research article

Detection and localization of hyperfunctioning parathyroid glands on [^{18}F]fluorocholine PET/CT using deep learning – model performance and comparison to human experts

Leon Jarabek[1], Jan Jamsek[2], Anka Cuderman[2], Sebastijan Rep[2,3], Marko Hocevar[4,5], Tomaz Kocjan[5,6], Mojca Jensterle[5,6], Ziga Spiclin[7], Ziga Macek Lezaic[8], Filip Cvetko[5], Luka Lezaic[2,5]

[1] Department of Radiology, General Hospital Novo Mesto, Slovenia
[2] Department for Nuclear Medicine, University Medical Centre Ljubljana, Slovenia
[3] Faculty of Health Sciences, University of Ljubljana, Ljubljana, Slovenia
[4] Department of Surgical Oncology, Institute of Oncology, Ljubljana
[5] Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
[6] Department for Endocrinology, Diabetes and Metabolic Diseases, University Medical Centre Ljubljana, Slovenia
[7] Faculty of Electrical Engineering, University of Ljubljana, Slovenia
[8] Rožna dolina, c. VI/8, Ljubljana, Slovenia

Radiol Oncol 2022; 56(4): 440-452.

Received 21 April 2022
Accepted 22 August 2022

Correspondence to: Assist. Prof. Luka Ležaić, M.D., Ph.D., Department for Nuclear Medicine, University Medical Centre Ljubljana, Slovenia.
E-mail: luka.lezaic@kclj.si

Disclosure: No potential conflicts of interest were disclosed.

Background. In the setting of primary hyperparathyroidism (PHPT), [^{18}F]fluorocholine PET/CT (FCH-PET) has excellent diagnostic performance, with experienced practitioners achieving 97.7% accuracy in localising hyperfunctioning parathyroid tissue (HPTT). Due to the relative triviality of the task for human readers, we explored the performance of deep learning (DL) methods for HPTT detection and localisation on FCH-PET images in the setting of PHPT.

Patients and methods. We used a dataset of 93 subjects with PHPT imaged using FCH-PET, of which 74 subjects had visible HPTT while 19 controls had no visible HPTT on FCH-PET. A conventional Resnet10 as well as a novel mPETResnet10 DL model were trained and tested to detect (present, not present) and localise (upper left, lower left, upper right or lower right) HPTT. Our mPETResnet10 architecture also contained a region-of-interest masking algorithm that we evaluated qualitatively in order to try to explain the model's decision process.

Results. The models detected the presence of HPTT with an accuracy of 83% and determined the quadrant of HPTT with an accuracy of 74%. The DL methods performed statistically worse ($p < 0.001$) in both tasks compared to human readers, who localise HPTT with the accuracy of 97.7%. The produced region-of-interest mask, while not showing a consistent added value in the qualitative evaluation of model's decision process, had correctly identified the foreground PET signal.

Conclusions. Our experiment is the first reported use of DL analysis of FCH-PET in PHPT. We have shown that it is possible to utilize DL methods with FCH-PET to detect and localize HPTT. Given our small dataset of 93 subjects, results are nevertheless promising for further research.

Key words: primary hyperparathyroidism, deep learning, nuclear medicine, fluorocholine, PET/CT

Introduction

Primary hyperparathyroidism (PHPT) is the third most common endocrine disorder with a reported prevalence ranging from 1 to 21 per 1,000 among the general population.[1] PHPT is the result of hyperfunctioning parathyroid tissue (HPTT), which becomes insensitive to the inhibitory effect of hy-

percalcemia. Histologically HPTT can be either an adenoma (in approximately 80% of cases), multiple adenomas, hyperplasia or rarely a carcinoma (in approximately 1% of cases).[2] The treatment of PHPT typically requires surgical removal of HPTT. Modern, minimally invasive surgical techniques require precise preoperative localization of HPTT. For this task, [18F]fluorocholine PET/CT (FCH-PET) is one of the most promising imaging modalities, with reported sensitivities of 94–100% and specificities of 88–100%.[3-13] Performance of FCH-PET was repeatedly shown to be superior to other HPTT localization methods, while at the same time having lower radiation exposure compared to other nuclear medicine modalities.[14]

Deep learning (DL) techniques with convolutional neural networks (CNN) have proven to be useful in various computer vision tasks, such as super-resolution, image synthesis, denoising, classification, segmentation and object detection.[15-22] In medical imaging, CNNs have shown promising performance, even exceeding experts in some specific cases, such as grading diabetic retinopathy from fundus images, detecting skin cancer from photographs and detecting abnormalities on chest X-ray images.[23-25] Research of CNNs in nuclear medicine showed its potential in reducing the PET radiation dose, improving image quality, lesion detection and segmentation as well as prediction of prognosis.[21-36]

Given the excellent human performance of analysing FCH-PET for the presence and localisation of HPTT, an interesting opportunity to challenge DL techniques is presented. An automated analysis pipeline of FCH-PET that would classify HPTT presence and location would allow for efficient surgical planning and could serve to double check the experts' reports. Such analysis would also allow for more accurate and objective comparison of potential follow-up studies; these are not often required, but unavoidable in cases of persistent or recurrent hyperparathyroidism. Furthermore, if the model could visualise the pathological uptake in the study, it would provide more visual feedback to the surgeon in axial images to allow for better visualisation of HPTT and would allow faster interpretation of interplay of surrounding anatomical structures. Our aim was to explore the performance of DL analysis of FCH-PET in the setting of PHPT, since the use of DL for FCH-PET analysis in PHPT has not yet been thoroughly investigated.

To this end, we developed a classification model which classifies whether HPTT is present in the study and its location. We also attempt to model in a novel unsupervised manner the regions-of-interest fed to the model. Furthermore, we aimed to provide a preliminary comparison of the diagnostic accuracy of the DL models to human experts to determine clinical applicability, as the model should be as accurate as an expert in evaluating FCH-PET studies to be clinically applicable.

Patients and methods

This was a retrospective analysis of prospective clinical trial data (NCT03203668) performed at the University Medical Centre Ljubljana and Institute of Oncology Ljubljana. The clinical trial was approved by the Medical Ethics Committee of the Republic of Slovenia (approval number 77/11/12). The trial only included patients with biochemically confirmed primary hyperparathyroidism; hypercalcemic patients had elevated or inappropriately normal parathormone (PTH) levels, whereas normocalcemic patients had inappropriately elevated PTH levels. All included patients were older than 18 years and had no clinical history of oncological, inflammatory, or infectious disease of the head and neck. No pregnant women were included in the trial. The retrospective use of the data was approved by the Medical Ethics Committee of the Republic od Slovenia (approval number 0120-582/2021/4) and the patient consent was waived due to the retrospective nature of the analysis.

The study only included images of patients with biochemically confirmed PHPT at time of FCH-PET imaging. Since the trial did not include healthy controls, data of patients with the following criteria were chosen as "controls": no visible HPTT in FCH-PET at time of imaging; have not undergone surgery in thyroid region; were biochemically normocalcemic at 6 months' follow-up.

