Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi
International Journal of Engineering Research and Development

UMAGD, (2022) 14(3), s268-s275.

10.29137/umagd.1215882

Cilt/Volume:14 Sayı/Issue:3 Aralık/December 2022 Özel Sayı / Special Issue

Araştırma Makalesi / Research Article

Adrenal Lesion Classification on T1-Weighted Abdomen Images with Convolutional Neural Networks

Ahmet Solak[1], **Rahime Ceylan**[1], **Mustafa Alper Bozkurt**[2] **Hakan Cebeci**[2], **Mustafa Koplay**[2]

[1]Konya Teknical University, Faculty of Engineering and Natural Sciences, Department of Electrical-Electronics Engineering,, 42250, Konya
[2]Selcuk University, Faculty of Medicine, Department of Radiology, 42250, Konya

***Başvuru/Received**: 07/12/2022* ***Kabul / Accepted**: 21/12/2022* ***Çevrimiçi Basım / Published Online:** 31/12/2022*
***Son Versiyon/Final Version**: 31/12/2022*

Abstract

Adrenal lesions are usually discovered incidentally during other health screenings and are usually benign. However, it is vital to take precautions when a malignant adrenal lesion is detected. Especially deep learning models developed in the last ten years give successful results on medical images. In this paper, adrenal lesion characterization on T1-weighted magnetic resonance abdomen images was aimed using convolutional neural network (CNN) which is one of the deep learning methods. Firstly, effects of important model parameters are assessed on performance of CNN, so optimum CNN model is obtained for classification of adrenal lesions. For a fixed number of convolution filters determined in the first stage of the study, CNN model implemented by different kernel sizes were trained. According to the best result obtained, this time the kernel size was kept constant, and experiments were made for different filter numbers. Finally, studies were carried out with CNN structures of different depths and the results were compared. As a result of the studies, when filter is selected as [5 20], the best results in the trainings conducted with a single-block CNN structure are obtained 0.97, 0.90, 0.98, 0.90, 0.90, and 0.94, for accuracy, sensitivity, specificity, precision, F1-score, and AUC score, respectively. The study was compared with the studies in the literature, and it was seen that it was superior to them.

Key Words

“Abdomen, Adrenal Lesion, Classification, Convolutional Neural Network, Deep Learning, Magnetic Resonance”

***Responsible Author:** asolak@ktun.edu.tr*

1. Introduction

Adrenal incidentalomas are masses found incidentally in the adrenal glands during examinations for the diagnosis of different diseases other than adrenal hormone diseases (e.g., Cushing's or Conn's syndrome). These are benign lesions, the majority of which are adrenal adenomas, and are clinically insignificant. However, lesions suspected to be malignant require further evaluation for patient health (Fassnacht et al., 2016). At this stage, cross-sectional imaging techniques (e.g., computed tomography, magnetic resonance) are of great importance in the separation of lesions.

The use of computer aided diagnostic (CAD) systems on medical images has increased considerably, especially in the last decade, with the developments in artificial intelligence and deep learning. It helps experts in the detection and diagnosis of diseases in different organs such as the brain (Alex, KP, Chennamsetty, & Krishnamurthi, 2017; Chen, Dou, Yu, Qin, & Heng, 2018; Moeskops, Veta, Lafarge, Eppenhof, & Pluim, 2017), colon (Kang & Gwak, 2019; Q. Li et al., 2017; Nguyen & Lee, 2018), chest (Albarqouni et al., 2016; Dhungel, Carneiro, & Bradley, 2015; Guan & Loew, 2017). Considering the human factors such as excessive workload of radiologists and accumulated fatigue, CAD systems helping specialists are very important for both the diagnosis of the disease and the health of the patient.

