Automated deep identification of radiopharmaceutical type and body region from PET images

Ali Ghafari[1, 2], Peyman Sheikhzadeh[1,3], Negisa Seyyedi[4], Mehrshad Abbasi[3], Shadab Ahamed[5,6], Mohammad Reza Ay[1,7], Arman Rahmim[5,6]

[1]Department of Medical Physics and Biomedical Engineering, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran
[2]Research Center for Evidence-Based Medicine, Tabriz University of Medical Sciences, Tabriz, Iran
[3]Department of Nuclear Medicine, Vali-Asr Hospital, Tehran University of Medical Sciences, Tehran, Iran
[4]Nursing Care Research Center, Iran University of Medical Sciences, Tehran, Iran
[5]Department of Integrative Oncology, BC Cancer Research Institute, Vancouver, BC, Canada
[6]Department of Physics and Astronomy, University of British Columbia, Vancouver, BC, Canada
[7]Research Center for Molecular and Cellular Imaging, Tehran University of Medical Sciences, Tehran, Iran

Corresponding author:
Dr. Peyman Sheikhzadeh
Department of Nuclear Medicine, Vali-Asr Hospital, IKHC, Tehran University of Medical Sciences, Tehran, Iran
E-mail: sheikhzadeh-p@sina.tums.ac.ir

Running title: PET image identification of radiopharmaceutical type and body region

Article History:
Received: 26 August 2023
Revised: 13 January 2024
Accepted: 14 January 2024
Published 17 March 2024

ABSTRACT

Introduction: A deep learning pipeline consisting of two deep convolutional neural networks (DeepCNN) was developed, and its capability to differentiate uptake patterns of different radiopharmaceuticals and to further categorize PET images based on the body regions was explored.

Methods: We trained two sets of DeepCNN to determine (i) the type of radiopharmaceutical ([^{18}F]FDG and [^{68}Ga]Ga-PSMA) used in imaging (i.e., a binary classification task), and (ii) body region including head and neck, thorax, abdomen, and pelvis (i.e., a 4-class classification task), using the 2D axial slices of PET images. The models were trained and tested for five different scan durations, thus studying different noise levels.

Results: The accuracy of the binary classification models developed for different scan duration levels was 98.9%–99.6%, and for the 4-class classification models in the range of 98.3%–99.9 ([^{18}F]FDG) and 97.8%–99.6% ([^{68}Ga]Ga-PSMA).

Conclusion: We were able to reliably detect the type of radiopharmaceutical used in PET imaging and the body region of the PET images at different scan duration levels. These deep learning (DL) models can be used together as a preliminary input pipeline for the use of models specific to a type of radiopharmaceutical or body region for different applications and for extracting appropriate data from unclassified images.

Keywords: Positron emission tomography; Deep learning; Categorization; [^{18}F]FDG; [^{68}Ga]Ga-PSMA

INTRODUCTION

Medical imaging is an essential tool for diagnosing and determining the prognosis of different diseases, which provides valuable information about patients' health status by acquiring quantitative, semi-quantitative, and qualitative data from the regions of interest. Nuclear medicine imaging utilizes nuclear interactions of the matter (i.e., injected radionuclide or radiopharmaceutical) as a medium for imaging and can thus provide molecular and metabolic information from the region scanned. Positron emission tomography (PET) is a subset of nuclear medicine imaging techniques commonly used to acquire metabolic data in a wide range of pathologies, including tumors, and offers numerous applications in oncology, neurology, and cardiology [1–4]. Fluorodeoxyglucose ([^{18}F]FDG) and gallium-68-prostate-specific membrane antigen ([^{68}Ga]Ga-PSMA) are examples of radiopharmaceuticals used in positron emission tomography/computed tomography (PET/CT) imaging to acquire images of different body parts and enable the monitoring of disease progression and treatment planning [5,6].

Artificial intelligence (AI) methods, including machine learning (ML) and deep learning (DL) [7], have found their role in medical imaging research with diverse applications such as classification, segmentation, super-resolution, and low-vision problems [4, 8–17].

Successful implementation of AI methods depends on the quality (i.e., balanced dataset, noise-free images, etc.) and quantity of the data. As for most medical imaging applications of AI, there has always been a lack of labeled data to be used in supervised learning tasks hindering the satisfactory development of AI models.

Increased dependence on medical imaging techniques for disease diagnosis and treatment planning over the years have resulted in acquiring massive collections of patient data, including images acquired via different imaging modalities, such as magnetic resonance imaging (MRI), computed tomography (CT), nuclear medicine imaging, etc. The manual organization of these large datasets requires considerable time and effort, and this tedious work is often prone to erroneous and disorganized datasets when performed manually because of the declining human performance with an increase in workload.

Categorizing PET images based on the radiopharmaceutical used and body region can efficiently organize the available image data for developing more robust and reliable AI models with the help of a more sophisticated dataset. Previously, DL models have been used to classify PET images based on several aspects. Wang et al. [18] investigated the strategies of adapting a previously developed automatic anatomy recognition (AAR) [19] system using fuzzy models to PET ([^{18}F]FDG) and low-dose CT in three categories of thoracic, abdominal, and pelvic regions. By evaluating size estimation and localization errors, they achieved noticeable results. Qayyum et al. [20] used a DeepCNN to classify multi-modality images into 24 organ-based classes and achieved an average accuracy of 99.77% and mean average precision of 0.69%. To the best of our knowledge, there has been no study classifying PET images based on the radiopharmaceutical type at different time scan levels with various peak signal-to-noise ratios (PSNR) and noise levels.

In the current study, we employed DeepCNNs to provide an input data pipeline capable of discriminating [^{18}F]FDG and [^{68}Ga]Ga-PSMA PET axial images (binary classification) and then categorizing these images into four anatomical regions including head and neck, thorax, abdomen, and pelvis (4-class classification). To simulate different noise level present in axial images, we developed different models using images post-reconstructed with various scan durations (standard, one-half, one-fourth, one-eighths, and one-sixteenth). Furthermore, we tested different combination of DL network hyperparameters to find the best combination appropriate for the specific application.