Dataset description and PET-CT image acquisition

We used the data of 79 participants (22 male, 57 female) with visible HPTT lesions on FCH-PET (referred below as *patients*) and 19 participants (7 male, 12 female) without visible HPTT lesions on FCH-PET (referred below as *controls*). Average age (± SD) of *patients* was 58.7 ± 12.7 years and average age of *controls* was 60.1 ± 11.8 years. Both *patients* and *control* groups were comparable in terms of age ($p = 0.659$) as well as male to female ratio ($p = 0.852$), as determined by Student *t*-test and normalised Chi-square test, respectively.[37,38]

FCH-PET imaging was performed at the Department for Nuclear Medicine of the University Medical Centre Ljubljana. The acquisition details were the same as in Cuderman *et al*.[3] The patients fasted 6 hours prior to the examination, were well hydrated and injected with 100 MBq of [18F] Fluorocholine (FCH). Acquisition was performed on a Siemens Biograph mCT® PET/CT (Siemens Healthineers AG, München, DE) 5 minutes and 60 minutes after the FCH application. The imaging region extended from the angle of mandible to the aortic arch. The imaging consisted of a low-dose CT (120 kVp, 25 mAs, CARE Dose 4D, FBP reconstruction), followed by PET imaging (one bed position of 4 minutes). PET images were reconstructed using Siemens HD PET software with iterative TrueX + TOF OSEM method (2 iterations, 21 subsets) with 400 × 400 matrix, zoom 1 and Gaussian filter with FWHM of 4 mm. To train and evaluate DL models, we used only images acquired 60 minutes after FCH application, where the balance of image quality and target-to-backround ratio is typically highest.

All *patients* with HPTT present on FCH-PET were surgically treated at Institute of Oncology Ljubljana. Ground truth HPTT presence and location for training the CNNs was based on the post-surgical histopathological results. Furthermore, our dataset included formatted information from FCH-PET reports as used by Cuderman *et al*. that we used to compare the performance of DL models with human experts.[3] These reports were used to guide the subsequent surgical removal of the HPTT.

For simplicity, we only used *patients* who had single gland disease and had HPTT in the typical anatomic location of parathyroid glands. HPTT was thus in one of 4 possible locations: upper left (UL, 21 *patients*), lower left (LL, 27 *patients*), upper right (UR, 5 *patients*) and lower right (LR, 26 *patients*). Since the UR location in our dataset contained only 5 patients, it was removed from the final analysis due to under-representation. For the final model development and evaluation, we used 19 *controls* and 74 *patients*, among them 21 with UL HPTT, 27 with LL HPTT and 26 with LR HPTT.

Image pre-processing

We used the same pre-processing pipeline for all analyzed images. First, we resampled the CT image using bivariate spline interpolation from *scipy* library to match the PET image matrix of 200 × 200 × 56.[39] 3D interpolation was not needed as CT was reconstructed at same slices as PET. Both images were concatenated to produce a 200 × 200 × 56 × 2 matrix representing the PET/CT. Next, we cropped the desired region of interest containing the hyperfunctioning parathyroid tissue to the matrix of size 64 × 64 × 32. For all patients, the region was cropped at same PET/CT coordinates, which were chosen empirically, such that it contained HPTT in all studies. In this way, there are lower memory requirements to run deep learning models.

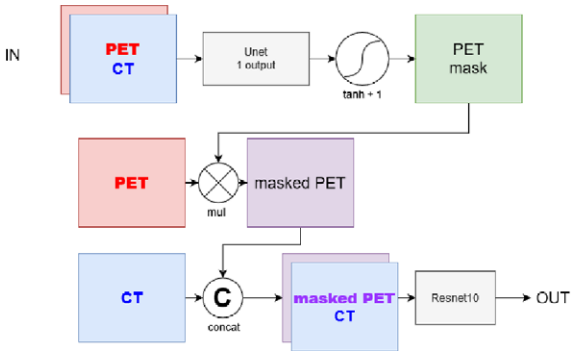

FIGURE 1. mPETResnet10 architecture. First, PET-CT images are fed into UNet with a single channel output and tanh+1 activation function. This output is the PET mask. This mask is elementwise multiplied with PET image to produce a masked PET image. Masked PET is concatenated with the original CT and the masked PET-CT is fed into the ResNet10 classifier. Gray boxes represent deep-learning models, coloured boxes represent data, and circles represent operations of *tanh+1*, *multiplication* (mul) by element and *concatenation* (concat).

The labels for an image were represented by a one-hot encoded vector of length 4, representing locations UL, LL, LR and a dummy variable representing "healthy" *controls*.

Modelling

For modelling, we defined 2 tasks: (i) a task of classifying whether the HPTT is present in the image or not (CPr, classification of presence) and (ii) a task of classifying in which quadrant the HPTT was present in the image (CLoc, classification of location). CPr is a simple binary classification task where *p(HPTT) = 1 – p(healthy)*. CLoc is a multi-class classification task where each output of the model is analogous to the probability of HPTT being present at one of three considered locations UL, LL and LR.

With normalized PET-CT images represented by a matrix of shape 200 × 200 × 56 × 2 as input, the output of the model was a vector of length 4, activated by *SoftMax* activation function, corresponding to *p(UL)*, *p(LL)*, *p(LR)* and *p(healthy)* (Figure 1).

The model was therefore trained for both CPr and CLoc simultaneously. Furthermore, the dataset was well balanced, containing a similar number of cases for each of the 4 classes, and thus ensured stable training using cross entropy as a loss function.[40] For training, batch size of 5 was used with stochastic gradient descent optimizer with momentum of 0.9 and weight decay of 0.005. The initial learning rate was determined by a grid search in log space and learning rate decay on plateau scheduling was used. Identical procedure was used for all models. All models were trained from scratch.

For both CPr and CLoc classification tasks, we performed baseline experiments using the 3D version of *Resnet10* (RN10) architecture and using our novel architecture as described below.[41,42] Our choice of architecture of Resnet10 was based on extensive experiments which included other state-of-the-art, and larger architectures, namely using 3D versions of *Densenet121*[43], *wideResNet101*[44], *PreActResnet101*[45], *Resnet101*[41] and *Resnet50.* For all architectures except our novel architecture, implementations from Kensho *et al*. were used.[42]

We provide comprehensive comparison between the performance of RN10 and proposed architecture "*masked-PET Resnet10*" (mRN10), as well as the comparison of mRN10 to experts' performance.

Masked-PET Resnet10

We developed a novel architecture designed to mask PET signals from unimportant (i.e., physiological uptake) regions with high signal (eg. muscle tissue, salivary glands) before entering the RN10 classifier. This is important as the FCH-PET images are heteroscedastic, with some regions - like muscle - having high variance between subjects and other regions - like air - having low variance. To mitigate this, and to improve conditioning of the data and therefore the stability of the classifier,[46] we decided to allow the model itself to optimize for differentiable masking of these potentially problematic regions. We named the proposed architecture "*masked-PET Resnet10*" (mRN10).

The mRN10 consisted of 2 parts. First, a *Unet* architecture was used to mask the PET-CT.[47] Next, *Resnet10* was used to classify the masked PET-CT. We decided on *Unet* architecture since it is commonly used in segmentation tasks[21] and we deemed the task of masking to be similar to segmentation of the region-of-interest. Masking was achieved by first activating per-voxel output of *Unet* with activation function $f(x) = tanh(x)+1$. These output values were in interval (0,2), such that regions where *Unet* output was negative were closer to 0, while regions where *Unet* output was positive were closer to 2. This matrix, representing the mask, was then multiplied elementwise by the PET matrix, to produce a masked PET image.

The architecture of mRN10 is depicted on Figure 1. Regions in PET image where *Unet* output was negative were multiplied by values close to 0 and were therefore effectively "masked" from the PET image. This masked PET was then concatenated with CT and the masked PET-CT was used as input for the *Resnet10* classifier. The entire mRN10 was trained end-to-end, therefore the masking was optimized for the lowest loss in the classification task of the downstream *Resnet10* classifier.