Classification of adrenal lesions on abdominal images has become possible with developed machine learning and deep learning algorithms. (Li, Guindani, Ng, & Hobbs, 2017) classified adrenal masses from 230 abdominal CT images (121 benign and 109 malignant). They extracted the features using gray level co-occurrence matrix and used Bayesian probability model for classification. As a result of the study, the classification accuracy was obtained as 0.80. (Romeo et al., 2018) performed texture analysis using 60 MR examinations, including 20 lipid-rich adenomas, 20 lipid-poor adenomas and 20 non-adenoma adrenal lesions. They achieved 0.80 classification accuracy with the J48 classifier. (Elmohr et al., 2019) performed a binary classification study from a dataset containing 54 CT images (25 adrenal adenomas, 29 adrenal carcinomas). In the study, they compared performance of logistic regression and random forest. The best results in studies performed with logistic regression were calculated as accuracy, specificity, sensitivity and AUC scores of 0.82, 0.83, 0.81 and 0.89, respectively. (Koyuncu, Ceylan, Asoglu, Cebeci, & Koplay, 2019) studied the subtype characterization of adrenal tumors in 114 abdominal CT images. In the study in which the performances of different feature extraction and classification algorithms were compared, the best results were found for accuracy, sensitivity, specificity, and area under curve (AUC) as 0.80, 0.75, 0.82 and 0.78, respectively. (Liu et al., 2022) used a dataset of 280 CT images to distinguish between lipid-poor adenoma and adrenal pheochromocytoma. They compared different machine learning algorithms for classification. In studies with logistic regression, the best results were obtained with accuracy, sensitivity, specificity, and AUC scores of 0.86, 0.81, 0.91 and 0.91, respectively.

In this study, lesion characterization was performed using a single-block convolutional neural network (CNN) from T1-weighted MR abdomen images. ROIs were extracted from the dataset in order to increase the accuracy. First, the effect of the filter size and filter number selected in the CNN's convolution layer on the training performance was examined. Then, the effect of increasing or decreasing the number of convolution blocks used in CNN on training was observed. The contributions of the study are as follows:

• It has been observed that using ROI instead of raw images has a positive effect on study performance.

• It has been observed that continuously increasing the filter size and the number of filters used in the convolution layer is not directly proportional to the training performance, and determination of optimum values directly affects the accuracy.

• It has been determined that using more than one convolution block does not always give better results.

2. Materials and Methods

2.1 Data Set

The data set used in the study was obtained from Selcuk University, Faculty of Medicine, Radiology Department. The device from which the images were taken is SIEMENS AREA 1.5 T, 2013. Images have been converted from DICOM format to JPEG format and each image is 1160x942 pixels. The data set consists of T1-weighted abdominal MR images of 122 patients, 112 of whom are benign and 10 are malignant. In order to increase the study performance, a common frame was determined to include the adrenal lesions in all images and ROIs were extracted from the abdomen images with the help of this frame. T1-weighted abdominal image and examples of the extracted ROI are presented in Figure 1. In the final case, ROI images size is 600x400 pixels.

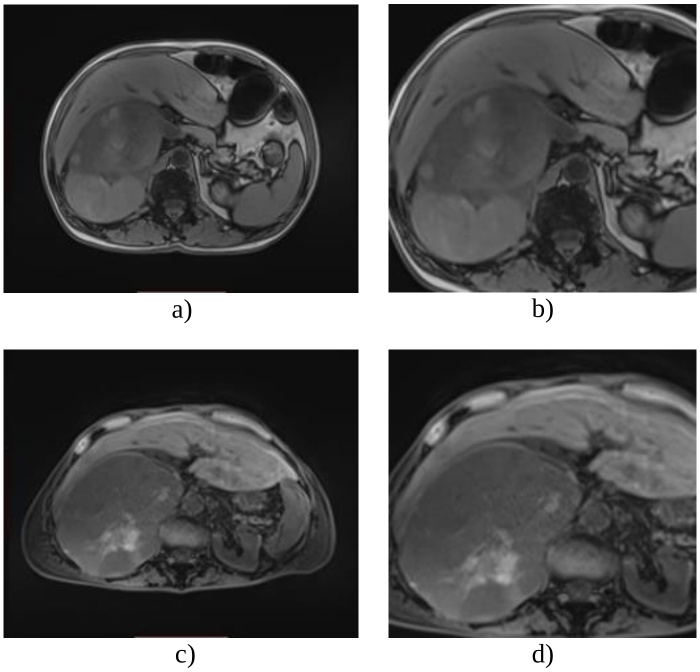

Figure 1. (a), (c) Original T1 Abdomen MR Images**; (b), (d)** Corresponding ROI Images

2.2 Convolutional Neural Network

Convolutional neural networks (CNN) became very popular a decade ago with the performance of AlexNet’s on ImageNet dataset (Krizhevsky, Sutskever, & Hinton, 2012). This has also paved the way for studies in the field of deep learning. The structure of CNN varies according to the study done and the desired properties. It basically consists of a convolution layer and an activation function layer.

