrowed to the classification of only specific subtypes in some works, e.g. clear cell renal cell carcinoma (ccRCC) / oncocytoma (Nikpanah et al. 2021), chromophobe RCC (chRCC) / oncocytoma (Baghdadi et al. 2020), ccRCC / chRCC / papillary RCC (pRCC) (Han, Hwang, and Lee 2019).

In our work, we employ colorization, which is rare for a medical deep learning pipeline. However, image colorization was implemented as a preprocessing step before segmentation in (Attique et al. 2012) and (Khan, Gotoh, and Nida 2017). Authors show that colorized medical images improve the contrast of anatomical structures and therefore facilitate precise segmentation and convey more precise anatomical information.

The key motivation for our study was decreasing the number of false-positive malignancy cases. An example of a study tackling a similar problem – achieving high recall rate in mammography is (Aboutalib et al. 2018). In this paper, the authors introduce a new class of "recalled-benign" tumors that represent cases referred for biopsy (i.e. previously assumed malignant) that were consequently labeled benign. It is also worth mentioning that the authors, despite the introduction of the third class, treat this as a binary classification problem. In our research, we used tumors from nephrectomy cases along with cases that did not require nephrectomy, so our classes can be seen as "true-benign", "recalled-benign" and "true-malignant".

Dataset

We collected 15485 CT images coming from 383 individual cases. These data came from two sources. 173 of the cases were collected by us, using historical data of the patients that had undergone the nephrectomy. Every single case, in addition to CT images, was paired with histopathological results from the postoperative biopsy. The usage of histopathological data instead of radiological diagnosis was motivated by its greater accuracy and reliability as it is considered the gold standard in renal tumor identification. Moreover, using data from patients on whom nephrectomy had already been performed, allowed us to focus on the false-positive results of radiological diagnosis, and therefore teach the network to correctly label such cases. We collected images from different phases of CT study, including non-enhanced, arterial, nephrogenic, and delayed. After the selection of the cases, each of them was anonymised, cropped, and labeled using a custom-build system. The crop size was selected to be 130x130 pixels as this area was considered to be sufficient to capture a kidney and a tumor. The example of an image and its cropped area is given in the Figure 2. In addition to the histopathological subtype, each image was labeled with the tumor size and patient's weight and height.

Figure 2: Visualization of the area cropped from full abdominal CT image.

To increase the amount of data, we also added 210 cases from the openly available dataset KiTS19 (Heller et al. 2019). This dataset was primarily designed for semantic segmentation challenge; however, in addition to the segmentation data, the authors have provided surgical outcome of the cases, including histopathological subtype of the tumors. Those images were collected only from the arterial phase of CT study. Since the images contained overlay data for tumors and kidneys, we chose the center of the tumor overlay as the center of the 130x130 crop to adjust it to our format. Detailed dataset description, distinguishing histopathological subtypes has been provided in the Table 1.

Next, we grouped the histopathological subtypes together into malignant and benign binary classes - ccRcc, chRcc and pRcc tumors were marked as malignant, and oncocytoma, angiomyolipoma (AML) and bening-other tumors were marked as benign.

As the KiTS19 dataset contained only the arterial phase of the study, we decided to use the very same phase during training and classification. This was also motivated by the fact that the arterial phase shows attenuation of tumors and, therefore, is suitable for malignancy prediction.

Tumor type	No. of cases	No. of images
ccRCC	214	10193
chRCC	26	1590
pRCC-type-1	10	488
pRCC-type-2	3	324
pRCC	27	769
malignant-other	1	75
oncocytoma	20	702
AML	78	1221
benign-other	4	123
malignant tumors	281	13439
benign tumors	102	2046
total	383	15485

Table 1: Dataset with respect to different phases.

In our baseline experiments, we use a single 2D slice per case where the visible tumor area is the largest (as such images are the best reference for the case). We test the effect of using all the available 2D slices per case, utilizing the full dataset.

Our Solution

Pre-processing

The images themselves were firstly processed in DICOM format, where image data is presented as a 2D array of Hounsfield units (HU). Those units, ranging from -1024 to 3071, represent the attenuation coefficient measurement (with respect to water and air) during a CT scan 1.

$$HU = 1000 \times \frac{\mu_x - \mu_{water}}{\mu_{water} - \mu_{air}} \quad (1)$$

In line with the current standard for viewing abdominal CT scans, we cropped this range with a window center of 60 HU and window width of 400 HU. After cropping the values, we scaled them to represent grayscale pixel values ranging from 0 to 255. The images were also resized to 256x256 pixels to fit the size of popular pre-trained architectures.