The models were written in *python 3.8.0* using *Pytorch 1.10* framework and trained on a single *GTX 1080Ti* graphics card (*Nvidia Corporation, Santa Clara, US*).[48,49] The code is freely available online at: https://github.com/ljarabek/AI_FCH

Training and evaluation

For training, we used 12-fold cross-validation with data split into a test set of 10 random subjects, with the remaining subjects being randomly split into a training set (90% of the remaining subjects) and validation set (10% of the remaining subjects). Data was normalised using z-score normalization upon splitting accordingly, such that the mean and standard deviation were computed only using the training set. Sets were sampled such that each set contained at least 1 subject from each class (UL, LL, LR and *control*). For testing, the model with the lowest validation loss was used. The confusion matrix for CPr evaluation was computed by summing the confusion matrices for the test set across the 12 data splits, providing 120 total samples. The confusion matrix for CLoc was obtained by summing the 3 confusion matrices for evaluated locations UL, LL, LR across the best performing 12 data splits, providing 360 "samples". Similarly, the area under the receiver operating characteristic curve (AUCROC) was computed.

We used *epiR* package for *R* to determine the diagnostic performance metrics and *McNemar* test from *DTComPair* package for determining statistically significant ($p < 0.05$) differences.[50-53] Only binary diagnostic performance metrics were used for evaluation, even though CLoc is theoretically a multi-class classification task. In this way, the results comparable to studies evaluating the performance of FCH-PET, since they also mostly used binary classification metrics.[3-13]

TABLE 1. Confusion matrices for CPr **(A)** and CLoc **(B)** for both RN10 and mRN10 models. Note that the confusion matrices for CLoc have more samples (360 in total), as they were computed by summing the confusion matrices for each of the three included locations (UL, LL, LR)

Ⓐ

	CPr task with RN10				CPr task with mRN10		
	HPTT present	HPTT not present	**sum**		HPTT present	HPTT not present	**sum**
Model output HPTT present	79	8	**87**	Model output HPTT present	90	11	**101**
Model output HPTT not present	20	13	**33**	Model output HPTT not present	9	10	**19**
sum	**99**	**21**	**120**	**sum**	**99**	**21**	**120**

Ⓑ

	CLoc task with RN10				CLoc task with mRN10		
	HPTT at GTLoc	HPTT not at GTLoc	**sum**		HPTT at GTLoc	HPTT not at GTLoc	**sum**
Predicted GTLoc	35	51	**86**	Predicted GTLoc	53	50	**103**
Not predicted GTLoc	61	213	**274**	Not predicted GTLoc	43	214	**257**
sum	**96**	**264**	**360**	**sum**	**96**	**264**	**360**

CPr = classification of presence; CLoc = classification of location; GTLoc = ground truth location based on postsurgical histopathological reports; HPTT = hyperactive parathyroid tissue; mRN10 = novel *masked-PET Resnet10* model; RN10 = baseline *Resnet10* model

Results

We determined the best performing models for both RN10 and mRN10 were trained using the initial learning rate of 0.013. The confusion matrices for RN10 and mRN10 are presented in Tables 1A and 1B, while the diagnostic performances for both tasks using the RN10 and mRN10 models are presented in Table 2. Both models had comparable performance in the CPr task. The mRN10 had a significantly higher accuracy for the CLoc task than the RN10 and was therefore used for comparison with human performance.

We performed a comprehensive comparison with human expert evaluation only for the CLoc task. Healthy controls had, by definition, no HPTT visible on FCH-PET (as reported by human experts), so the comparison could not be made for the CPr task, as human performance for CPr was 100%. Comparison of performance metrics for the CLoc task between the mRN10 model and human performance (based on the same subset of 83 patients used for the DL model development) is shown in Table 3.

Studies with different architectures

Studies across multiple models were performed to determine the use of RN10 as the base architecture. The results of other models are stated below, as well as the number of trainable parameters and optimal initial learning rate. Mean CPr AUCROC and 95% confidence intervals were computed as population statistics of 50 models obtained from 5 runs of 10-fold cross-validation at optimal learning rate. The highest performance among the models tested was achieved with RN10 and mRN10. The performance of other models is noted in the table below.

PET masking qualitative results

Qualitative results were evaluated across all subjects and using an iteration of the model trained from a single data split. The qualitative results did not change in a significant manner with repeated training. In qualitative analysis of PET masking results, the region-of-interest mask correctly identified the foreground, while we have found that in

TABLE 2. Diagnostic performance metrics of RN10 and mRN10 as well as *p*-values as determined by McNemar test comparing both models for each task (except AUCROC)

	CPr ***RN10***	**CPr** ***mRN10***	*CPr p-value*	**CLoc** ***RN10***	**CLoc** ***mRN10***	*CLoc p-value*
Sensitivity [95% CI]	0.800 [0.719; 0.877]	**0.909** [0.852; 0.965]	**0.028**	0.365 [0.268; 0.460]	**0.552** [0.453; 0.652]	**0.018**
Specificity [95% CI]	**0.619** [0.411; 0.827]	0.476 [0.263; 0.690]	0.257	0.807 [0.759; 0.854]	**0.811** [0.763; 0.858]	0.910
Positive predictive value [95% CI]	**0.908** [0.847; 0.969]	0.891 [0.830; 0.951]	0.507	0.407 [0.303; 0.511]	**0.515** [0.418; 0.611]	0.089
Negative predictive value [95% CI]	0.394 [0.227; 0.560]	**0.526** [0.302; 0.751]	0.205	0.777 [0.728; 0.827]	**0.833** [0.787; 0.878]	**0.021**
Accuracy [95% CI]	0.767 [0.681; 0.839]	**0.833** [0.756; 0.895]	**0.050**	0.689 [0.638; 0.736]	**0.742** [0.693 0.786]	**0.031**
AUCROC	0.815	**0.849**	/	0.702	**0.770**	/

AUCROC = area under the receiver operating characteristic curve; CPr = classification of presence; CLoc = classification of location; mRN10 = novel *masked-PET Resnet10* model; RN10 = baseline *Resnet10* model

all but 3 subjects, 1 with LL HPTT and 2 LR HPTT, that the mask completely obscured (masked) the original location of HPTT on masked PET. In the 3 subjects with visible HPTT in the masked PET in the original location, the mask still partially obscured the HPTT, as seen in Figure 3, rows d), f) and g).

Figure 2 shows a typical example of mRN10 masking, where HPTT was masked and cannot be distinguished in masked PET image. The network correctly classified the subject in Figure 2 as having lower right HPTT. The region of air outside the patient is masked to approximately 25% of the original PET signal, with mask having a value of approximately 0.25. The high signal from the salivary glands is masked in all cases, whereas signal from the thyroid gland is only partially masked in all cases, as seen in Figure 3.

Discussion

The aim of the study was to evaluate the potential of DL models in classifying HPTT presence and location in FCH-PET studies in the setting of PHPT. For our experiments to be representative of results of such a model in practice, we used data of representative cohort of subjects with PHPT. Classification of FCH-PET studies was performed using multiple common DL models and we found that the simplest among the models tested, RN10, achieved the highest performance. Furthermore, we improve the model's performance by modifying the architecture to include a region-of-interest masking step, which produced a region-of-interest mask, which successfully identified the foreground of PET. The mRN10 achieved superior performance to models of similar size. Overall, given the size of our dataset and achieved performance, we found that the use of deep learning is highly promising in potential evaluation of FCH-PET in PHPT.