The basic CNN block created for this study is given in Figure 2. Here, the structure consists of convolution, batch normalization and activation function layers. For this study, the pooling layer was not preferred because it reduced the data size. Instead, batch normalization is used, which normalizes the extracted feature maps. In Figure 3, structures consisting of one or more basic CNN blocks used in this study are given.

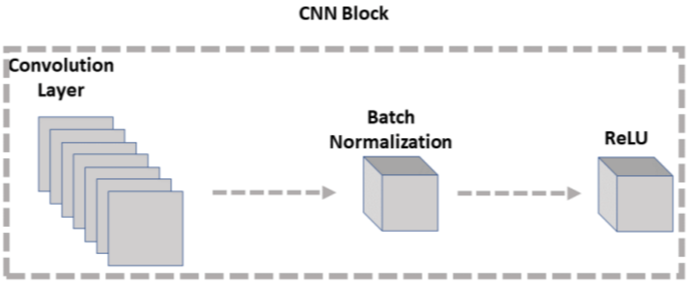

Figure 2. Basic CNN Block

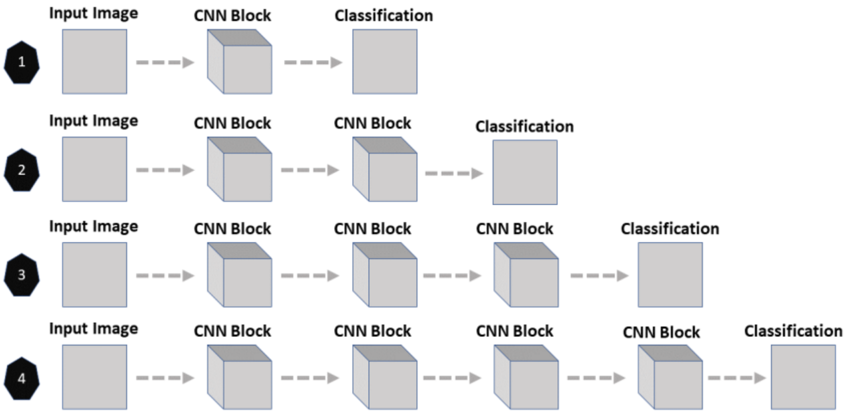

Figure 3. CNN Network Structures Used in This Study

In the convolution layer, the effects of model parameters on the model's classification performance were observed by using different filter sizes and filter numbers. In the batch normalization section, the feature maps extracted in the previous section are normalized according to the mean and standard deviation. In the last part, it is passed through the rectified linear unit (ReLU) activation function and the classification stage is started. In the classification phase, images are separated into related classes with the help of the Softmax activation function.

2.3 Evaluation Metrics

Evaluation metrics are quantitative characteristics used to interpret the performance of the study. The number of metrics used is important both in terms of evaluating the study from different aspects and comparing it with previous studies. In this study, accuracy, confusion matrix, specificity, sensitivity, precision, F1-score and AUC metrics were used, and their formulas are given in Equation 1-6, respectively. In the equations, TP, TN, FP, FN represent True Positive, True Negative, False Positive, False Negative, respectively. In this study, malignant lesions were determined as positive and benign lesions as negative.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{1}$$

$$Confusion\ Matrix = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$F - 1\ Score = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity} \tag{6}$$

3. Experiments

Data augmentation was applied to the ROI images extracted primarily in the study to prevent overfitting. Here, methods such as height shift, width shift, rotation, which are suitable for medical images, are used. In this way, the data set was expanded by obtaining a total of 2000 images, 1000 for each class. At all stages of the study, 75% of this dataset was randomly reserved for training and 25% for validation. In addition, in order to test the performance of the models after the training, tests were performed with 59 benign and 10 malignant images that the network did not see during the training phase. The entire training and testing processes were carried out on MATLAB.