The proposed models can determine the type of radiopharmaceutical used in imaging and categorize axial images based on the body region; thus, they can be used for automatic categorization and archiving of PET images available at different noise levels for two radiopharmaceuticals, alleviating the problem of the considerable amount of unused data. Automating the tedious but straightforward process of image labeling based on radiopharmaceuticals and the body regions. As for the different noise levels covered in this study, the developed models can be adopted to categorize images acquired from various PET imaging devices, thereby broadening the generalizability of the proposed models.

This paper is structured as follows: Section 2 elaborates the specific procedures undertaken to prepare the dataset, briefly introduces the necessary concepts, and explains the methods in detail. Results are presented in Section 3 and discussed in Section 4. Finally, conclusions are drawn in Section 5.

METHODS

Dataset

The dataset is consisted of PET axial images of 20 patients (10 for each radiopharmaceutical), acquired with two radiopharmaceuticals ([^{18}F]FDG or [^{68}Ga]Ga-PSMA) using a GE Discovery PET/CT scanner. For each radiopharmaceutical, the patients were scanned using the standard scan duration (denoted with S_1) and retrospectively reconstructed using one-half ($S_{1/2}$), one-fourth ($S_{1/4}$), one-eighth ($S_{1/8}$), and one-sixteenth ($S_{1/16}$) of the standard scan duration to mitigate different noise levels. For 4-class classification, each dataset was further split into four balanced categories using corresponding computed tomography (CT) images of PET slices: head and neck, thorax, abdomen, and pelvis. For binary classification, the corresponding sub-series of each radiopharmaceutical (e.g., S_1) were compiled to prepare a balanced dataset comprising two classes: [^{18}F]FDG and [^{68}Ga]Ga-PSMA. More details on patient data and datasets are presented in Tables 1 and 2. Peak signal-to-noise ratio (PSNR) was calculated by considering sub-series S_1 as Ground Truth.

Architecture of the input pipeline and deep learning models

This input pipeline consists of two separate sets of DeepCNNs developed for binary classification specifying the type of radiopharmaceutical used in scanning and 4-class classification of images acquired with each radiopharmaceutical into categories of head and neck (abbreviated as H_N), thorax, abdomen, and pelvis.

Based on Figure 1, the architectures of the models developed for binary and 4-class classification tasks differ. Python programming language (Version 3.7.10) was used to harness Keras API (Version 2.3.1) on the TensorFlow Backend (Version 2.1.0). Image files were originally DICOM files, but for the training and testing purposes and privacy concerns, only the pixel data of the files were extracted and saved as JPEG images and then used in training and testing. The models received two-dimensional axial images with three channels (192, 192, 3) as the input without any specific pre-processing. The models used in this study were composed of four consecutive blocks, each block containing two components: 1) a 2-D convolutional layer (with rectified linear unit (ReLU) activation function) followed by 2) a maximum pooling layer with a stride of 2 and pooling size of (2, 2). These four consecutive layers were followed by two fully-connected (dense) layers. The activation function of the first dense layer was ReLU, while for outputting a probability, the activation function of the second dense layer was the Softmax function for 4-class classification and the sigmoid function for binary classification. Another difference between the models developed for binary and 4-class classification was the filter values of each successive convolutional layer. Unlike 4-class classification models, an L2 layer weight regularizer was used for each convolutional layer of binary classification network. For binary classification, the convolutional kernel size was constant (3, 3), whereas, for the 4-class classification task, two kernel sizes (3, 3) and (5, 5) were explored.

The models were implemented using an NVIDIA GeForce GTX 950M Graphics Processing Unit (GPU) with 4 GB of dedicated memory and an Intel(R) Core (TM) i7-4720HQ CPU with a base frequency of 2.60GHz.

Training and testing strategies for deep learning methods

Binary classification

Models with the same architecture were developed for each sub-series (i.e., S_1 –$S_{1/_{16}}$) prepared for this task. A 5-fold nested cross-validation strategy [21] was followed for training and testing, i.e., in each fold 70% of the data was dedicated to training, 10% to validation, and the remaining 20% to testing. Details about the multiple hyperparameter value settings of the implemented models and other training and testing properties are given in Table 3.

4-Class classification

For this 4-class classification task, we followed a different path. For each sub-series of radiopharmaceuticals, a model with the same basic architecture but different hyperparameters (batch size, learning rate, and kernel size) was repeatedly trained for three times to find the combination offering better results which resulted in 360 trained models (when repeated for three times). A 5-fold nested cross-validation strategy was used (70% of the data for training, 10% for validation, and the remaining 20% for testing) (Refer to Table 3 for further details including training times).

Classification evaluation metrics

The metrics used for evaluating the performance of DL the classification models measure various aspects of their performance. Thus, we evaluated ten metrics (accuracy, precision, sensitivity, specificity, negative predictive value, false positive rate, false negative rate, F1-score, Matthews correlation coefficient, and area under the receiver operating characteristic curve (AUC)) for each of the binary/4-class classification model; however, for the sake of brevity, we will only report average accuracy, recall (sensitivity), F1-score, and the (AUC) metrics for all folds, and plot the confusion matrix and ROC curves of the folds with the best results in each sub-series.

Accuracy measures how well and precisely the model predict the class of all the samples present in the dataset, representing the proportion of correct predictions [22, 23] (Equation 1).

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

In equation 1, TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative instances (the acceptance threshold was set to 0.5).

The ratio of the true positive predictions to the total positive predictions is reported using the recall or sensitivity metric [23] (Equation 2).

$$Sensitivity\ (recall) = \frac{TP}{TP+FN} \qquad (2)$$

The harmonic mean of the precision and recall metrics results in another classification performance metric called the F1-score [23] (Equation 3). ROC curve is a tool for measuring the accuracy of methods, and the AUC is a value summarizing it. An AUC value of 0.5 shows that the method has no diagnosis capability, while values close to 1 are preferred [24].

$$F1 - score = \frac{2 \times (recall \times precision)}{(recall + precision)} \quad (3)$$

RESULTS

Binary Classification Results

The purpose of binary classification is to determine the radiopharmaceutical used in PET imaging ([^{18}F]FDG or [^{68}Ga]Ga-PSMA) based on PET images in each noise level (e.g., S_1). We followed a 5-fold nested cross-validation approach for training, validation, and testing. Table 4 summarizes the macro average results of all folds of the binary classification task for each sub-series. Figure 2 presents the confusion matrices of the folds with the best results in the binary classification task in each sub-series. ROC curves are present in Figure 3. The GradCam [12, 25] is a helpful tool to improve the transparency of the decision made by CNNs by showing the regions of the images triggering the CNN to decide. In Figure 4, GradCam illustrations of the binary classification models for different scan durations are plotted.

