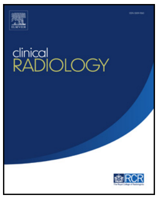
Clinical Radiology 78 (2023) 584–589

Contents lists available at ScienceDirect

Clinical Radiology

journal homepage: www.clinicalradiologyonline.net



Differentiation between normal and abnormal kidneys using ^{99m}Tc-DMSA SPECT with deep learning in paediatric patients

C. Lin[a,b], Y.-C. Chang[a,c], H.-Y. Chiu[a], C.-H. Cheng[d,e], H.-M. Huang[f,*]

[a] *Department of Nuclear Medicine, Chang Gung Memorial Hospital, No. 5, Fuxing Street, Gueishan District, Taoyuan 33305, Taiwan*

[b] *School of Chinese Medicine, Chang Gung University, No. 259, Wenhua 1st Rd, Guishan District, Taoyuan 33302, Taiwan*

[c] *Department of Medical Imaging and Radiological Sciences, College of Medicine, Chang Gung University, No. 259, Wenhua 1st Rd, Guishan District, Taoyuan 33302, Taiwan*

[d] *Department of Pediatrics, Chang Gung University, No. 259, Wenhua 1st Rd, Guishan District, Taoyuan 33302, Taiwan*

[e] *Department of Pediatrics, Chang Gung Memorial Hospital, No. 5, Fuxing Street, Gueishan District, Taoyuan 33305, Taiwan*

[f] *Institute of Medical Device and Imaging, College of Medicine, National Taiwan University, No. 1, Sec. 1, Jen Ai Rd, Zhongzheng District, Taipei City 100, Taiwan*

ARTICLE INFORMATION

Article history:
Received 31 January 2023
Received in revised form
24 April 2023
Accepted 26 April 2023

AIM: To investigate the feasibility of using deep learning (DL) to differentiate normal from abnormal (or scarred) kidneys using technetium-99m dimercaptosuccinic acid (^{99m}Tc-DMSA) single-photon-emission computed tomography (SPECT) in paediatric patients.

MATERIAL AND METHODS: Three hundred and one ^{99m}Tc-DMSA renal SPECT examinations were reviewed retrospectively. The 301 patients were split randomly into 261, 20, and 20 for training, validation, and testing data, respectively. The DL model was trained using three-dimensional (3D) SPECT images, two-dimensional (2D) maximum intensity projections (MIPs), and 2.5-dimensional (2.5D) MIPs (i.e., transverse, sagittal, and coronal views). Each DL model was trained to determine renal SPECT images into either normal or abnormal. Consensus reading results by two nuclear medicine physicians served as the reference standard.

RESULTS: The DL model trained by 2.5D MIPs outperformed that trained by either 3D SPECT images or 2D MIPs. The accuracy, sensitivity, and specificity of the 2.5D model for the differentiation between normal and abnormal kidneys were 92.5%, 90% and 95%, respectively.

CONCLUSION: The experimental results suggest that DL has the potential to differentiate normal from abnormal kidneys in children using ^{99m}Tc-DMSA SPECT imaging.

[*] Guarantor and correspondent: H.-M. Huang, Institute of Medical Device and Imaging, College of Medicine, National Taiwan University, No. 1, Sec. 1, Jen Ai Rd, Zhongzheng District, Taipei City 100, Taiwan.
E-mail address: b9003205@gmail.com (H.-M. Huang).

https://doi.org/10.1016/j.crad.2023.04.015
0009-9260/© 2023 Published by Elsevier Ltd on behalf of The Royal College of Radiologists.