First, a fixed number of filters was determined, and network trainings were carried out for different kernel sizes. The goal here is to find the kernel size that gives the best results. The test results obtained in the test phase performed after the training process are given in Table 1. As can be seen in this table, when kernel sizes 5 and 9 are selected, the test accuracy reaches 0.97, and the most successful results are obtained. When compared with other metrics, they cannot provide a clear advantage over each other. Since training time is an important criterion along with accuracy in classification studies, it would be more accurate to compare these two parameters with the same accuracy values according to the training time. Therefore, the same accuracy was achieved in a shorter time when was selected 5 which had a smaller kernel size compared to the training times. Therefore, in the next part of the study, kernel size 5 will be selected and the process will continue. Accuracy, sensitivity, specificity, precision, F1-score, and AUC score for this kernel size were 0.97, 0.90, 0.98, 0.90, 0.90, and 0.94, respectively.

Table 1. Test Results on Different Kernel Sizes for Single CNN Block

[Kernel Size Filter Numbers]	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC	Training Time
[3 20]	0.91	0.70	0.95	0.70	0.70	0.82	3 min 5 s
[4 20]	0.54	1	0.46	0.24	0.38	0.73	3 min 14 s
[5 20]	**0.97**	**0.90**	**0.98**	**0.90**	**0.90**	**0.94**	**3 min 25 s**
[6 20]	0.85	0.80	0.86	0.50	0.62	0.83	3 min 33 s
[7 20]	0.55	1	0.47	0.24	0.39	0.74	3 min 48 s
[8 20]	0.83	0.90	0.81	0.45	0.60	0.86	4 min 48 s
[9 20]	**0.97**	**0.80**	**1**	**1**	**0.89**	**0.90**	**4 min 43 s**
[10 20]	0.90	0.70	0.93	0.64	0.67	0.82	5 min 7 s
[11 20]	0.83	0.90	0.81	0.45	0.60	0.86	6 min 24 s
[12 20]	0.75	0.80	0.75	0.35	0.45	0.77	6 min 49 s
[13 20]	0.88	0.70	0.92	0.58	0.64	0.81	7 min 16 s

After the kernel size to be used was decided, this time studies were carried out on the number of filters. The results obtained on the test images after the trainings with different filter numbers are shared in Table 2. This table shows that the best results are obtained for both 20 and 25 filter numbers. Since the metric results are the same for both filter number values, the training time will again play a decisive role. It is more appropriate to choose 20 as the number of filters since it achieves similar results in a shorter time.

Table 2. Test Results on Different Filter Numbers for Single CNN Block

[Kernel Size Filter Numbers]	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC	Training Time
[5 5]	0.95	1	0.95	0.77	0.87	0.97	1 min 14 s
[5 10]	0.90	0.90	0.90	0.60	0.72	0.90	1 min 46 s
[5 15]	0.90	0.80	0.91	0.62	0.70	0.86	2 min 27 s
[5 20]	**0.97**	**0.90**	**0.98**	**0.90**	**0.90**	**0.94**	**3 min 25 s**
[5 25]	**0.97**	**0.90**	**0.98**	**0.90**	**0.90**	**0.94**	**4 min 10 s**
[5 30]	0.93	0.60	0.98	0.86	0.71	0.79	4 min 38 s
[5 35]	0.90	0.80	0.92	0.62	0.70	0.86	5 min 25 s
[5 40]	0.88	0.90	0.88	0.56	0.69	0.89	6 min 18 s
[5 45]	0.88	0.70	0.92	0.58	0.64	0.81	8 min 16 s
[5 50]	0.913	0.60	0.97	0.75	0.67	0.78	8 min 43 s

As a result of the studies made with the single-block CNN, it has been seen that the best performance is obtained with the [5 20] filter. In the next section, the effect of using networks consisting of multiple CNN blocks instead of a single CNN block, as shown in Figure 3, on performance is assessed. For this purpose, network structures obtained by cascading two, three and four CNN blocks (Figure 3) were used. The test results obtained at this stage are given in Table 3. The model numbers in Table 3 represent the networks indicated in Figure 3. It is seen that the test accuracy is 0.94 when two CNN blocks are used, 0.91 when three CNN blocks are used, and 0.97 when four CNN blocks are used. When the two and three CNN blocks are used, it is observed that test accuracy decreased as %3 and %6, respectively. On the other hand, when four CNN blocks are used, the same result is obtained as when using a single CNN block. However, when compared in terms of training times, four CNN blocks are deeper and require more computation than a single CNN block, so they reach the same result with a larger training time. There is an approximately seven-fold difference between the two networks in terms of training time. This clearly shows that usage of a single CNN block is more advantageous between two networks with the same test accuracy.