Dataset and patient characteristics

Both our *patients* and the *controls* had representative demographic characteristics of patients with PHPT, with male-to-female ratio in literature being 1:3 to 1:4 and the peak incidence of 62 ± 13 years.[54-57] Therefore, the models were more likely to have learned the correct features to classify HPTT presence and were trained on a relatively representative dataset that would be encountered in real-life

TABLE 3. Comparison of mRN10 and human performance for the CLoc task. *p*-values were determined by using the McNemar test

	CLoc *mRN10*	**CLoc** *human*	*p-value*
Sensitivity [95% CI]	0.552 [0.453; 0.652]	**0.917** [0.857; 0.958]	**< 0.001**
Specificity [95% CI]	0.811 [0.763; 0.858]	**0.997** [0.986; 0.999]	**< 0.001**
Positive predictive value [95% CI]	0.515 [0.418; 0.611]	**0.992** [0.945; 0.999]	**< 0.001**
Negative predictive value [95% CI]	0.833 [0.787; 0.878]	**0.972** [0.952; 0.984]	**< 0.001**
Accuracy [95% CI]	0.742 [0.693; 0.786]	**0.977** [0.960; 0.988]	**< 0.001**

CLoc = classification of location; mRN10 = novel *masked-PET Resnet10* model; RN10 = baseline *Resnet10* model

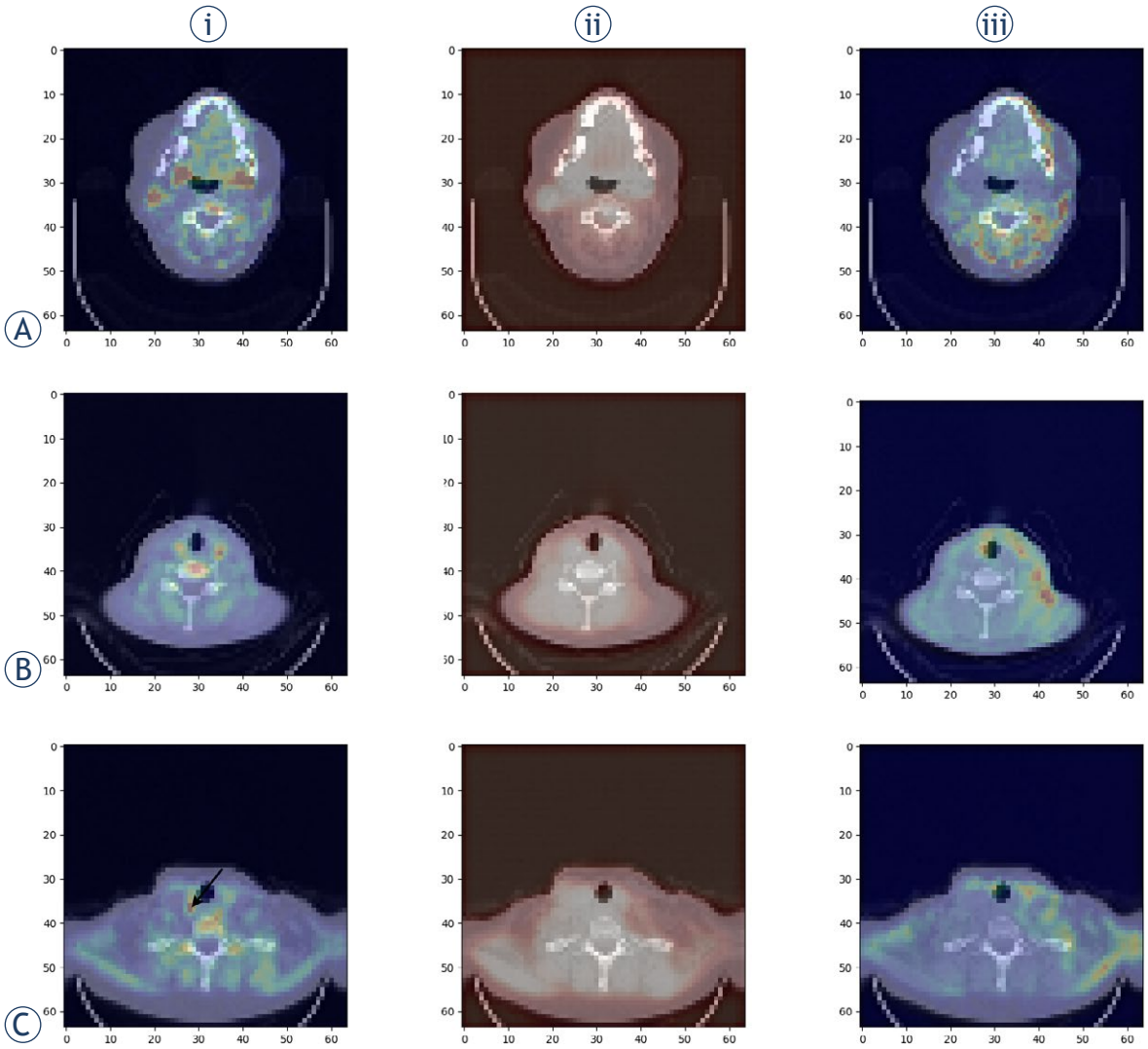

FIGURE 2. Example of novel masked-PET Resnet10 model (mRN10) masking of PET signal in a subject with parathyroid adenoma in the region of lower right parathyroid gland (black arrow in row c). Each row represents a different slice through the pre-processed [18F]fluorocholine PET/CT (FCH-PET) images (**(A)** – mandibular region, **(B)** – upper neck region **(C)** – lower neck region containing parathyroid adenoma). The first column shows a pre-processed PET/CT image (64 × 64 × 32 matrix), where colours toward the "warm" (red) part of the spectrum indicate higher PET signal and colours toward the "cool" (blue) part of the spectrum indicate lower PET signal. The second column shows the mask, where regions coloured toward the red part of the spectrum have higher weights (non-masked) and regions toward the yellow part of the spectrum have lower weights (masked). The third column represents the final masked PET/CT images computed by multiplying the mask with the original PET/CT. The image was correctly classified as containing the adenoma in the lower right region.

application. Representation per quadrant of HPTT in our cohort was also congruous to numbers reported in the literature. Marzouki *et al*. provide 95% confidence intervals of HPTT ratio per site as follows: lower left 32–51%, lower right 25–42%, upper left 10–23% and upper right 4–15%.[58-60]

Unfortunately, the dataset was imbalanced with respect to patients vs "controls". However, obtaining negative FCH-PET studies is difficult due to high positivity rate of finding HPTT in FCH-PET, since only patients with biochemically confirmed PHPT are imaged. Such patients are highly likely to have visible HPTT, as reported in studies exploring the effectiveness of FCH-PET.[3-13] Since healthy subjects are generally not referred to undergo FCH-PET imaging, the best attempt was made to select the criteria for choosing "controls" among patients with negative visual assessment of FCH-PET. Our controls therefore had negative imaging findings and biochemical criteria for PHPT