Introduction

Single-photon-emission computed tomography (SPECT) with technetium-99m dimercaptosuccinic acid (^{99m}Tc-DMSA) is a commonly used imaging method for diagnosing kidney diseases such as kidney transplantation, renal scars, renal fibrosis and acute pyelonephritis.[1–5] It is also widely used in paediatric patients for evaluating renal scars, acute pyelonephritis, and renal cortical fibrosis for their small-sized kidneys.[3–5] Additionally, early detection of significant cortical defects might prompt surgical decisions in cases of high-grade vesicoureteral reflux. In general, nuclear medicine physicians performed image interpretation through visual reading on the three-dimensional (3D) SPECT images, displayed as transaxial, coronal, and sagittal views; however, the experience of nuclear medicine physicians may lead to different diagnostic results. One previous study reported that the interobserver agreement was not very high (around 73%) in the interpretation of ^{99m}Tc-DMSA renal SPECT imaging.[6] In addition, Cohen's kappa coefficient showed moderate agreement (around 0.59) between observers.[6] An automated computer-aided diagnostic system that improves interobserver agreement in the differentiation between normal and abnormal kidneys in terms of presence of cortical defects may be of clinical interest.

Recently, deep learning (DL) has been applied in many medical imaging applications.[7,8] For example, DL in combination with SPECT myocardial perfusion imaging was developed to diagnose coronary artery diseases.[9,10] Moreover, DL with SPECT images had shown promising results for the diagnosis of thyroid diseases.[11] Similarly, DL was proposed to detect metastatic and arthritic lesions in SPECT bone scintigraphy.[12] Similarly, DL was applied to detect metastatic lymph node for thyroid cancer on ^{131}I post-ablation whole-body planar imaging.[13] More importantly, one previous study showed that a deep convolutional neural network (CNN) model trained by [^{123}I]*N*-ω-fluoropropyl-2β-carbomethoxy-3β-(4-iodophenyl)nortropane ([^{123}I]FP-CIT) SPECT images could be robust with respect to different camera settings and scan-specific image characteristics.[14] Although further validation is required, DL has the potential to assist nuclear medicine physicians in the diagnosis of various diseases.

To the authors' knowledge, there is only one study that has used an artificial neural network to classify renal ^{99m}Tc-DMSA planar images.[15] No study has evaluated the diagnostic accuracy of depicting paediatric renal cortical defects using ^{99m}Tc-DMSA SPECT and CNN. The aim of the present study was therefore to investigate the feasibility of using DL to differentiate normal and abnormal (or scarred) kidneys from ^{99m}Tc-DMSA SPECT in paediatric patients. Consensus reading results by two nuclear medicine physicians served as the reference standard. In particular, the DL model was trained using different datasets including 3D SPECT images, two-dimensional (2D) maximum intensity projections (MIPs), and 2.5-dimensional (2.5D) MIPs (i.e., transverse, sagittal, and coronal views). Different data augmentation methods were used to reduce overfitting. A dataset consisting of a group of paediatric ^{99m}Tc-DMSA SPECT images was used to assess the performance of the DL-based model.

Materials and methods

DL model for kidney differentiation

In this study, the DL mode used for differentiating normal from abnormal kidneys was based on a CNN. In brief, a CNN typically consists of four different layers: the convolutional layer, the pooling layer, the activation layer, and the fully-connected layer. The convolutional layer is used to extract features from the input images. The number of convolutional filters and kernel size were determined in each convolutional layer. The pooling layer is used to reduce the dimension of the extracted features. For example, the maximum pooling layer reports the maximum value for patches of a feature map and produces a down-sampled feature map. The activation layer is a mathematical function that transfers the values from one layer to another. It adds a non-linear transformation to the CNN network. The fully-connected layer maps the extracted features into the final output by multiplying the input with a weight matrix and adding a bias vector.