Colorization

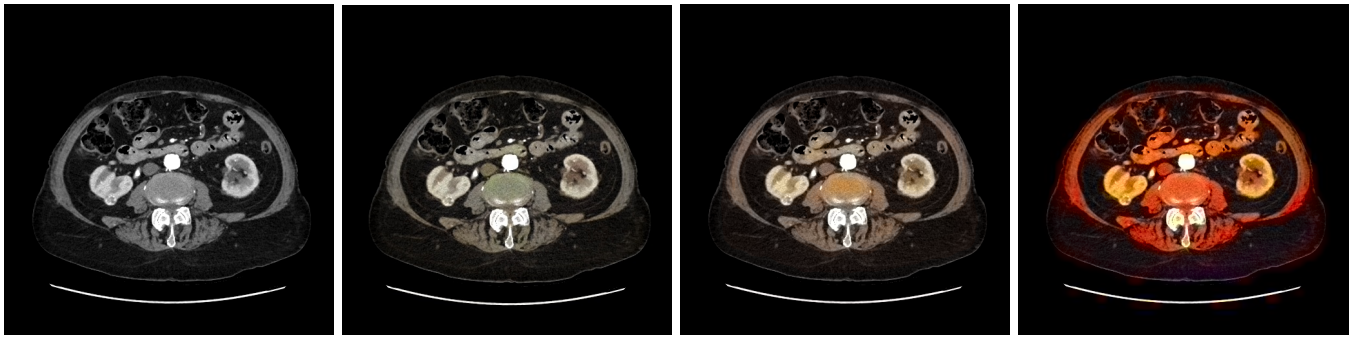
Currently, ImageNet is the most popular dataset for pre-training large convolutional models. Results show that despite the significant differences in modality, models pre-trained on ImageNet can still achieve better results in medical image classification than models trained solely for this task. However, medical images such as CT, MRI or X-ray are processed in grayscale format and must be converted to a 3-channel format in order to be processed by a pre-trained network. The most common technique for doing this is to copy the grayscale values across different channels. This can, however, lead to a sub-optimal utilization of filters learned from color images in transfer learning (Xie and Richmond 2018). In our solution, we tackled this problem by pre-processing the images using colorization models. Those models deal with image-to-image prediction problems by reconstructing the image in RGB color space based on its grayscale equivalent. Most of the popular models are based on pre-trained image classification models that are later adopted to the colorization task by conversion to fully convolutional networks.

In the initial experiments, we tested 3 popular open source image colorization models. The first model — Let there be Color (LTBC) (Iizuka, Simo-Serra, and Ishikawa 2016) uses an end-to-end network that jointly learns global and local features of an image. This is done through utilization of 2 convolutional networks — one for detection of global features and the other for detection of local features. These network outputs are then concatenated and used by the decoder network to produce colorized versions of the image. The second model, described in Learning Representations for Automatic Colorization (LRAC) (Larsson, Maire, and Shakhnarovich 2016), uses architecture based on a deep convolutional network — VGG16. It takes spatially localized multi-layer slices as per-pixel description, predicting chroma distribution of pixels given its hypercolumn descriptor. The third model — Colorful Image Colorization (CIC) (Zhang, Isola, and Efros 2016) uses a VGG-styled network with added depth and dilated convolutions to map a grayscale image to its colored version. It is also noteworthy that this version of the network has no pooling layers and changes in resolution are achieved through spatial downsampling or upsampling between convolutional blocks.