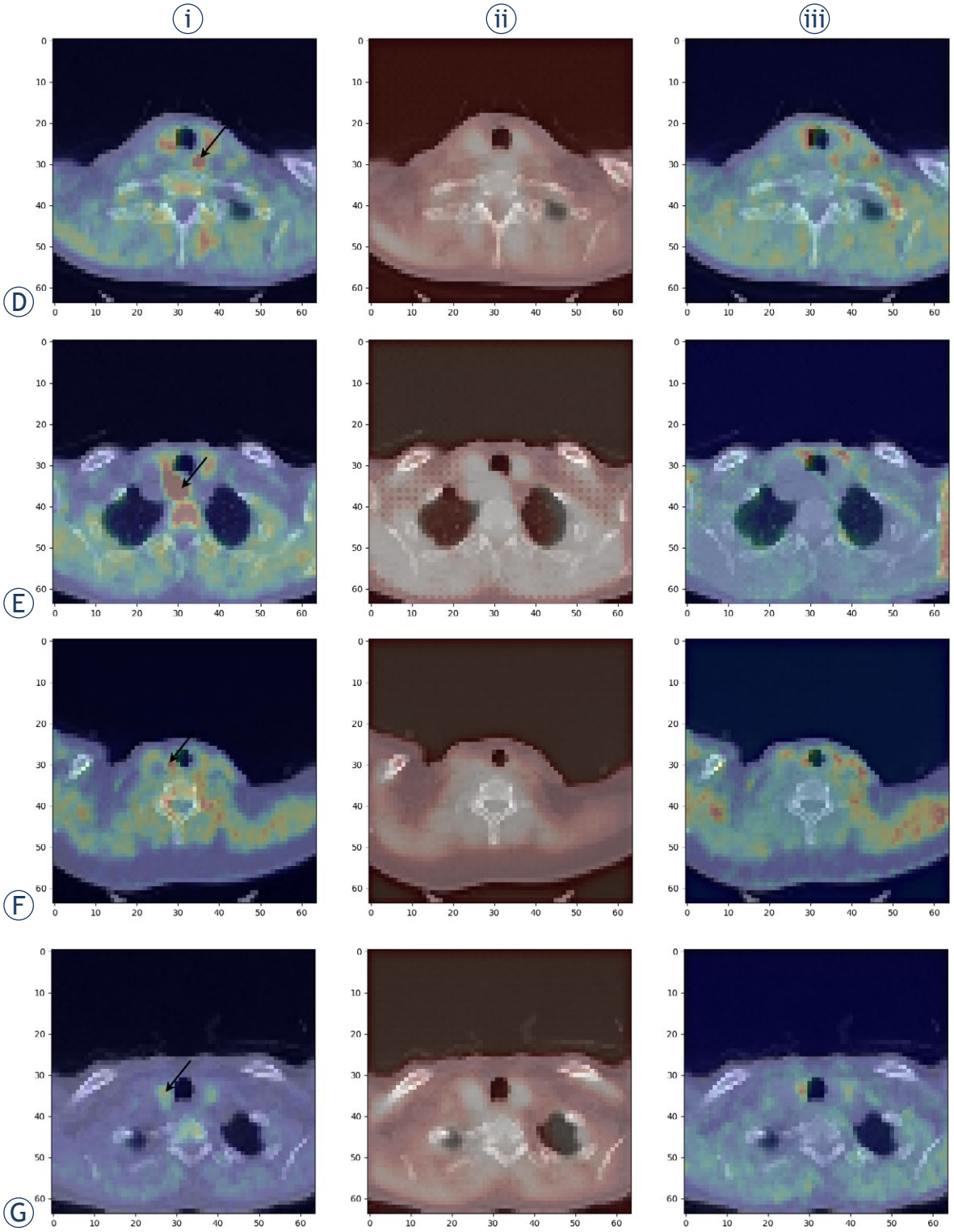

FIGURE 3. Some examples of masking of hyperactive parathyroid tissue (HPTT), which is indicated by an arrow in column **(I)**. The images are shown in the same format as in Figure 2. Rows **(D)**, **(F)** and **(G)** represent the only 3 cases where HPTT was not completely masked.

TABLE 4. Performance of several models on CPr task

Model name	mRN10	RN10	*Resnet50*	*Resnet101*	*Densenet101*	*PreActResnet101*	*WideResnet101*
# Trainable parameters (millions)	33.5	14.3	46.2	85.2	112.9	85.2	85.2
Optimal initial learning rate	0.0136	0.0136	$2.15*10^{-3}$	$1.47*10^{-4}$	0.316	$1.47*10^{-4}$	$2.15*10^{-3}$
Mean CPr AUCROC [95% CI]	**0.850 [0.734; 0.998]**	**0.812 [0.716; 0.994]**	0.754 [0.624; 0.980]	0.527 [0.410; 0.639]	0.703 [0.606; 0.905]	0.739 [0.486; 0.998]	0.752 [0.653; 0.966]

AUCROC = area under the receiver operating characteristic curve; CPr = classification of presence; mRN10 = novel *masked-PET Resnet10* model; RN10 = baseline *Resnet10* model

resolved at follow-up after 6 months without surgical treatment.

For ground truth location, histopathological results were used as opposed to expert visual assessment of FCH-PET, in order to simulate real-world use of the models in guiding surgical removal of HPTT.

Deep-learning model architecture

We have chosen the 3D *Resnet10* as our baseline model since multiple research groups have shown it provides promising results in classification tasks on both medical and non-medical images and is the basis of modern architectures.[41,61-63] *Resnet10* also achieved the highest performance among the models tested. The other tested models with more parameters performed worse, as they seemed overparameterized and likely learned aberrant features, thus overfitting to the training data. Not many studies explore this phenomenon in detail, but a similar phenomenon was noted in the results of a recent study of Bailly *et al.*[64] studying the effects of dataset size, dataset complexity, and model complexity on performance.

The main motivation behind the design of mRN10 and implementation of masking is the way experts interpret FCH-PET. Experienced nuclear medicine physicians know that HPTT usually appears around the thyroid region, and we wanted to allow for the model to learn to mask regions that were deemed unimportant for classification. Furthermore, these unimportant regions (e.g., muscle) commonly produced high intensity PET signal that might affect the classifier. Using end-to-end training with only cross-entropy classification loss, we allowed the network to learn to mask these unimportant regions in an unsupervised manner by carefully tailoring the architecture. Given how experts interpret FCH-PET, mRN10 was an attempt to integrate expert knowledge into the model to improve the *Resnet10* classifier.

The *Unet* was chosen as the masking architecture as we deem our masking to be a task that is comparable to segmentation. For the activation function, we used *tanh* (hyperbolic tangent), since it was shown to be more stable in backpropagation compared to sigmoid function.[65] Since our initial goal was to mask unimportant parts of the image, and *tanh* is a function bound between –1 and 1, we used *tanh* + 1, such that regions where the *Unet* output was very negative were close to 0 and subsequently masked when multiplied by the PET signal intensity. The use of batch normalisation layers in the downstream *Resnet10* in mRN10 ensures stable training even when masked PET is the input, which is not explicitly normalized apriori. The masking *Unet* was trained end-to-end along with *Resnet10* in the mRN10 architecture for optimal performance of the classification task. This was an attempt to explain the classification decision of the classifier by allowing it to optimize for masking of unimportant parts of the image as well as increase the performance by improving the conditioning of the input data to the classifier.[46]

Classification results

One of the goals of the study was to compare the model's performance to nuclear medicine experts. The task of detecting and localizing HPTT on FCH-PET is relatively "trivial" for human experts, with reported accuracies of up to 98%.[3-13] We therefore feel that a small dataset is sufficient for training a model to similar performance. However, the results differed from our expectations, as the achieved performance was significantly below the one of humans for both of our tasks. It is most likely that by increasing the dataset to several hundred subjects, the performance gap would be closed.