During the training process, the final output was fed into a feed-forward neural network, and the weights of the neural network were updated via back propagation. Over a series of epochs, the DL model learns to map the input data into the target. More details about the concept of CNN can be found in Alzubaidi *et al.*[16]

Fig 1 shows the flow chart of the DL training models. The DL model was trained using three different input data types including 3D SPECT images, 2D MIPs, and 2.5D MIPs (i.e. transverse, sagittal, and coronal views). In particular, the 2.5D training method concatenated all features obtained from three independent CNN models. For each input data type, image features extracted from the CNN model were fed into a 512-unit dense fully connected layer followed by a dropout layer with a rate equal to 0.5. The final output layer was a dense layer with one node and a sigmoid activation. The DL model was trained to predict a value ranging from 0 (i.e., normal) to 1 (i.e., abnormal). Note that the input data contained one kidney. In other words, the right and left kidneys were considered as two different datasets.

As shown in Fig 2, the CNN model used in this study consisted of several convolutional layers, batch normalisation layers, rectified linear unit (ReLu) activation layers, and max-pooling layers. Three skip connections (i.e. Add) were also added that were designed to perform an element-wise addition. The final layer of the CNN model was a global average pooling layer. For both the 2D and 2.5D data types, the kernel size of each convolutional (and pooling) layer was 3 × 3. For the 3D data type, the kernel size of each convolutional (and pooling) layer was 3 × 3 × 3. The number of filters was set to 32, 64, 64, 128, 128, 256, and 256 for the first to the seventh convolutional layers, respectively.

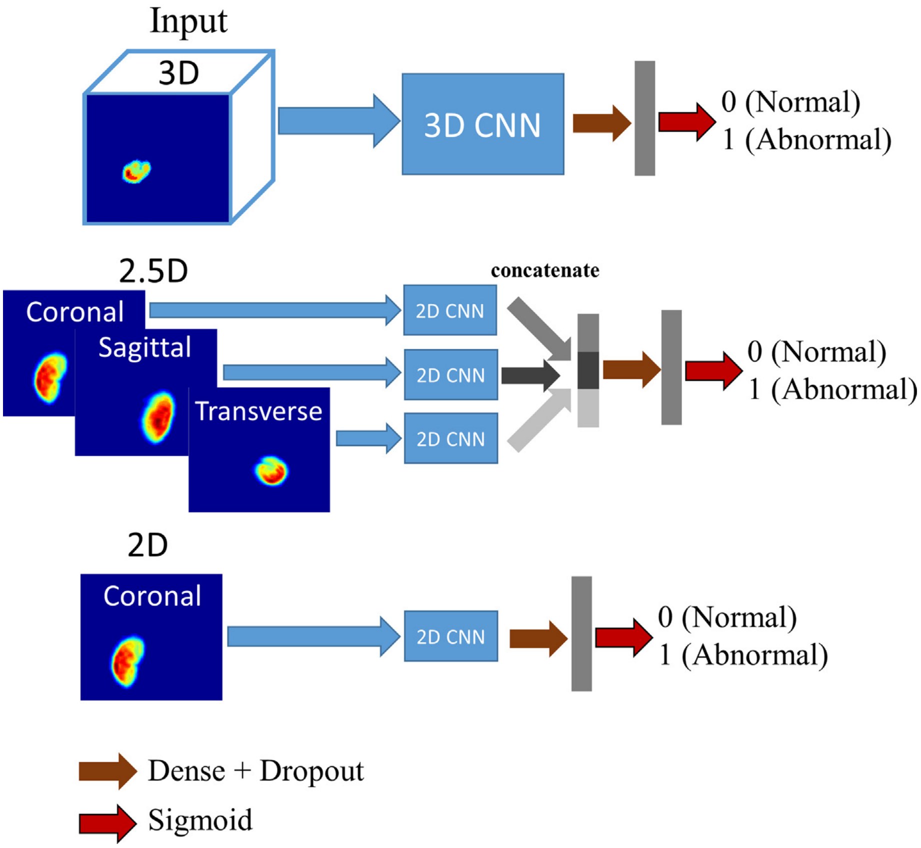


Figure 1 Flow charts of training DL models.