4-Class classification

For each sub-series of the two radiopharmaceuticals, 12 DeepCNN models were developed using various combinations of hyperparameters, and each model was trained and evaluated three times using a one-held-out validation approach. In Table 5, hyperparameters settings and classification metrics are reported only for the model with the best performance based on confusion matrices. In Figure 5, Confusion Matrices are displayed for each radiopharmaceutical. Figure 6 presents the GradCam illustration of the 4-class classification models in multiple sub-series and various regions of the body. ROC curves are also present in Figure 7. Some combinations of the hyperparameters led to the model collapsing (classifying all classes as one) are summarized in Table 6.

DISCUSSION

We developed an input pipeline consisting of two sets of Deep CNN models for two purposes: 1) determining the radiopharmaceutical used in PET imaging based on PET images (binary classification), 2) categorizing PET images into four anatomical categories, namely head and neck, thorax, abdomen, and pelvis (4-class classification).

Although sub-series $S_{1/_{16}}$ was reconstructed using 1/16th of the standard scan duration of sub-series S_1 in the binary classification task, accuracy decreased from 99.55 to 98.94 (5-fold nested cross-validated and averaged), representing model robustness in dealing with noise level since the PSNR of the acquired images can differ based on scan duration, different patient physiology, etc. Similar to accuracy, recall, and F1-score did not change significantly, from S_1 to $S_{1/_{16}}$, despite the differences in PSNR and scan duration (Table 4).

In Figure 4, the GradCam illustration of the models used for binary classification is depicted in each sub-series and for both radiopharmaceuticals. There are two images for each radiopharmaceutical in this figure, the original image (on the left) and the same image but with GradCam heat maps (on the right). The CNN model decides to classify images based on regions with a noticeable uptake specific to that region. For instance, in sub-series S_1, the model decides to classify images based on radiopharmaceutical uptake in temporal regions of the brain; in the case of [^{68}Ga]Ga-PSMA in sub-series $S_{1/2}$, the model decides based on the uptake in the hepatic region.

Categorizing images into different body regions can help develop more sophisticated tools such as computer-aided diagnosis (CAD) systems and different DL models for super-resolution and segmentation. So 4-class classification results are reported in Table 5. In terms of model hyperparameters, it is evident that a learning rate of 0.0001 and batch size of 20 mostly led to better results. Although the results reported in this table may not be the best quantitative results, they offer a good insight into the outcomes. Figure 8 depicts the scatter plots of 4-class classification models and classification metrics, indicating that most models with different combinations of hyperparameter values had a similar performance.

Table 5 reports the classification metrics of 4-class classification for each sub-series of radiopharmaceuticals. Accuracy decreased from 99.89% (S_1) to 98.30% ($S_{1/16}$) for the [^{18}F]FDG, and from 99.16% (S_1) to 97.84% ($S_{1/16}$) for the [^{68}Ga]Ga-PSMA. These values are the average of all the classes (head and neck, thorax, etc.). Note that these values were higher for classes without any overlapping regions; nevertheless, since there is not a definitive separation between body regions, and there is some tissue overlap (e.g., the liver can be seen in the same axial slice as the lungs), it is expectable for the models to mistake some thoracic slices as abdominal and can be a reason the overall accuracy of classification can decrease to some extent.

Some 4-class classification models collapsed meaning that all classes were classified as one (Table 6). The shared trait of all the collapsed models was a learning rate of 0.001, which could explain their inability to learn. However, the results are not conclusive for other hyperparameters. Moreover, 70% of the crashed models had a batch size of 30, and 82% of them had a kernel size of (5, 5) which was 100% in the case of the models used for classifying [^{18}F]FDG images. Larger kernel sizes mean that the convolutional kernel's scope is broader, which could explain why these models failed to identify delicate local patterns.

This study has some limitations. Firstly, this study is a part of the series of studies to employ DL for PET image processing, and the aim may be confusing at first sight. Nevertheless, the capability to identify the radiopharmaceutical and body sections on images is important because the findings highlight the possibility of DL employment for this purpose and enable the researchers to use DL preprocessing methods for further image manipulations. Second, the [^{68}Ga]Ga-PSMA data came from male patients only, while the [^{18}F]FDG data included both male (5) and female (5) patients. This difference is a source of imbalance in the data because of the different physiologies of the male and female bodies. For an ideal comparison of the performance of classification models, images should be acquired from the same patients imaged with both [^{68}Ga]Ga-PSMA and [^{18}F]FDG, but this is not achievable. Although the randomness of the CNN weights and initialization was revoked by setting the seed of TensorFlow and other packages used in developing the models, it is still evident that the results suffer from randomness, and different images trained and tested on during train and test split procedure (although models of 4-class classification with the same hyperparameters were trained three times, some of them collapsed while the others did not).

There are multiple possibilities to further expand this study. These models were two-dimensional DeepCNNs, and accordingly, their input was two-dimensional (2D) axial images. A three-dimensional (3D) DeepCNN and multi-slice input strategy could be considered, and additional gains from 3D DeepCNNs be explored. Transfer learning and ensemble methods could prove to be much more efficient by increasing the number of images acquired with radiopharmaceuticals other than [^{18}F]FDG and [^{68}Ga]Ga-PSMA, making the classification categories much more sophisticated classifying the images present in a body region group based on different organs. PET images do not provide enough anatomical information which is one of the reasons to use accompanying CT images to provide anatomical context about the scan region of interest. At the same time, even without such anatomical context, our deep identification task performances were excellent. Finally, the 4-class classification task could be further improved to classify all axial and all anatomical regions.