Given the size of our dataset, our results are comparable to other published studies on other medical imaging related tasks. The study with a similarly sized dataset (85 subjects) in the classification of cardiac sarcoidosis by Togo *et al*. achieved sensitivity and specificity of 84% and 87%.[66] In line with the established best practice, Lu *et al*. explored the diagnosis of Alzheimer disease from PET and MRI images using a multimodal approach on a dataset of 397 subjects and achieved 93% accuracy at detecting Alzheimer disease; Ma *et al*. used a DL method to classify thyroid diseases from SPECT with a dataset of more than 2000 subjects and achieved accuracy of up to 100% for some tasks.[67,68] Because the aforementioned tasks are different and generally have different difficulty compared to ours, these comparisons and potential conclusions are hypothetical, but they give us a rough estimate of the number of subjects needed to substantially improve the performance of our model.

We feel that by increasing the size of our dataset to several hundred patients, similar levels of performance metrics to human performance could most likely be achieved. One supporting data point for this assumption is that the upper-bound of the 95% CI of AUC in the population statistics of 50 model iterations used in experiments was 0.998. Given the right data split, the model could perfectly classify the test set.

PET mask discussion

Qualitatively, we observed interesting properties of the mask created using the *UNet*, with examples depicted in Figures 2 and 3. In Figure 2 row a), we can see that the physiological signal from the salivary glands was masked, and the weak signal of the paravertebral musculature is amplified. In row b), the physiological signal from the red marrow in the vertebral body was masked and signal from the neck musculature on the left was enhanced. In row c), the physiological signal from the thyroid gland and paravertebral musculature were masked, contradicting findings in row a). The model likely learns to amplify the weak signal from the musculature with low uptake of FCH and to suppress strong signal from salivary glands and certain muscle groups with high uptake.

The physiologically high PET activity in salivary glands and the thyroid were correctly masked. This is likely because there is usually high PET activity in these regions. The masking of the thyroid region is especially problematic since the signal from HPTT can also be masked along with the thyroid. This resulted in HPTT being masked in all but 3 cases, as shown in Figure 3. Still, this did not always result in a false classification of the HPTT location. The parathyroid adenoma in row c) is crucial to the task for experts and yet it was masked in this case by the network. Even though the model masked the adenoma, the mRN10 model output in this case was still correct (lower right adenoma location). It is likely that *UNet* learns to encode the information of adenoma into the mask that is passed to the *Resnet10*.

Regions near the skin and the skin itself were always enhanced – we assumed that this was an important signal to the model, as skin-air interface exhibits high contrast on PET and CT and acts as a rough anatomical landmark. It is also much higher in contrast than soft tissue interfaces of the structures in the parathyroid region and produces stronger gradients in training. The region outside the patient (air) was not masked to 0, but to approximately 25% of the signal (value of mask was 0.25), since it is irrelevant to the classification and likely does not produce a gradient in training, so the *Unet* output for this region is closer to the initialization state.

We find the obtained masks to be interpretable in terms of optimizing downstream *Resnet10*, yet they did not enhance HPTT signal on masked PET as could be expected. Highly active PET regions were therefore always masked (thyroid, salivary glands). The regions which produced high PET activity only in some subjects (musculature) were masked only if they produced high PET activity (Figure 2, row c), if not, these regions were enhanced (Figure 2 row a), introducing noise to masked PET. This further makes the masked PET uninterpretable as the intensity of the introduced noise is higher than the masked signal from the parathyroid adenoma, which can itself be masked. However, in terms of optimizing the *Resnet10* classification performance, these findings make sense, since the mechanism acts to adaptively scale the inputs to stabilize *Resnet10* classifier.

While the proposed mRN10 model, using *Unet* and *Resnet* sequentially for region-of-interest identification and classification tasks, respectively, somewhat resembles the state-of-the-art region proposal algorithms, we have not found such a model presented in existing literature. Firstly, it is unlikely that such architecture would achieve superior performance on other tasks as *Resnet* is a good classifier on its own if it is trained on a large enough database.[20,41] Secondly, the masking results we achieved did not appear to consistently add

value to FCH-PET interpretation when explored by humans, however, according to our results, the mask can be clearly interpreted in terms of optimizing downstream *Resnet10* performance.

Namely, we found the mRN10 to be superior in performance to the RN10 in CLoc task. This is probably due to the improved conditioning of the masked input to *Resnet10* in mRN10, leading to increased stability, which in turn increases the performance of the trained model.[46]

Limitations of the study

In the model selection, we found that the model with lowest number of parameters performed the best. This is one limitation of our study since experiments with even simpler models were not carried out. Another potential performance improvement could be using transfer learning, but we have not found suitable pretrained models for the FCH-PET images.

Our PET masking was an attempt to make the model more interpretable. Most notable similar mechanisms that exist within literature are the attention mechanisms.[69] The main problem with most attention mechanisms is that they rely on weighing of the image features, which are obtained by embedding a small image patch into a vector. Because of this, the spatial resolution of the attention map is limited by the size of the image patch, which is commonly 16 × 16 in visual transformers.[70] In analogy, if we used 16 × 16 × 16 for our theoretical attention, the feature map of our entire image would be of spatial dimensions 4 × 4 × 2, which is too low detailed enough interpretation. Another method of explaining the model output is the class activation mapping (CAM), which also relies on feature embeddings before fully connected layers and therefore entails a loss of spatial resolution;[71] in case of the RN10, the CAM resolution would be 4 × 4 × 2. Gradient-based attribution methods, which do provide pixel-level (or in our case voxel-level) input attribution to the model output, have received criticism due to their inconsistency and poor theoretical foundations.[72]

Conclusions

We provide extensive experiments in deep learning analysis of FCH-PET using standard classification model RN10 and a novel architecture tailored to the task. As deep learning for FCH-PET analysis in PHPT has to our knowledge not yet been described in literature, our experiments provide a baseline for future work. Even though inferior performance to human experts was achieved, the results seem very promising considering the small dataset and the achieved accuracy of 83% for detecting HPTT and 74% accuracy for localizing the quadrant of HPTT.