Patient data

The present study retrospectively recruited 301 paediatric patients with at least one episode of prior urinary tract infection who underwent renal ^{99m}Tc-DMSA SPECT between January 2013 and April 2020. The studied population comprised of 150 male patients aged 0.2–17 years (mean age, 3.5 years) and 151 female patients aged 0.3–24 years (mean age, 4.0 years). Only one patient was >18 years but still attended the paediatric department. The institutional review board approved this study and informed consent was waived. All DICOM images were de-identified before using them. For the scan, each patient received an intravenous injection of ^{99m}Tc-DMSA (dose range, 18.5–370 MBq) with 1.85 MBq/kg body weight and minimum dose of 18.5 MBq. Image acquisition was started approximately 3 h after the injection. Each ^{99m}Tc-DMSA SPECT examination was performed with a dual-head gamma camera (ECAM, Siemens, Germany) and a low-energy high-resolution collimator. The data acquisition parameters were as follows: 128 × 128 matrix, zoom factor of two, 1.748-mm pixel, 180° rotation per detector, 60 views per detector (total of 120 views) with a 5-second acquisition per view (total imaging time, 30 minutes = 5 minutes per cycle × 6 cycles), and non-circular orbit in continuous acquisition. All SPECT images were reconstructed using the 2D ordered-subsets expectation-maximisation algorithm with eight subsets and six iterations. All reconstructed SPECT images were then smoothed using a Gaussian filter of 6-mm full width at half maximum. Finally, the 3D renal SPECT images of each patient were examined visually by two experienced nuclear medicine physicians in consensus, which served as the ground truth, and each kidney was classified as either

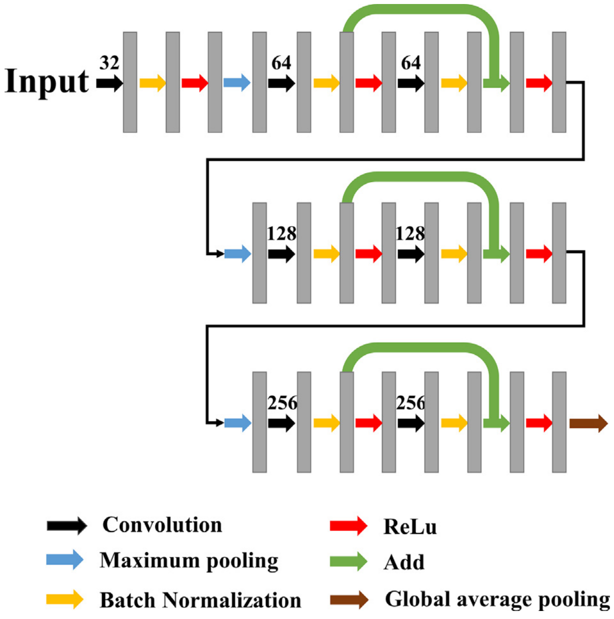


Figure 2 The architecture of the CNN model. The number on the top of each convolutional layer is the number of filters.

normal or abnormal. A kidney was designated as abnormal if presence of any cortical defect.

Data pre-processing and model training

A total of 301 patients were split randomly into 261 for training (i.e., 520 kidneys, two kidneys were excluded), 20 for validation (i.e., 40 kidneys), and 20 for testing (i.e., 40 kidneys). Because the number of sections varied between patients, each 3D SPECT dataset was zero padded to 128 sections. Then, each 3D SPECT dataset was normalised by its maximum value. Finally, the 3D SPECT images containing the kidney were split manually into the left and right side, and the individual left and right kidneys were used as input data.

In the present study, each DL model was trained using a binary cross-entropy loss function. The Adam optimisation algorithm was used to minimise the loss function. The learning rate was 10^{-4}. The hyper-parameters β 1 and $\beta 2$ were set to 0.9 and 0.999, respectively. The number of epochs was set to 300, and the batch size was 4. During the training phase, the model with the highest accuracy over all validation datasets was selected from the epoch. Each DL model was implemented with TensorFlow 2.8.0 and ran on a NVIDIA GeForce RTX 3090 GPU.