CONCLUSIONS

We developed an input data pipeline consisting of two sets of DeepCNNs for categorizing PET images based on radiopharmaceuticals used in imaging and categorizing images into specific body region groups (head and neck, thorax, abdomen, and pelvis). The models included in the pipeline provided promising results at different noise levels denoted by sub-series S_1 to $S_{1/16}$ to achieve decreasing scan durations during reconstruction. Covering multiple noise levels ensured that trained models were generalized enough to be applicable on the real-world PET images. The smaller number of parameters decreases the possibility of the model memorizing data instead of learning patterns required for seamless image classification, thereby promoting model generalization. DeepCNNs presented in the pipeline were successfully capable of classifying PET images based on radiopharmaceutical used and further categorization according to anatomical region without the help of additional anatomical data from corresponding CT images.

Acknowledgments

This work was supported by the Tehran University of Medical Sciences (grant number: 49551).

REFERENCES

1. Kang J, Gao Y, Shi F, Lalush DS, Lin W, Shen D. Prediction of standard-dose brain PET image by using MRI and low-dose brain [18F]FDG PET images. Med Phys. 2015 Sep;42(9):5301–9.
2. Lei Y, Dong X, Wang T, Higgins K, Liu T, Curran WJ. Whole-body PET estimation from low count statistics using cycle-consistent generative adversarial networks. Phys Med Biol. 2019 Nov;64(21):215017.
3. Kaplan S, Zhu YM. Full-dose PET image estimation from low-dose PET image using deep learning: a pilot study. J Digit Imaging. 2019 Oct;32(5):773–8.
4. Xu J, Gong E, Pauly J, Zaharchuk G. 200x low-dose PET reconstruction using deep learning. arXiv preprint arXiv:171204119. 2017;
5. an Leeuwen PJ, Donswijk M, Nandurkar R, Stricker P, Ho B, Heijmink S, Wit EMK, Tillier C, van Muilenkom E, Nguyen Q, van der Poel HG, Emmett L. Gallium-68-prostate-specific membrane antigen (^{68} Ga-PSMA) positron emission tomography (PET)/computed tomography (CT) predicts complete biochemical response from radical prostatectomy and lymph node dissection in intermediate- and high-risk prostate cancer. BJU Int. 2019 Jul;124(1):62-8.
6. Almuhaideb A, Papathanasiou N, Bomanji J. 18F-FDG PET/CT imaging in oncology. Ann Saudi Med. 2011 Jan-Feb;31(1):3-13.
7. Bradshaw TJ, McMillan AB. Anatomy and physiology of artificial intelligence in PET imaging. PET Clin. 2021 Oct;16(4):471-82.
8. Ghafari A, Monsef A, Sheikhzadeh P. F-18-FDG PET Image quality improvement using a pix2pix conditional generative adversarial network combined with the Bayesian Penalized Likelihood (BPL) reconstruction algorithm. In: Springer One New York Plaza, Suite 4600, New York, NY, United States; 2022. p. S625–6.
9. Ghafari A, Monsef A, Sheikhzadeh P. Image augmentation for image-to-image translation in F-18-FDG PET imaging: does it make a difference? In: Springer One New York Plaza, Suite 4600, New York, NY, United States; 2022. p. S621–S621.
10. Ghafari A, Sheikhzadeh P, Seyyedi N, Abbasi M, Ay MR. Realizing 32-time scan duration reduction of 18F-FDG PET using deep learning model with image augmentation. Frontiers Biomed Technol. 2023 Mar;10(2):195–203.

11. Ghafari A, Sheikhzadeh P, Seyyedi N, Abbasi M, Farzenefar S, Yousefirizi F, Ay MR, Rahmim A. Generation of^{18}F-FDG PET standard scan images from short scans using cycle-consistent generative adversarial network. Phys Med Biol. 2022 Oct 19;67(21):215005.
12. Azar AS, Ghafari A, Najar MO, Rikan SB, Ghafari R, Khamene MF, Sheikhzadeh P. Covidense: providing a suitable solution for diagnosing covid-19 lung infection based on deep learning from chest X-ray images of patients. Frontiers Biomed Technol. 2021 Jun;8(2):131–42.
13. Babaei Rikan S, Sorayaie Azar A, Ghafari A, Bagherzadeh Mohasefi J, Pirnejad H. covid-19 diagnosis from routine blood tests using artificial intelligence techniques. Biomed Signal Process Control. 2022 Feb;72:103263.
14. Liu CC, Qi J. Higher SNR PET image prediction using a deep learning model and MRI image. Phys Med Biol. 2019 May 23;64(11):115004.
15. Liu J, Malekzadeh M, Mirian N, Song TA, Liu C, Dutta J. Artificial intelligence-based image enhancement in PET imaging: noise reduction and resolution enhancement. PET Clin. 2021 Oct;16(4):553–76.
16. Yousefirizi F, Jha AK, Brosch-Lenz J, Saboury B, Rahmim A. Toward high-throughput artificial intelligence-based segmentation in oncological PET imaging. PET Clin. 2021 Oct;16(4):577–96.
17. Yousefirizi F, Pierre Decazes null, Amyar A, Ruan S, Saboury B, Rahmim A. AI-based detection, classification and prediction/prognosis in medical imaging: towards radiophenomics. PET Clin. 2022 Jan;17(1):183–212.
18. Wang H, Udupa JK, Odhner D, Tong Y, Zhao L, Torigian DA. Automatic anatomy recognition in whole-body PET/CT images. Med Phys. 2016 Jan;43(1):613–29.
19. Udupa JK, Odhner D, Zhao L, Tong Y, Matsumoto MM, Ciesielski KC, Falcao AX, Vaideeswaran P, Ciesielski V, Saboury B, Mohammadianrasanani S, Sin S, Arens R, Torigian DA. Body-wide hierarchical fuzzy modeling, recognition, and delineation of anatomy in medical images. Med Image Anal. 2014 Jul;18(5):752-71.
20. Qayyum A, Anwar SM, Awais M, Majid M. Medical image retrieval using deep convolutional neural network. Neurocomputing. 2017 Nov;266:8–20.
21. Bradshaw TJ, Boellaard R, Dutta J, Jha AK, Jacobs P, Li Q, Liu C, Sitek A, Saboury B, Scott PJH, Slomka PJ, Sunderland JJ, Wahl RL, Yousefirizi F, Zuehlsdorff S, Rahmim A, Buvat I. Nuclear medicine and artificial intelligence: best practices for algorithm development. J Nucl Med. 2022 Apr;63(4):500-10.
22. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv preprint arXiv:200805756. 2020;
23. Erdaw Y, Tachbele E. Machine learning model applied on chest X-ray images enables automatic detection of covid-19 cases with high accuracy. Int J Gen Med. 2021 Aug 28;14:4923-4931.
24. Obuchowski NA, Bullen JA. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. Phys Med Biol. 2018 Mar 29;63(7):07TR01.
25. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). 2017. p. 618–26.