Table 3. Test Results on Different CNN Blocks

Model Number	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC	Training Time
1	**0.97**	**0.90**	**0.98**	**0.90**	**0.90**	**0.94**	**3 min 25 s**
2	0.94	0.60	1	1	0.75	0.80	9 min 27 s
3	0.91	0.80	0.93	0.67	0.72	0.87	15 min 39 s
4	0.97	0.80	1	1	0.89	0.90	22 min 46 s

All in all, it is clear from the tables that the best results among the three different scenarios applied are obtained with [5 20] filter when a single CNN block is used. Accuracy, sensitivity, specificity, precision, F1-score, and AUC scores acquired for these parameters were 0.97, 0.90, 0.98, 0.90, 0.90, and 0.94, respectively. In addition to these, the confusion matrix obtained on the test images is given in Figure 4. As seen in the confusion matrix, only one of 59 benign test images was misclassified, and only 1 of 10 malignant test images was misclassified. This supports the superiority of the model.

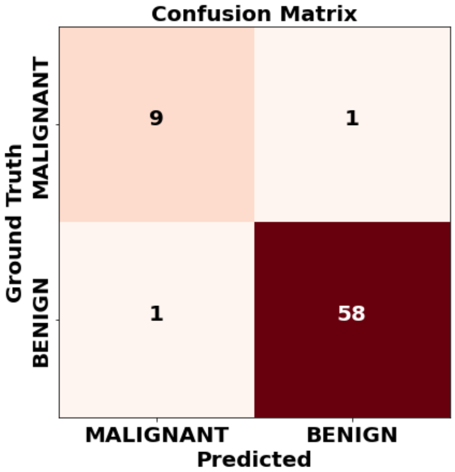

Figure 4. Confusion Matrix for [5 20] Filter

4. Discussion

In parallel with the developments in the field of artificial intelligence, the detection of adrenal masses from radiological sectional images using CAD has also increased. Both machine learning and deep learning methods lead these studies. Table 4 presents the results obtained in this study, as well as the results of previous studies on dual adrenal lesion classification from radiological images. In terms of the accuracy metric, which is considered primarily in the classification studies, the 0.97 accuracy achieved in this study was far superior to the others. When compared with other metrics, it is clearly seen that this study is superior in each metric. The difference of this study from previous studies is that CNN, one of the deep learning structures, was used instead of machine learning algorithms. This revealed the difference between the studies.

Table 4. Comparison With Previous Studies

Study	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC
(X. LI ET AL., 2017)	0.80	---	---	---	---	---
(ROMEO ET AL., 2018)	0.80	0.79	0.80	---	---	0.79
(ELMOHR ET AL., 2019)	0.82	0.81	0.83	---	---	0.89
(KOYUNCU ET AL., 2019)	0.80	0.75	0.82	---	---	0.78
(LIU ET AL., 2022)	0.86	0.81	0.91	---	---	0.91
THIS STUDY	0.97	0.90	0.98	0.90	0.90	0.94

5. Conclusion

In this study, lesion characterization was performed from MR abdominal images using CNN. ROIs were extracted in the direction of the determined frame before the images were given to the network, and the images were augmented with data augmentation. First of all, different kernel sizes were investigated in a fixed number of filters. According to the best kernel size obtained from here, experiments were carried out with different filter numbers this time. Finally, the kernel size-filter numbers, where the best results were obtained, were tested with CNN networks at different depths. As a result of the studies, the best results were obtained when the kernel size was 5 and the number of filters was 20, and in a structure consisting of a single CNN block. The accuracy, sensitivity, specificity, precision, F1-score, and AUC score for these selected values were 0.97, 0.90, 0.98, 0.90, 0.90, and 0.94, respectively. Afterwards, the study was compared with the previous studies, and it was seen that the best results were obtained in all metrics compared to these studies.