To reduce the overfitting problem, data augmentation was performed that synthesised new data from the existing training data. Specifically, the training data were augmented through random flip (both horizontal and vertical), translation (± 25 pixels), rotation ($\pm 90^{o}$), and zoom (20% zoom-in and -out). Note that the data augmentation technique was implemented only in the training process. For the training data (i.e., 520 kidneys), the number of normal kidneys was 368, and the number of abnormal kidneys was 152. To deal with the class imbalance problem, class weights (i.e., the total number of cases divided by the total number of cases in that class) were added to the loss function. Note that both validation and testing datasets had 20 patients. Among these patients, five patients had no defect in both kidneys, five patients had defects in both kidneys, five patients had no defect in the right kidney (but defects in the left kidney), and five patients had no defect in the left kidney (but defects in the right kidney). In other words, 20 kidneys were labelled as normal, and 20 kidneys were labelled as abnormal.

Evaluation metrics

To evaluate model performance on the testing dataset, a receiver operating characteristic (ROC) curve was plotted that was used to choose the optimal threshold of each trained DL model. Then, the accuracy, sensitivity, specificity, precision, and F1 score at the optimal threshold were calculated.

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

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Specifity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{F1 score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{5}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively. Cohen's kappa coefficient was also calculated, which was a measure of how closely the SPECT images classified by the DL model matched the results labelled by the two nuclear medicine physicians.

Results

Fig 3 shows ROC curves of the DL models trained by 3D, 2.5D, 2D (coronal), 2D (sagittal), and 2D (transverse) images. The area under curve was 0.91, 0.94, 0.92, 0.87, and 0.92 for 3D, 2.5D, 2D (coronal), 2D (sagittal), and 2D (transverse) models, respectively. The results indicate that the 2.5D model exhibited better differentiation performance than the other models. Table 1 summarises the diagnostic accuracy, sensitivity, specificity, precision, and F1-score of five DL models for differentiating between normal and abnormal kidneys. The DL model trained by 2.5D MIPs outperformed that trained by 3D SPECT images or 2D MIPs. Fig 4 shows the confusion matrix of differentiation using the 2.5D model. Only three kidneys (i.e., two FN and one FP cases) were categorised incorrectly by the 2.5D model. Fig 5 shows the coronal MIPs of these three

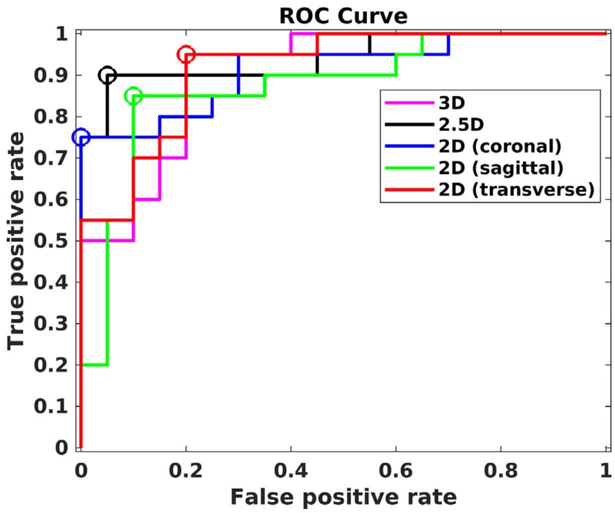


Figure 3 ROC curves of the DL models trained by 3D images, 2.5D MIPs, 2D (coronal) MIPs, 2D (sagittal) MIPs, and 2D (transverse) MIPs. The circle marker denotes the optimal point.

Table 1
Diagnostic accuracy, sensitivity, specificity, precision, and F1-score of 5 deep-learning (DL) models for differentiating normal and abnormal kidneys.