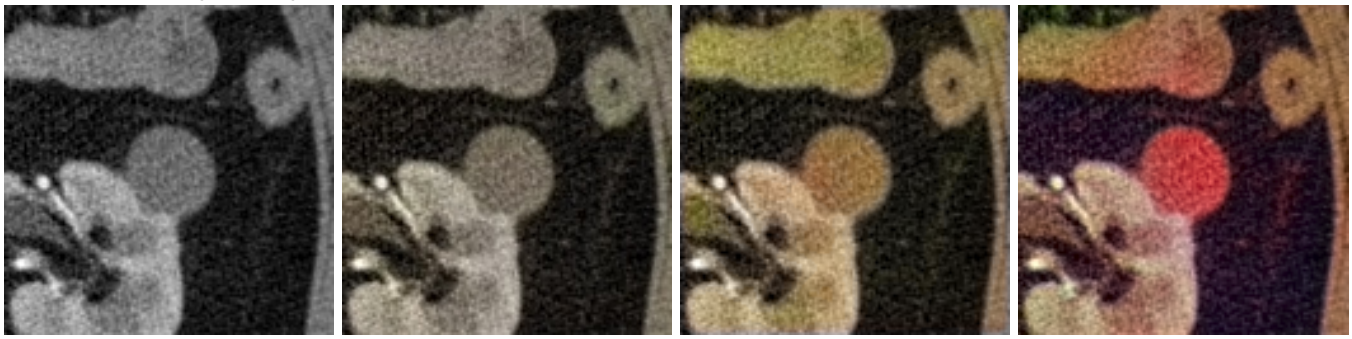
Results of colorization of an abdominal CT scan using different methods are shown in Figure 3. Figures 3d and 3e show the original image in grayscale format, 3b, 3f show the result of colorization using Let There Be Color method, 3c, 3g show results of Learning Representations for Automatic Colorization model, while 3d, 3h show the output of Colorful Image Colorization model. Having tested different solutions through visual examination of how well they improve the contrast of anatomical structures, we decided to use the Colorful Image Colorization framework. The authors claim that their model produces colorization that is more vibrant and perceptually realistic than the other approaches. This can be particularly useful in our case, as seen in the Figure 3. Other image colorization models produce results that are much less vibrant and, therefore, there is no clear semantic separation between different organs. Figure 3d shows that this model colorizes the image in a way that clearly distinguishes kidney (orange), bones (white) and other organs (red), while Figure 3h shows a separation between the tumor(red) and the rest of the kidney. This is likely due to the fact that authors show that their solution is suitable not only for colorization but also for semantic segmentation task - what we show is also true for the medical domain. In the section , we show that this improves the classification performance.

Architectures

Pre-trained networks. In our initial experiments, we tested architectures using popular pre-trained convolutional networks, including VGG16 (Simonyan and Zisserman 2014) networks that replace large kernel-sized filters with multiple subsequent 3×3 kernel-sized filters, Xception (Chollet 2017) based on pointwise convolution followed by



(a) Abdomen: Original Image (b) Abdomen: LTBC (c) Abdomen: LRAC (d) Abdomen: CIC



(e) Tumor: Original Image (f) Tumor: LTBC (g) Tumor: LRAC (h) Tumor: CIC

Figure 3: Comparison of different colorization models: Let There be Color! (Iizuka, Simo-Serra, and Ishikawa 2016) (b, f), Learning Representations for Automatic Colorization (Larsson, Maire, and Shakhnarovich 2016) (c, g) and Colorful Image Colorization (Zhang, Isola, and Efros 2016) (d, h). Models based on (b, f) and (c, g) show little color difference between various parts of the CT scan, while the results based on Colorful Image Colorization (d) show clear distinction between bone (white), kidney (orange) and the rest of the organs (red). In Figure (h), there can also be observed a separation between tumor (red) and the rest of the kidney.

a depthwise convolution, ResNet (He et al. 2015) based on deep residual connections and DenseNet (Iandola et al. 2014) based on connecting each layer to every other layer in a feed-forward fashion.

All those models were pre-trained on ImageNet before the fine-tuning task of tumor malignancy prediction. Additionally, we also used the CheXNet (Rajpurkar et al. 2017) model based on DenseNet architecture and pre-trained on NIH dataset containing 112 120 frontal-view X-ray images that has also been proven to achieve accuracy in medical image analysis higher than models pre-trained on natural images.

In all cases, the model's classification layers were replaced by a custom unified classifier described in the following section.

Classification layers. In each case, the base networks were followed by the global average pooling layer with dropout of 0.2 and batch normalization. Next, the inputs were fed to two dense layers with 2048 neurons each and ReLU activation function. Those layers were also regularized with L2 type regularization. Subsequently, we applied another dropout of 0.2. The dropouts were crucial in the architectures in order to prevent the overfitting of the network with the class imbalance in the dataset. Ultimately, after the two hidden layers, there was a final classification layer with softmax activation function to map the inputs into tumor malignancy. The visualization of the architecture is shown in Figure 4.

We used a binary cross-entropy loss function as each image should be mapped to exactly one tumor type. For the optimization, we used Adam algorithm (Kingma and Ba 2017). To deal with the class imbalance problem, the disproportionate amount of malignant tumors in the dataset, we applied class weight of 0.1 to the malignant cases. This has prevented the network from over-fitting and additional bias towards malignant tumors.

Experiments and Results

Dataset

Motivated by the limited size of our dataset, to test our solution we constructed a 5-fold cross-validation set stratified by the tumor malignancy. Cases were chosen from the historical data of patients that underwent nephrectomy based on the radiological diagnosis of the tumor being malignant. Each of the cases was paired with the postoperative histopathological results (considered a gold standard in renal tumor prediction) dictating whether the tumor was, as previously assumed, malignant, or in fact benign.