References

1. Fraser WD. Hyperparathyroidism. *Lancet* 2009; **374:** 145-58. doi: 10.1016/s0140-6736(09)60507-9
2. Grimelius L, Akerström G, Johansson H, Bergström R. Anatomy and histopathology of human parathyroid glands. *Pathol Annu* 1981; **16(Pt 2)**: 1-24. PMID: 7036057
3. Cuderman A, Senica K, Rep S, Hocevar M, Kocjan T, Sever et al. ^{18}F-Fluorocholine PET/CT in primary hyperparathyroidism: superior diagnostic performance to conventional scintigraphic imaging for localization of hyperfunctioning parathyroid glands. *J Nucl Med* 2019; **61:** 577-83. doi: 10.2967/jnumed.119.229914
4. Lezaic L, Rep S, Sever MJ, Kocjan T, Hocevar M, Fettich J. 18F-Fluorocholine PET/CT for localization of hyperfunctioning parathyroid tissue in primary hyperparathyroidism: a pilot study. *Eur J Nucl Med Mol Imaging* 2014; **41:** 2083-9. doi: 10.1007/s00259-014-2837-0
5. Graves CE, Hope TA, Kim J, Pampaloni MH, Kluijfhout W, Seib CD, et al. Superior sensitivity of 18F-fluorocholine: PET localization in primary hyperparathyroidism. *Surgery* 2022; **171:** 47-54. doi: 10.1016/j.surg.2021.05.056
6. Michaud L, Balogova S, Burgess A, Ohnona J, Huchet V, Kerrou K, et al. A pilot comparison of 18F-fluorocholine PET/CT, ultrasonography and 123I/99mTc-sestaMIBI dual-phase dual-isotope scintigraphy in the preoperative localization of hyperfunctioning parathyroid glands in primary or secondary hyperparathyroidism. *Medicine* 2015; **94:** e1701. doi: 10.1097/md.0000000000001701
7. Kluijfhout WP, Vorselaars WM, van den Berk SA, Vriens MR, Borel Rinkes IH, Valk GD, et al. Fluorine-18 fluorocholine PET-CT localizes hyperparathyroidism in patients with inconclusive conventional imaging. *Nucl Med Commun* 2016; **37:** 1246-52. doi: 10.1097/mnm.0000000000000595
8. Kluijfhout WP, Pasternak JD, Drake FT, Beninato T, Gosnell JE, Shen WT, et al. Use of PET tracers for parathyroid localization: a systematic review and meta-analysis. *Langenbecks Arch Surg* 2016; **401:** 925-35. doi: 10.1007/s00423-016-1425-0
9. Thanseer N, Bhadada SK, Sood A, Mittal BR, Behera A, Gorla A K R, et al. Comparative effectiveness of ultrasonography, 99mTc-sestamibi, and 18F-fluorocholine PET/CT in detecting parathyroid adenomas in patients with primary hyperparathyroidism. *Clin Nucl Med* 2017; **42:** e491-7. doi: 10.1097/rlu.0000000000001845
10. Whitman J, Allen IE, Bergsland EK, Suh I, Hope TA. Assessment and comparison of ^{18}F-Fluorocholine PET and ^{99m}Tc-sestamibi scans in identifying parathyroid adenomas: a metaanalysis. *J Nucl Med* 2021; **62:** 1285-91. doi: 10.2967/jnumed.120.257303
11. Beheshti M, Hehenwarter L, Paymani Z, Rendl G, Imamovic L, Rettenbacher R, et al. 18F-Fluorocholine PET/CT in the assessment of primary hyperparathyroidism compared with 99mTc-MIBI or 99mTc-tetrofosmin SPECT/CT: a prospective dual-centre study in 100 patients. *Eur J Nucl Med Mol Imaging* 2018; **45:** 1762-71. doi: 10.1007/s00259-018-3980-9
12. Broos WAM, Wondergem M, Knol RJJ, Van der Zant FM. Parathyroid imaging with 18F-fluorocholine PET/CT as a first-line imaging modality in primary hyperparathyroidism: a retrospective cohort study. *EJNMMI Res* 2019; **9:** 72. doi: 10.1186/s13550-019-0544-3
13. Hope TA, Graves CE, Calais J, Ehman EC, Johnson GB, Thompson D, et al. Accuracy of 18 F-fluorocholine PET for the detection of parathyroid adenomas: prospective single-center study. *J Nucl Med* 2021; **62:** 1511-6. doi: /10.2967/jnumed.120.256735

14. Rep S, Hocevar M, Vaupotic J, Zdesar U, Zaletel K, Lezaic L. ^{18}F-choline PET/CT for parathyroid scintigraphy: significantly lower radiation exposure of patients in comparison to conventional nuclear medicine imaging approaches. *J Radiol Prot* 2018; **38:** 343-56. doi: 10.1088/1361-6498/aaa86f

15. Li Y, Sixou B, Peyrin F. A review of the deep learning methods for medical images super resolution problems. *IRBM* 2021; **42:** 120-33. doi: 10.1016/j.irbm.2020.08.004

16. Yang W, Zhang X, Tian Y, Wang W, Xue J-H, Liao Q. Deep learning for single image super-resolution: a brief review. *IEEE Trans Multimedia* 2019; **21:** 3106-21. doi: 10.1109/tmm.2019.2919431

17. Wang L, Chen W, Yang W, Bi F, Yu FR. A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access* 2020; **8:** 63514-37. doi: 10.1109/access.2020.2982224

18. Liu B, Liu J. Overview of image denoising based on deep learning. *J Phys Conf Ser* 2019; **1176:** 022010. doi: 10.1088/1742-6596/1176/2/022010

19. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep learning: a primer for radiologists. *RadioGraphics* 2017; **37:** 2113-31. doi: 10.1148/rg.2017170077

20. Al-Saffar AAM, Tao H, Talab MA. Review of deep convolution neural network in image classification. In: *2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunication.* IEEE 2017. p. 26-31. doi: 10.1109/icramet.2017.8253139

21. Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: a survey. [Internet]. *arXiv:* 2001.05566 2020. Available from: https://doi.org/10.48550/arXiv.2001.05566

22. Jiao L, Zhang F, Liu F, Yang S, Li L, Feng Z, et al. A survey of deep learning-based object detection. [Internet]. *arXiv:* 2019. Available from: http://arxiv.org/abs/1907.09408

23. Sahlsten J, Jaskari J, Kivinen J, Turunen L, Jaanio E, Hietala K, et al. Deep learning fundus image analysis for diabetic retinopathy and macular edema grading. *Sci Rep* 2019; **9:** 10750. doi: 10.1038/s41598-019-47181-w

24. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115-8. doi: 10.1038/nature21056

25. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proc Conf AAAI Artif Intell* 2019; **33:** 590-7. doi: 10.1609/aaai.v33i01.3301590

26. Nie D, Cao X, Gao Y, Wang L, Shen D. Estimating CT image from MRI data using 3D fully convolutional networks. *Deep Learn Data Label Med Appl* 2016; **2016:** 170-8. doi: 10.1007/978-3-319-46976-8_18

27. Torrado-Carvajal A, Vera-Olmos J, Izquierdo-Garcia D, Catalano OA, Morales MA, Margolin J, et al. Dixon-VIBE Deep Learning (DIVIDE) pseudo-CT synthesis for pelvis PET/MR attenuation correction. *J Nucl Med* 2019; **60:** 429-35. doi: 10.2967/jnumed.118.209288

28. Guo R, Hu X, Song H, Xu P, Xu H, Rominger A, et al. Weakly supervised deep learning for determining the prognostic value of 18F-FDG PET/CT in extranodal natural killer/T cell lymphoma, nasal type. *Eur J Nucl Med Mol Imaging* 2021; **48:** 3151-61. doi: 10.1007/s00259-021-05232-3

29. Hwang D, Kang SK, Kim KY, Seo S, Paeng JC, Lee DS, et al. Generation of PET attenuation map for whole-body time-of-flight ^{18}F-FDG PET/MRI using a deep neural network trained with simultaneously reconstructed activity and attenuation maps. *J Nucl Med* 2019; **60:** 1183-9. doi: 10.2967/jnumed.118.219493

30. Liu F, Jang H, Kijowski R, Bradshaw T, McMillan AB. Deep learning MR imaging-based attenuation correction for PET/MR imaging. *Radiology* 2018; **286:** 676-84. doi: 10.1148/radiol.2017170700

31. Leynes AP, Yang J, Wiesinger F, Kaushik SS, Shanbhag DD, Seo Y, et al. Zero-Echo-Time and Dixon Deep Pseudo-CT (ZeDD CT): direct generation of pseudo-CT images for pelvic PET/MRI attenuation correction using deep convolutional neural networks with multiparametric MRI. *J Nucl Med* 2018; **59:** 852-8. doi: 10.2967/jnumed.117.198051

32. Blanc-Durand P, Van Der Gucht A, Schaefer N, Itti E, Prior JO. Automatic lesion detection and segmentation of 18F-FET PET in gliomas: a full 3D U-Net convolutional neural network study. *PLoS One* 2018; **13:** e0195798 doi: 10.1371/journal.pone.0195798