Input data types	Accuracy	Sensitivity	Specificity	Precision	F1-score
3D images	0.850	0.900	0.800	0.818	0.857
2.5D MIPs	0.925	0.900	0.950	0.947	0.923
2D MIP (coronal)	0.850	0.750	0.950	0.938	0.833
2D MIP (sagittal)	0.850	0.800	0.900	0.889	0.842
2D MIP (transaxial)	0.850	0.900	0.800	0.818	0.857

The 5 DL models were trained by five different input data types.
MIP, maximum intensity projection.

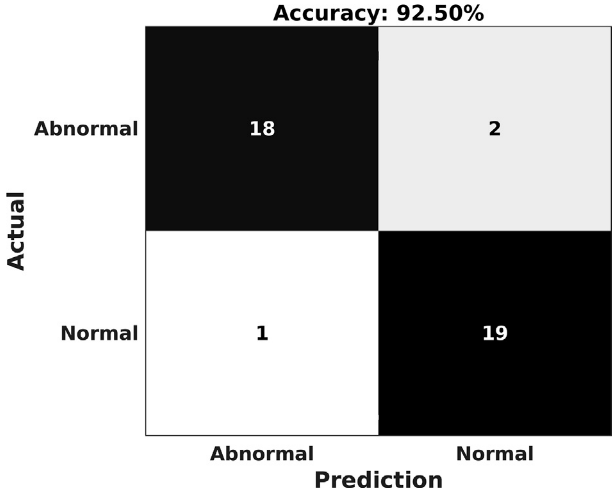


Figure 4 Confusion matrix of differentiation using the 2.5D DL model.

false-predicted kidneys. Cohen's kappa coefficient was 0.70, 0.85, 0.70, 0.70, and 0.70 for 3D, 2.5D, 2D (coronal), 2D (sagittal), and 2D (transverse) models, respectively.

Discussion

The present study applied DL to the automatic differentiation of paediatric normal from abnormal kidneys from ^{99m}Tc-DMSA SPECT images. The 2.5D model outperformed both 2D and 3D models in terms of accuracy, precision, and F1-score. DL in combination with 2.5D MIPs achieved 92.5% accuracy, 90% sensitivity, and 95% specificity in the differentiation between normal and abnormal kidneys. Moreover, a kappa coefficient of 0.85 obtained from the 2.5D model indicated that DL could provide an almost perfect agreement with the visual reading results of the two experienced nuclear medicine physicians. These preliminary results suggest that DL may be used as a computer-aided diagnosis system that assists less-experienced nuclear medicine physicians in the diagnosis of renal cortical defects.

To the authors' knowledge, this is the first study using DL with ^{99m}Tc-DMSA SPECT images to evaluate differentiating normal from abnormal kidneys; however, a previous study used an artificial neural network to create an automatic classifier system for renal ^{99m}Tc-DMSA planar images.[15] Based on the pre-defined heuristic features extracted from posterior images, the artificial neural network could achieve a differentiation accuracy of 95.9%.[15] In contrast, the proposed DL method could achieve similar accuracy (92.5%) without manual feature selection. In addition, the proposed DL method was designed to process one single kidney. This is different from the previous study that required heuristic features (e.g., perimeter difference, counts difference, and area difference) obtained from both right and left kidneys.

One interesting finding of the present study was that the 3D model did not show better differentiation performance than the other 2D models. One possible reason is that only some SPECT sections reflect renal cortical function impairment. In fact, each 3D SPECT dataset was zero-padded along the transverse axis to the size of 128 sections. SPECT sections that do not show impaired renal function may affect the differentiation performance. Moreover, the 3D CNN model used in this study may be too simple. Further improvement may be obtained by using important SPECT image sections or more complex 3D CNN models.

Despite promising results, this study has several limitations. First, the number of patients used in this study was smaller than that used in previous studies.[9–14] Further validation using large datasets is required. Moreover, only paediatric patients were recruited; however, the proposed method should be applicable to adults including renal transplant donor candidates.