Table 1. Demographic information for patients including Sex, Age, and Body Mass Index (BMI) present in our study

Radiopharmaceutical	**Number of Patients**	**Sex**	**Age**			**BMI**		
			Min	Mean ± STD	Max	Min	Mean ± STD	Max
[^{18}F]F-FDG	5	M	18	51 ± 21	75	17.3	25.0 ± 5.1	30.3
	5	F	40	55 ± 12	72	15.0	26.1 ± 7.3	34.1
[^{68}Ga]Ga-PSMA	10	M	59	73 ± 7	86	22.9	25.7 ± 2.7	29.9

BMI: Body Mass Index, Min: Minimum, STD: Standard Deviation, Max: Maximum

Table 2. Datasets for the different classification tasks. For each dataset, scan duration and peak signal-to-noise ratio is present

Classification Task	**Radiopharmaceutical**	**Sub-series**	**Number of axial slices for each category**	**Total Scan Duration (Second) (Mean ± STD)**	**PSNR (Mean ± STD)**			
					Head and Neck	Thorax	Abdomen	Pelvis
4-class	[^{18}F]F-FDG	S_1	550	674.5 ± 101.5	GT	GT	GT	GT
		$S_{1/2}$	550	331.9 ± 36.7	47.30±5.88	47.10±6.14	44.60±6.51	45.51±6.06
		$S_{1/4}$	550	166.9 ± 17.9	42.74±6.43	42.41±6.81	39.98±7.10	40.99±6.52
		$S_{1/8}$	550	85.1 ± 8.6	39.53±6.22	39.28±6.82	36.87±6.97	37.86±6.48
		$S_{1/16}$	550	43.3 ± 4.4	35.70±6.90	35.68±35.68	33.35±7.45	34.60±6.70
	[^{68}Ga]Ga-PSMA	S_1	520	600	GT	GT	GT	GT
		$S_{1/2}$	520	300	53.48 ± 4.09	51.50 ± 1.98	43.30±1.56	50.83±2.24
		$S_{1/4}$	520	150	50.70 ± 4.18	47.41±3.60	40.50±2.81	46.07±3.59
		$S_{1/8}$	520	75	47.42 ± 4.40	44.12±3.64	36.82±2.86	42.88±3.63
		$S_{1/16}$	520	40	44.87 ± 4.34	41.42±3.52	34.10±2.78	40.37±35.70
					[^{18}F]F-FDG		**[^{68}Ga]Ga-PSMA**	
Binary	[^{18}F]F-FDG /[^{68}Ga]Ga-PSMA	S_1	2600	637.25±79.62	GT		GT	
		$S_{1/2}$	2600	315.95±30.12	46.15±6.21		51.60±5.30	
		$S_{1/4}$	2600	158.45±15.05	41.54±6.76		46.39±5.11	
		$S_{1/8}$	2600	80.05±7.84	38.41±6.65		43.04±5.29	
		$S_{1/16}$	2600	41.65±3.45	34.83±7.09		40.43±5.28	

PSNR: Peak signal-to-noise ratio; GT: Ground Truth

Table 3. Model architectures and implementation details

Classification Task	Radiopharmaceutical	Sub-series	Optimizer	Loss Function	Learning Rate	Batch Size	Kernel Size	Epochs	Regularizer	Average Training Time (s) (mean ± STD)	Average Testing Time (s) (mean ± STD)
Binary	[^{18}F]F-FDG /[^{68}Ga]Ga-PSMA	S_1	ADAM	Binary Cross-Entropy	0.0001	30	(3,3)	70	L2	491.08±1.35	1.21±0.06
		$S_{1/2}$								502.92±15.13	1.34±0.22
		$S_{1/4}$								509.53±5.31	1.56±0.47
		$S_{1/8}$								493.71±3.33	1.31±0.11
		$S_{1/16}$								496.73±3.08	1.26±0.06
4-class	[^{18}F]F-FDG	S_1	ADAM	Categorical Cross-entropy	0.001/0.0001 /0.00001	20/30	(3,3)/(5,5)	100	-	454.56±60.63	0.59±0.07
		$S_{1/2}$								456.64±64.70	0.61±0.09
		$S_{1/4}$								454.18±59.25	0.60±0.07
		$S_{1/8}$								452.56±57.66	0.64±0.28
		$S_{1/16}$								460.45±63.96	0.66±0.23
	[^{68}Ga]Ga-PSMA	S_1	ADAM	Categorical Cross-entropy	0.001/0.0001 /0.00001	20/30	(3,3)/(5,5)	100	-	423.18±56.33	0.56±0.07
		$S_{1/2}$								424.72±56.40	0.58±0.09
		$S_{1/4}$								430.11±57.17	0.59±0.09
		$S_{1/8}$								432.57±59.96	0.63±0.16
		$S_{1/16}$								426.88±57.12	0.59±0.08

Table 4. Binary classification results.*

Sub-series	Accuracy	Sensitivity	F1-score	AUC	Training time (s)	Prediction time (s)
$\boldsymbol{S_1}$	0.996±0.003	0.996±0.002	0.996±0.003	0.996±0.003	491.08±1.35	1.21±0.06
$\boldsymbol{S_{1/2}}$	0.994±0.002	0.995±0.002	0.994±0.002	0.994±0.002	502.92±15.13	1.34±0.22
$\boldsymbol{S_{1/4}}$	0.994±0.003	0.994±0.003	0.994±0.003	0.994±0.004	509.53±5.31	1.56±0.47
$\boldsymbol{S_{1/8}}$	0.990±0.002	0.992±0.002	0.992±0.002	0.992±0.002	493.71±3.33	1.31±0.11
$\boldsymbol{S_{1/16}}$	0.989±0.004	0.989±0.004	0.989±0.004	0.989±0.004	496.73±3.08	1.26±0.06

AUC: Area under the curve

*: In order to differentiate values, the accuracy, sensitivity, f1-score, and AUC results presented in this table are reported with more fraction digits than the rest of the results presented in this study

Table 5. Results for 4-class classification, and the corresponding hyperparameters settings. Results are reported as average over all classes.