33. Zhao X, Li L, Lu W, Tan S. Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network. *Phys Med Biol* 2018; **64:** 015011 doi: 10.1088/1361-6560/aaf44b

34. Zhong Z, Kim Y, Plichta K, Allen BG, Zhou L, Buatti J, et al. Simultaneous cosegmentation of tumors in PET-CT images using deep fully convolutional networks. *Med Phys* 2019; **46(2):** 619-33. doi: 10.1002/mp.13331

35. Schwyzer M, Ferraro DA, Muehlematter UJ, Curioni-Fontecedro A, Huellner MW, von Schulthess GK, et al. Automated detection of lung cancer at ultralow dose PET/CT by deep neural networks – initial results. *Lung Cancer* 2018; **126:** 170-3. doi: 10.1016/j.lungcan.2018.11.001

36. Hatt M, Laurent B, Ouahabi A, Fayad H, Tan S, Li L, et al. The first MICCAI challenge on PET tumor segmentation. *Med Image Anal* 2018; **44:** 177-95. doi: 10.1016/j.media.2017.12.007

37. Student. The probable error of a mean. *Biometrika* 1908; **6:** 1. doi: 10.2307/2331554

38. Pearson K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond Edinb Dublin Philos Mag J Sci* 1900; **50:** 157-75. doi: 10.1080/14786440009463897

39. Jones E, Oliphant T, Peterson P, Others. SciPy.org. *SciPy Open source Sci. tools Python2*. 2001.

40. Good IJ. Rational decisions. *J R Stat Soc Ser B* 1952; **14:** 107-14. doi: 10.1111/j.2517-6161.1952.tb00104.x

41. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016. doi: 10.1109/cvpr.2016.90

42. Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* 2017. doi: 10.1109/iccvw.2017.373

43. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2017. doi: 10.1109/cvpr.2017.243

44. Zagoruyko S, Komodakis N. Wide residual networks. *Procedings of the British Machine Vision Conference 2016;* 2016. doi: 10.5244/c.30.87

45. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. *Computer Vision – ECCV 2016.* **2016:** 630-45. doi: 10.1007/978-3-319-46493-0_38

46. Full stack deep learning. Lecture 1: DL fundamentals [Internet]. Fullstackdeeplearning.com. [cited 2022 Aug 28]. Available from: https://fullstackdeeplearning.com/spring2021/lecture-1/

47. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A. editors. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science* 2015; **9351:** 234-41. Cham: Springer.doi: 10.1007/978-3-319-24574-4_28

48. Rossum G Van, Drake FL. Python Tutorial, Technical Report CS-R9526. *Cent voor Wiskd en Inform* 1995.

49. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, et al.. Automatic differentiation in PyTorch. *31st Conf Neural Inf Process Syst* 2017.

50. Stevenson M, Sergeant E, Nunes T, Heuer C, Marshall J, Sanchez J, et al. *epiR: Tools for the analysis of epidemiological data.* v1.0-15. 2020. [cited 2022 Mar 15]. Available at: https://CRAN.R-project.org/package=epiR

51. R Development Core Team. *R: a language and environment for statistical computing*. Vienna; R Foundation for Statistical Computing. Available at: . http://www.R-project.org

52. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; **12:** 153-7. doi: 10.1007/bf02295996

53. Stock C, Hielscher T. DTComPair: comparison of binary diagnostic tests in a paired study design. R package version 1.0.3. [Internet]. 2014. Available from: http://cran.r-project.org/package=DTComPair

54. Rao SD. Epidemiology of parathyroid disorders. *Best Pract Res Clin Endocrinol Metab* 2018; **32:** 773-80. doi: 10.1016/j.beem.2018.12.003

55. Somnay YR, Craven M, McCoy KL, Carty SE, Wang TS, Greenberg CC, et al. Improving diagnostic recognition of primary hyperparathyroidism with machine learning. *Surgery* 2017;**161:** 1113-21. doi: 10.1016/j.surg.2016.09.044

56. Press DM, Siperstein AE, Berber E, Shin JJ, Metzger R, Monteiro R, et al. The prevalence of undiagnosed and unrecognized primary hyperparathyroidism: a population-based analysis from the electronic medical record. *Surgery* 2013; **154:** 1232-8. doi: 10.1016/j.surg.2013.06.051

57. Bilezikian JP, Marcus R, Levine MA, Marcocci C, Silverberg SJ, Potts JT, editors. *Parathyroids: basic and clinical concepts*. 3rd edition. 2014. Elsevier, Academic Press.

58. Marzouki HZ, Chavannes M, Tamilia M, Hier MP, Black MJ, Levental M, et al. Location of parathyroid adenomas: 7-year experience. *J Otolaryngol Head Neck Surg* 2010; **39:** 551-4. PMID: 20828518

59. Filser B, Uslar V, Weyhe D, Tabriz N. Predictors of adenoma size and location in primary hyperparathyroidism. *Langenbeck's Arch Surg* 2021; **406:** 1607. doi: 10.1007/s00423-021-02179-9

60. Shah VN, Bhadada SK, Bhansali A, Behera A, Mittal BR. Changes in clinical & biochemical presentations of primary hyperparathyroidism in India over a period of 20 years. *Indian J Med Res* 2014; **139:** 694-9. PMID: 25027078

61. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. *Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017* 2017. doi: 10.1109/cvpr.2017.634

62. Gao S, Cheng MM, Zhao K, Zhang XY, Yang MH, Torr PHS. Res2Net: a new multi-scale backbone architecture. *IEEE Trans Pattern Anal Mach Intell* 2019. doi: 10.1109/TPAMI.2019.2938758

63. Chen S, Tan X, Wang B, Hu X. Reverse attention for salient object detection. *Computer Vision – ECCV 2018* 2018; 236-52. doi: 10.1007/978-3-030-01240-3_15

64. Bailly A, Blanc C, Francis É, Guillotin T, Jamal F, Wakim B, et al. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. *Comput Methods Programs Biomed* 2022; **213:** 106504 doi: 10.1016/j.cmpb.2021.106504

65. Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Comput* 1997; **9:** 1735-80. doi: 10.1162/neco.1997.9.8.1735

66. Togo R, Hirata K, Manabe O, Ohira H, Tsujino I, Magota K, et al. Cardiac sarcoidosis classification with deep convolutional neural network-based features using polar maps. *Comput Biol Med* 2019; **104:** 81-6. doi: 10.1016/j.compbiomed.2018.11.008

67. Lu D, Popuri K, Ding GW, Balachandar R, Beg MF. Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimer's disease. *Med Image Anal* 2018; **46:** 26-34. doi: 10.1016/j.media.2018.02.002

68. Ma L, Ma C, Liu Y, Wang X. Thyroid diagnosis from SPECT images using convolutional neural network with optimization. *Comput Intell Neurosci* 2019; **2019:** 6212759. doi: 10.1155/2019/6212759

69. Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. *Neurocomputing* 2021; **452:** 48-62. doi: 10.1016/j.neucom.2021.03.091

70. Liu Y, Zhang Y, Wang Y, Hou F, Yuan J, Tian J, et al. A survey of visual transformers. *arXiv* [csCV] [Internet]. 2021 [cited 2022 Aug 28]; Available from: http://arxiv.org/abs/2111.06091

71. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. *arXiv* [csCV] [Internet]. 2015 [cited 2022 Aug 28]; Available from: http://arxiv.org/abs/1512.04150

72. Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv* [csLG] [Internet]. 2017 [cited 2022 Aug 28]; Available from: http://arxiv.org/abs/1711.06104