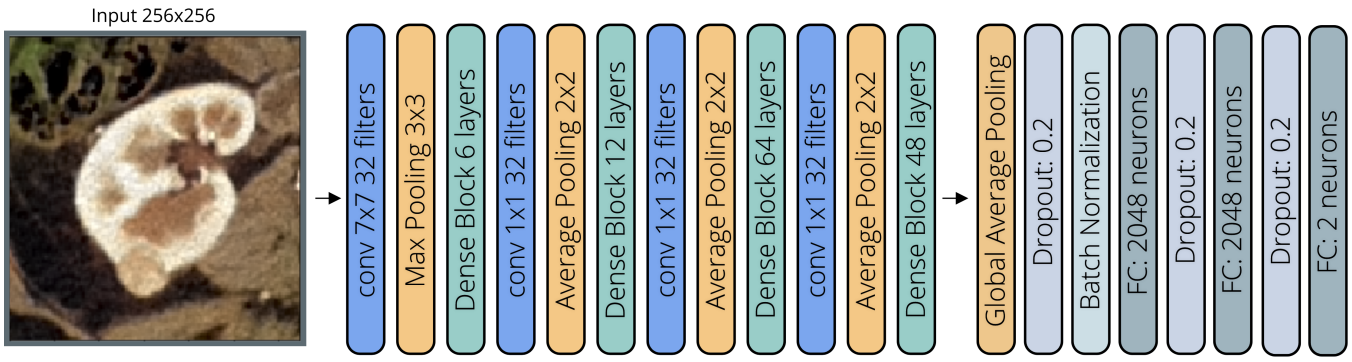


Figure 4: Architecture of the network. DenseNet encoder is followed by a classification block consisting of global average pooling, 2 hidden layers and the classification layer mapping the output to binary/malignancy prediction.

Pre-trained networks

In the initial experiments, we tested the performance of popular pre-trained architectures described in the previous section. In each case, the hyper-parameters of the networks were identical as described in the previous section. The results of the evaluation are shown in Table 2. The results show

Architecture	F1-score
CheXNet	0.7233
ResNetV2	0.769
Xception	0.7772
VGG16	0.8011
DenseNet	**0.8046**

Table 2: Comparison of the pre-trained models.

that the best pre-trained network turned out to be DenseNet, achieving the F1-score of 80%. Another noteworthy fact is that CheXNet achieved the lowest score of all tests of pre-trained encoders indicating that knowledge transfer from chest X-ray images was not beneficial over Image-Net, despite the fact that X-ray images and CT scans might be considered more similar than CT scans and natural images from ImageNet.

Colorization

Based on the experiments described in the previous subsection, we chose DenseNet-based network as the baseline for further experiments. Comparing its performance with and without the colorization pre-processing (described in section 3), we can see in Table 3 that colorization improves the F1 score by 1.8 pp.

Model	F1-score
DenseNet (without colorization)	0.8046
DenseNet (with colorization)	**0.8228**

Table 3: Effect of image colorization

Adding additional slices per tumor

Additionally, we also tested the effect of using additional 2D slices per single case in the training phase. This can be seen as an augmentation method where instead of providing the network with one reference image per case we use multiple CT slices per case depicting the tumor from different depths. This increased our dataset size from 383 images to 15485 images. To test the effect of adding additional slices, in the testing phase we used a single image per case where the tumor is best visible, similarly to previous subsections.

Model	F1-score
DenseNet (with colorization, single slice)	0.8228
DenseNet (with colorization, all slices)	**0.8444**

Table 4: The effect of adding additional CT slices per case.

In Table 4, we can see that providing the networks with additional CT slices increases its F1-score by up to 2.2 pp.

Final Solution

For the final solution, based on the results obtained in sections above, we chose the pre-trained DenseNet network fine-tuned on the colorized CT images from the full dataset. This network is able to achieve 0.84 F1-score, 0.86 accuracy, 0.79 precision and 0.86 recall. The high recall is especially important as it depicts the model's ability to recognize cases misclassified in the initial radiological diagnosis.

Conclusions and Future Work

In the paper, we present a deep learning model for kidney tumor malignancy classification. This model's role is to serve as a second opinion system, catching incorrectly classified malignant tumors in order to reduce the number of unnecessary surgeries. We show that medical image colorization is able to increase the F1-score up to 1.8pp by improving the knowledge transfer from pre-trained networks. We also show using additional CT slices in the training phase can be beneficial to the network's performance improving its F1-score by up to 2.2 pp.