Note that the DL model trained by small datasets may be overfitted. The overfitting problem can be reduced by using the data augmentation technique described in Section 2.2.

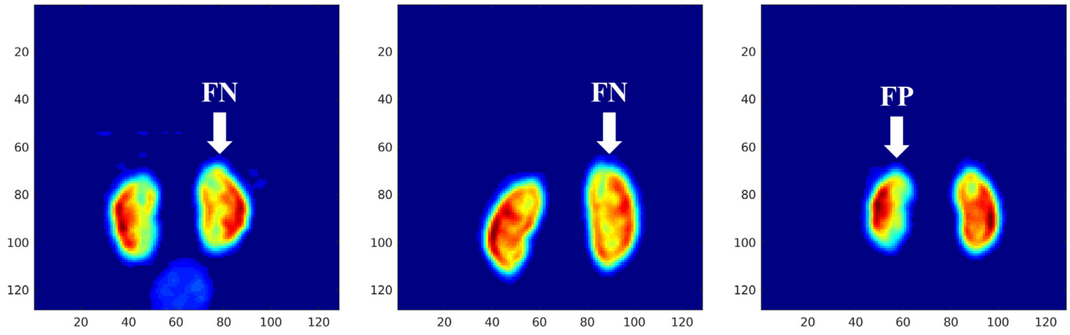


Figure 5 Coronal MIPs of the three false-predicted cases predicted by the 2.5D model. FN: DL model failed to recognise abnormal kidney considered by expert consensus reading and FP: DL model erroneously considered the presence of abnormal kidney function.

Second, the ground truth used for training the DL model was based on the visual reading results of the two experienced nuclear medicine physicians with consensus. There might be a chance that the two nuclear medicine physicians incorrectly categorised normal and abnormal kidneys. This represents the clinical scenario and it would be clinically infeasible to obtain pathological proof for each cortical defect. As a result, it is difficult to explain the clinical relevance of false findings shown in Fig 5. The present study can only demonstrate the discrepancy between DL model and expert consensus reading. Any potential clinical impact may require further testing in a larger scaled and prospective study. Third, the CNN model used in the present study may not be optimal. Well-known classification models such as DenseNet[17] and InceptionV3[18] were not used in the present study. The reason is that these well-known models are designed to classify natural colour images. Compared to natural images, SPECT images are grey-scale and less complex. It may be worthwhile testing advanced DL techniques such as label smoothing,[19] generative adversarial network based data augmentation,[20] and attention mechanism.[21] Moreover, the pre-defined heuristic features[15] and pertinent clinical information[10] we could be integrated into the CNN model.

In conclusion, present study investigated the feasibility of using DL to differentiate between normal and abnormal (or scarred) kidneys from ^{99m}Tc-DMSA SPECT in paediatric patients. The DL model was trained using different data types including 3D SPECT images, 2D MIPs, and 2.5D MIPs. Each DL model was trained to determine renal SPECT images into normal or abnormal. Compared to 3D SPECT images and 2D MIPs, 2.5D MIPs had the best performance in the differentiation between normal and abnormal kidneys. The diagnostic accuracy obtained from the 2.5D model was 92.5%. These preliminary results suggest that DL has the potential to differentiate normal from abnormal kidneys in paediatric patients using ^{99m}Tc-DMSA SPECT imaging.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This work was supported by MOST 107-2221-E-002-081, 109-2314-B-182A-071, from Ministry of Science Technology, Taiwan and CMRPG5J0132 from the Research Fund of Chang Gung Memorial Hospital, Taiwan.