Radiopharmaceutical	Sub-series	Kernel Size	Learning Rate	Batch Size	Accuracy	Sensitivity	F1-score	AUC	Training Time (s)	Prediction Time (s)
[^{18}F]F-FDG	S_1	(3,3)	0.001	20	0.999	0.998	0.999	1.0	395.84	0.80
	$S_{1/2}$	(5,5)	0.0001	20	0.996	0.991	0.998	1.0	524.07	0.61
	$S_{1/4}$	(3,3)	0.0001	20	0.998	0.996	0.997	1.0	399.49	0.55
	$S_{1/8}$	(5,5)	0.001	30	0.993	0.986	0.995	1.0	494.99	0.61
	$S_{1/16}$	(3,3)	0.0001	20	0.983	0.966	0.993	1.0	399.02	0.55
[^{68}Ga]Ga-PSMA	S_1	(3,3)	0.0001	20	0.992	0.998	0.998	1.0	371.22	0.50
	$S_{1/2}$	(5,5)	0.0001	20	0.996	0.991	0.991	1.0	496.90	0.58
	$S_{1/4}$	(3,3)	0.0001	20	0.993	0.996	0.996	1.0	378.72	0.52
	$S_{1/8}$	(5,5)	0.0001	30	0.988	0.986	0.986	0.996	480.10	0.61
	$S_{1/16}$	(5,5)	0.0001	30	0.978	0.966	0.966	0.995	469.27	0.60

Table 6. Hyperparameter setting of collapsed models

Radiopharmaceutical	Sub-series	Kernel Size	Learning Rate	Batch Size
[^{68}Ga]Ga-PSMA	S_1	(5,5)	0.001	30
	$S_{1/2}$	(3,3)	0.001	30
		(5,5)	0.001	30
	$S_{1/4}$	(3,3)	0.001	20
		(5,5)	0.001	30
	$S_{1/8}$	(5,5)	0.001	30
	$S_{1/16}$	(3,3)	0.001	30
		(5,5)	0.001	30
[^{18}F]F-FDG	S_1	(5,5)	0.001	30
		(5,5)	0.001	20
	$S_{1/2}$	(5,5)	0.001	20
		(5,5)	0.001	30
		(5,5)	0.001	20
	$S_{1/4}$	(5,5)	0.001	30
	$S_{1/8}$	(5,5)	0.001	20
		(5,5)	0.001	30
	$S_{1/16}$	(5,5)	0.001	30