Although our research is limited by the fact that our solution is shown working in pair-with human diagnosis and additional research would need to be done to test it in a stand-alone fashion and compare it to radiological diagnosis directly, we show that such a system achieving high recall score is suitable for post-radiological diagnosis reevaluation.

In the future, we are going to study further medical image colorization and its influence on medical image classification. We are also planning to extend the machine-learning pipeline with segmentation pre-processing allowing us to use multiple CT slices in the prediction phase which would enable us to facilitate majority-voting based methods that could further improve the network's accuracy.

References

Aboutalib, S. S.; Mohamed, A. A.; Berg, W. A.; Zuley, M. L.; Sumkin, J. H.; and Wu, S. 2018. Deep learning to distinguish recalled but benign mammography images in breast cancer screening. *Clinical Cancer Research* 24(23):5902–5909.

Attique, M.; Gilanie, G.; Mehmood, M. S.; Naweed, M. S.; Ikram, M.; Kamran, J. A.; Vitkin, A.; et al. 2012. Colorization and automated segmentation of human t2 mr brain images for characterization of soft tissues. *PloS one* 7(3):e33616.

Baghdadi, A.; Aldhaam, N. A.; Elsayed, A. S.; Hussein, A. A.; Cavuoto, L. A.; Kauffman, E.; and Guru, K. A. 2020. Automated differentiation of benign renal oncocytoma and chromophobe renal cell carcinoma on computed tomography using deep learning. *BJU Int* 125(4):553–60.

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251–1258.

Cokkinides, V., A. J. S. A. e. a. 2020. American cancer society: Cancer facts and figures.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Erdim, C.; Yardimci, A. H.; Bektas, C. T.; Kocak, B.; Koca, S. B.; Demir, H.; and Kilickesmez, O. 2020. Prediction of benign and malignant solid renal masses: machine learning-based ct texture analysis. *Academic radiology* 27(10):1422–1429.

Han, S.; Hwang, S. I.; and Lee, H. J. 2019. The classification of renal cancer in 3-phase ct images using a deep learning method. *Journal of digital imaging* 32(4):638–643.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition.

Heller, N.; Sathianathen, N.; Kalapara, A.; Walczak, E.; Moore, K.; Kaluzniak, H.; Rosenberg, J.; Blake, P.; Rengel, Z.; Oestreich, M.; et al. 2019. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*.

Iandola, F.; Moskewicz, M.; Karayev, S.; Girshick, R.; Darrell, T.; and Keutzer, K. 2014. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*.

Iizuka, S.; Simo-Serra, E.; and Ishikawa, H. 2016. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG)* 35(4):1–11.

Kay, F. U., and Pedrosa, I. 2018. Imaging of solid renal masses. *Urologic Clinics* 45(3):311–330.

Khan, M. U. G.; Gotoh, Y.; and Nida, N. 2017. Medical image colorization for better visualization and segmentation. In *Annual Conference on Medical Image Understanding and Analysis*, 571–580. Springer.

Kingma, D. P., and Ba, J. 2017. Adam: A method for stochastic optimization.

Larsson, G.; Maire, M.; and Shakhnarovich, G. 2016. Learning representations for automatic colorization. In *European conference on computer vision*, 577–593. Springer.

Nikita, B. K.; Sadawarti, H.; and Singla, J. 2020. A neuro-fuzzy based intelligent system for diagnosis of renal cancer. *International Journal of Scientific and Technology Research* 9(1).

Nikpanah, M.; Xu, Z.; Jin, D.; Farhadi, F.; Saboury, B.; Ball, M. W.; Gautam, R.; Merino, M. J.; Wood, B. J.; Turkbey, B.; et al. 2021. A deep-learning based artificial intelligence (ai) approach for differentiation of clear cell renal cell carcinoma from oncocytoma on multi-phasic mri. *Clinical Imaging* 77:291–298.

Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Thompson, R. H.; Boorjian, S. A.; Lohse, C. M.; Leibovich, B. C.; Kwon, E. D.; Cheville, J. C.; and Blute, M. L. 2008. Radical nephrectomy for pt1a renal masses may be associated with decreased overall survival compared with partial nephrectomy. *The Journal of urology* 179(2):468–473.

Xie, Y., and Richmond, D. 2018. Pre-training on grayscale imagenet improves medical image classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 0–0.

Yu, H.; Scalera, J.; Khalid, M.; Touret, A.-S.; Bloch, N.; Li, B.; Qureshi, M. M.; Soto, J. A.; and Anderson, S. W. 2017. Texture analysis as a radiomic marker for differentiating renal tumors. *Abdominal Radiology* 42(10):2470–2478.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European conference on computer vision*, 649–666. Springer.