References

1. Groshar D, Embon OM, Frenkel A. Front D. Renal function and technetium-99m-dimercaptosuccinic acid uptake in single kidneys: the value of in vivo SPECT quantitation. *J Nucl Med* 1991;**32**(5):766–8.
2. Cairns HS, Spencer S, Hilson AJW, *et al*. ^{99m}Tc-DMSA imaging with tomography in renal transplant recipients with abnormal lower urinary tracts. *Nephrol Dial Transplant* 1994;**9**(8):1157–61.
3. Yen TC, Chen WP, Chang SL, *et al*. A comparative study of evaluating renal scars by ^{99m}Tc-DMSA planar and SPECT renal scans, intravenous urography, and ultrasonography. *Ann Nucl Med* 1994;**8**(2):147–52.
4. Yen TC, Chen WP, Chang SL, *et al*. Technetium-99m-DMSA renal SPECT in diagnosing and monitoring paediatric acute pyelonephritis. *J Nucl Med* 1996;**37**(8):1349–53.
5. Chiou YY, Wang ST, Tang MJ, *et al*. Renal fibrosis: prediction from acute pyelonephritis focus volume measured at ^{99m}Tc dimercaptosuccinic acid SPECT. *Radiology* 2001;**221**(2):366–70.
6. Beslic N, Milardovic R, Sadija A, *et al*. Interobserver variability in interpretation of planar and SPECT Tc-99m-DMSA renal scintigraphy in children. *Acta Inform Med* 2017;**25**(1):28–33.
7. Anaya-Isaza A, Mera-Jiménez L, Zequera-Diaz M. An overview of deep learning in medical imaging. *Inform Med Unlocked* 2021;**26**:100723.
8. Cai L, Gao J, Zhao D. A review of the application of deep learning in medical image classification and segmentation. *Ann Transl Med* 2020;**8**(11):713.
9. Betancur J, Commandeur F, Motlagh M, *et al*. Deep learning for prediction of obstructive disease from fast myocardial perfusion SPECT: a multicenter study. *JACC Cardiovasc Imaging* 2018;**11**(11):1654–63.
10. Apostolopoulos ID, Apostolopoulos DI, Spyridonidis TI, *et al*. Multi-input deep learning approach for cardiovascular disease diagnosis using myocardial perfusion imaging and clinical data. *Phys Med* 2021;**84**:168–77.
11. Ma L, Ma C, Liu Y, *et al*. Thyroid diagnosis from SPECT images using convolutional neural network with optimization. *Comput Intell Neurosci* 2019:6212759.
12. Lin Q, Cao C, Li T, *et al*. dSPIC: a deep SPECT image classification network for automated multi-disease, multi-lesion diagnosis. *BMC Med Imaging* 2021;**21**(1):1–16.
13. Kavitha MS, Lee CH, Shibudas KS, *et al*. Deep learning enables automated localization of the metastatic lymph node for thyroid cancer on ^{131}I post-ablation whole-body planar scans. *Sci Rep* 2020;**10**(1):1–12.
14. Wenzel M, Milletari F, Krüger J, *et al*. Automatic classification of dopamine transporter SPECT: deep convolutional neural networks can be trained to be robust with respect to variable image characteristics. *Eur J Nucl Med Mol Imaging* 2019;**46**(13):2800–11.
15. Wright JW, Duguid R, McKiddie F, *et al*. Automatic classification of DMSA scans using an artificial neural network. *Phys Med Biol* 2014;**59**(7):1789–800.
16. Alzubaidi L, Zhang J, Humaidi AJ, *et al*. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 2021;**8**:53.
17. Huang G, Liu Z, Van Der Maaten L, *et al*. Densely connected convolutional networks. *ArXiv* 2016. 1608.06993.
18. Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision. *ArXiv* 2015. 1512.00567.
19. Müller R, Kornblith S, Hinton G. When does label smoothing help? *ArXiv* 2019. 1906.02629.
20. Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial networks. *ArXiv* 2015. 1406.2661.
21. Wang F, Jiang M, Qian C, *et al*. Residual attention network for image classification. *ArXiv* 2017. 1704.0690.