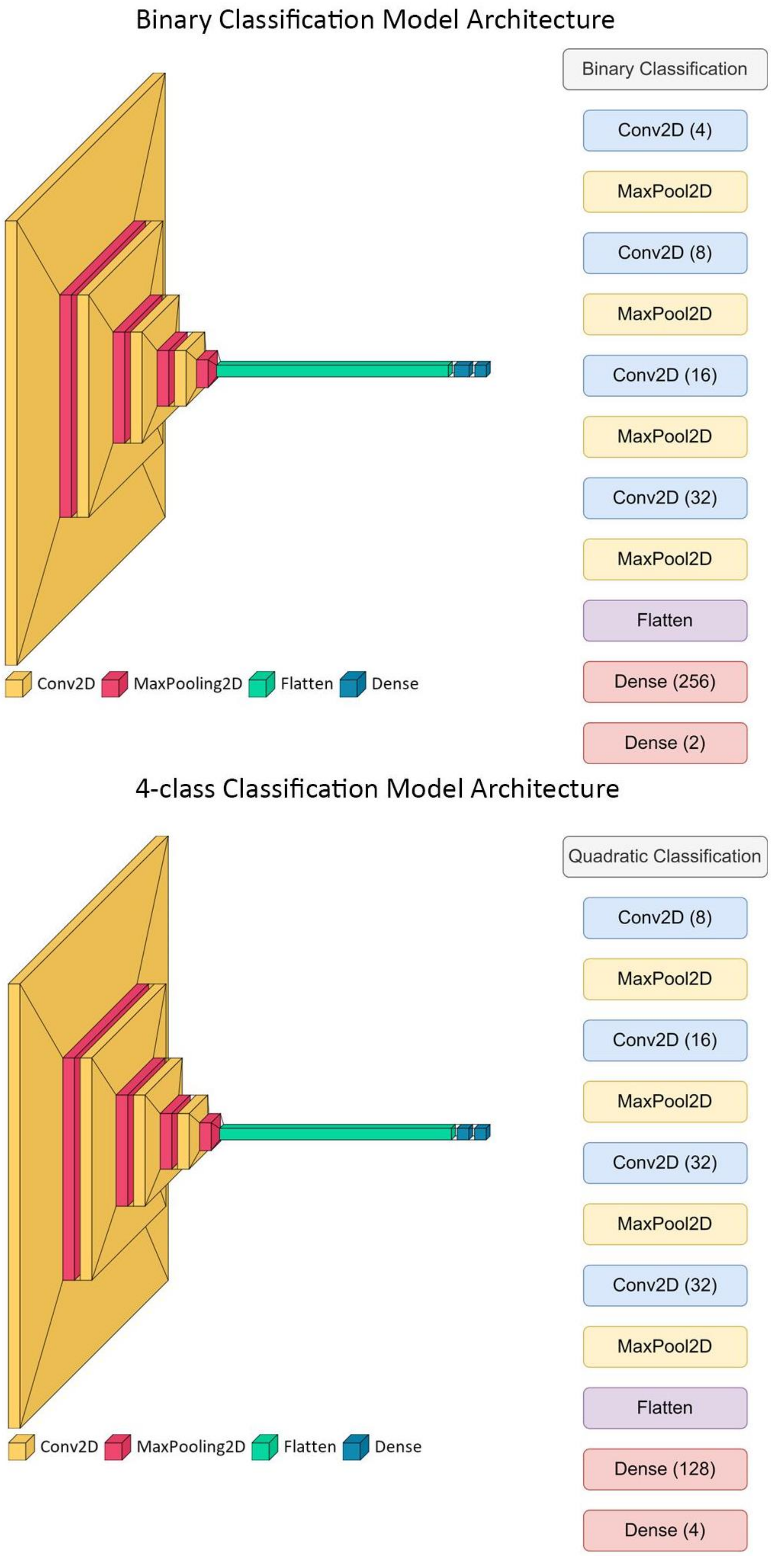

Fig 1. Model architectures used for binary and 4-class classification tasks. Four blocks of convolution and maximum pooling layers are followed by a flattening layer. Finally, two fully-connected (dense) layers are added to the end of the models

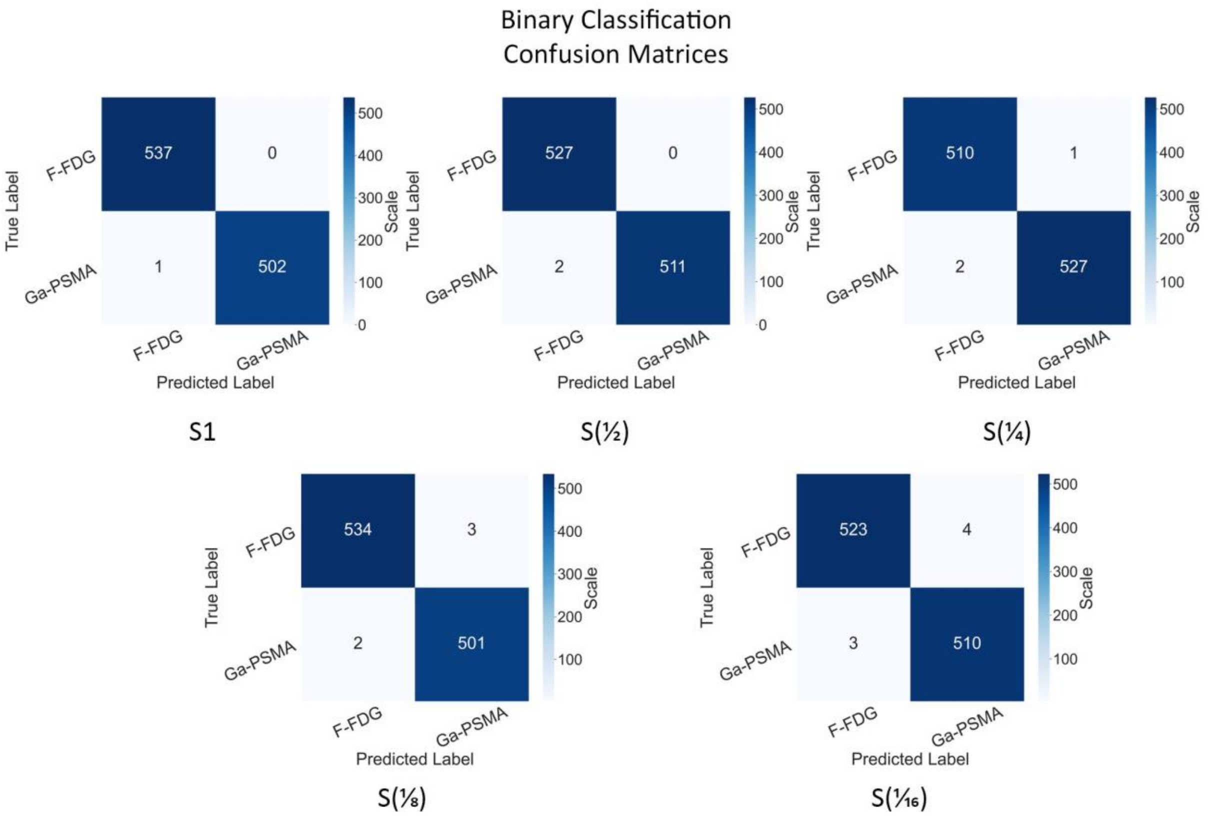

Fig 2. Confusion matrices of folds with the best results in each sub-series of binary classification task. S_1, $S_{1/2}$, $S_{1/4}$, $S_{1/8}$ and $S_{1/16}$ denote different scan duration fractions at which binary classification task was explored