The lack of a publicly available data set in the detection of adrenal lesion draws attention as the biggest limitation. In order to compare the studies with previous studies, comparisons could only be made according to the methods used in different data sets and the results obtained. On the other hand, since the image size and network depths were adjusted according to the hardware used in the study, certain limits could not be exceeded. With a stronger hardware infrastructure, it is possible to work on both deeper network structures and higher image sizes.

Future studies will focus on adrenal lesion detection in 3D images as opposed to 2D images and their performance in different network structures.

Acknowledgment

This study was supported by Konya Technical University OYP Project No. 2018-OYP-033.

This study has been evaluated within the scope of the special issue of engineering sciences research and development applications.

References

Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., & Navab, N. (2016). Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging, 35*(5), 1313-1321.

Alex, V., KP, M. S., Chennamsetty, S. S., & Krishnamurthi, G. (2017). *Generative adversarial networks for brain lesion detection.* Paper presented at the Medical Imaging 2017: Image Processing.

Chen, H., Dou, Q., Yu, L., Qin, J., & Heng, P.-A. (2018). VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage, 170,* 446-455.

Dhungel, N., Carneiro, G., & Bradley, A. P. (2015, 2015//). *Deep Learning and Structured Prediction for the Segmentation of Mass in Mammograms.* Paper presented at the Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015, Cham.

Elmohr, M., Fuentes, D., Habra, M., Bhosale, P., Qayyum, A., Gates, E., . . . Elsayes, K. (2019). Machine learning-based texture analysis for differentiation of large adrenal cortical tumours on CT. *Clinical radiology, 74*(10), 818. e811-818. e817.

Fassnacht, M., Arlt, W., Bancos, I., Dralle, H., Newell-Price, J., Sahdev, A., . . . Dekkers, O. M. (2016). Management of adrenal incidentalomas: European society of endocrinology clinical practice guideline in collaboration with the European network for the study of adrenal tumors. *European journal of endocrinology, 175*(2), G1-G34.

Guan, S., & Loew, M. (2017). *Breast Cancer Detection Using Transfer Learning in Convolutional Neural Networks.* Paper presented at the 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR).

Kang, J., & Gwak, J. (2019). Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. *IEEE Access, 7*, 26440-26447.

Koyuncu, H., Ceylan, R., Asoglu, S., Cebeci, H., & Koplay, M. (2019). An extensive study for binary characterisation of adrenal tumours. *Medical & biological engineering & computing, 57*(4), 849-862.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks.* Paper presented at the Advances in neural information processing systems.

Li, Q., Yang, G., Chen, Z., Huang, B., Chen, L., Xu, D., . . . Wang, T. (2017, 14-16 Oct. 2017). *Colorectal polyp segmentation using a fully convolutional neural network.* Paper presented at the 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI).

Li, X., Guindani, M., Ng, C., & Hobbs, B. (2017). *Classification of adrenal lesions through spatial Bayesian modeling of GLCM.* Paper presented at the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017).

Liu, H., Guan, X., Xu, B., Zeng, F., Chen, C., Yin, H. L., . . . Chen, B. T. (2022). Computed Tomography-Based Machine Learning Differentiates Adrenal Pheochromocytoma From Lipid-Poor Adenoma. *Frontiers in endocrinology*, *13*, 833413.

Moeskops, P., Veta, M., Lafarge, M. W., Eppenhof, K. A., & Pluim, J. P. (2017). Adversarial training and dilated convolutions for brain MRI segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 56-64): Springer.

Nguyen, Q., & Lee, S.-W. (2018). *Colorectal segmentation using multiple encoder-decoder network in colonoscopy images*. Paper presented at the 2018 IEEE first international conference on artificial intelligence and knowledge engineering (AIKE).

Romeo, V., Maurea, S., Cuocolo, R., Petretta, M., Mainenti, P. P., Verde, F., . . . Brunetti, A. (2018). Characterization of adrenal lesions on unenhanced MRI using texture analysis: a machine-learning approach. *Journal of Magnetic Resonance Imaging, 48*(1), 198-204.