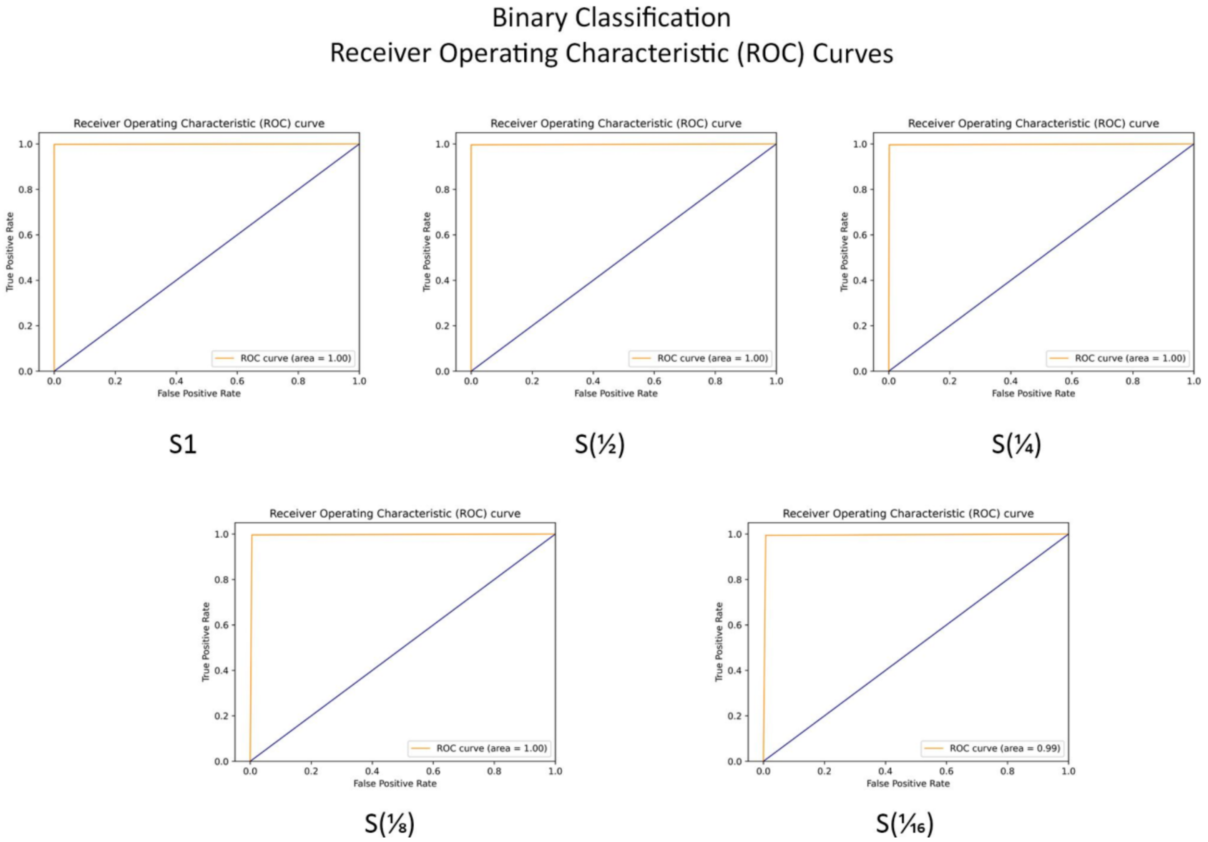

Fig 3. Receiver Operating Characteristic Curves (ROC) of folds with the best results in each sub-series of binary classification task

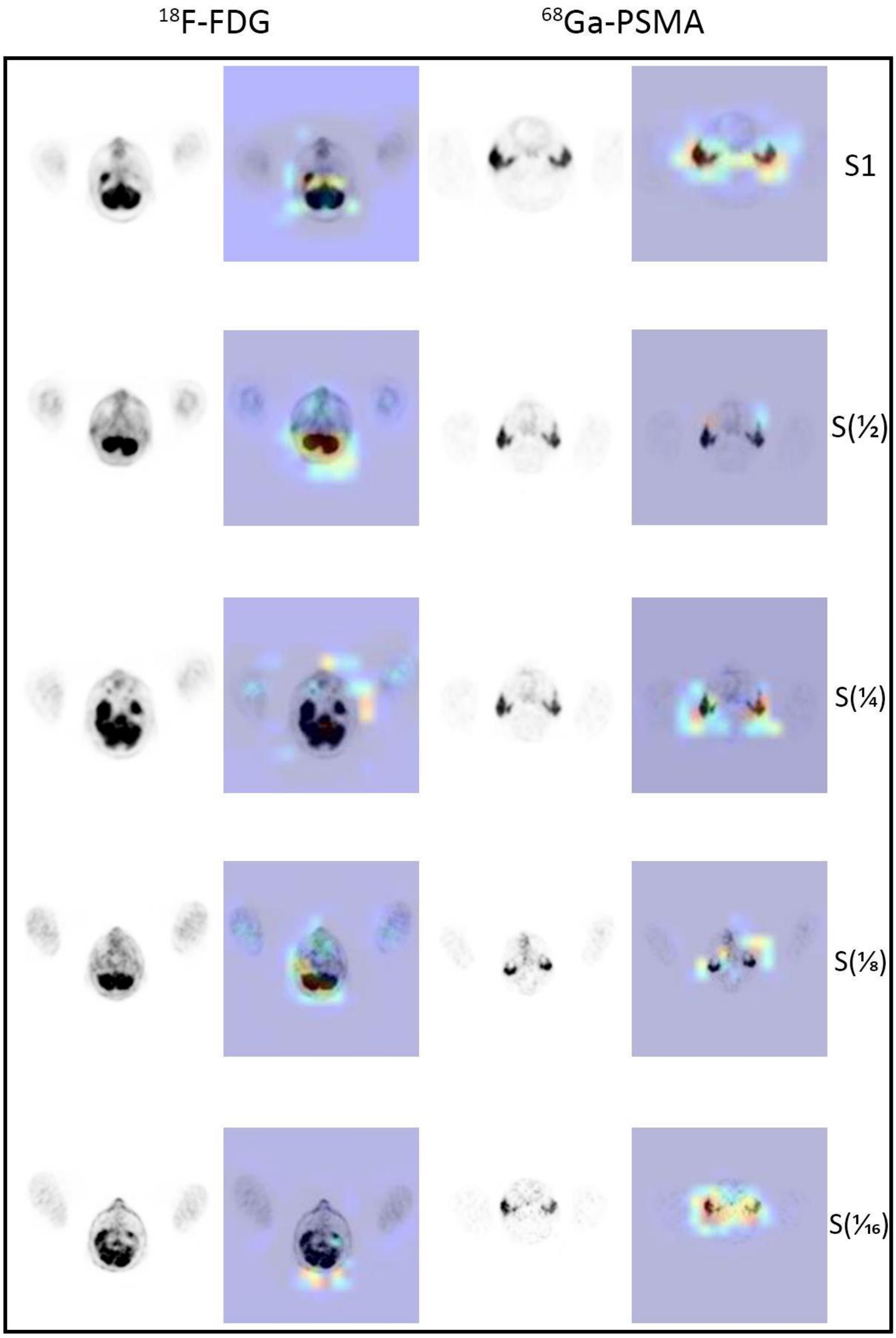

Fig 4. GradCam illustration for binary classification models in different sub-series and radiopharmaceuticals

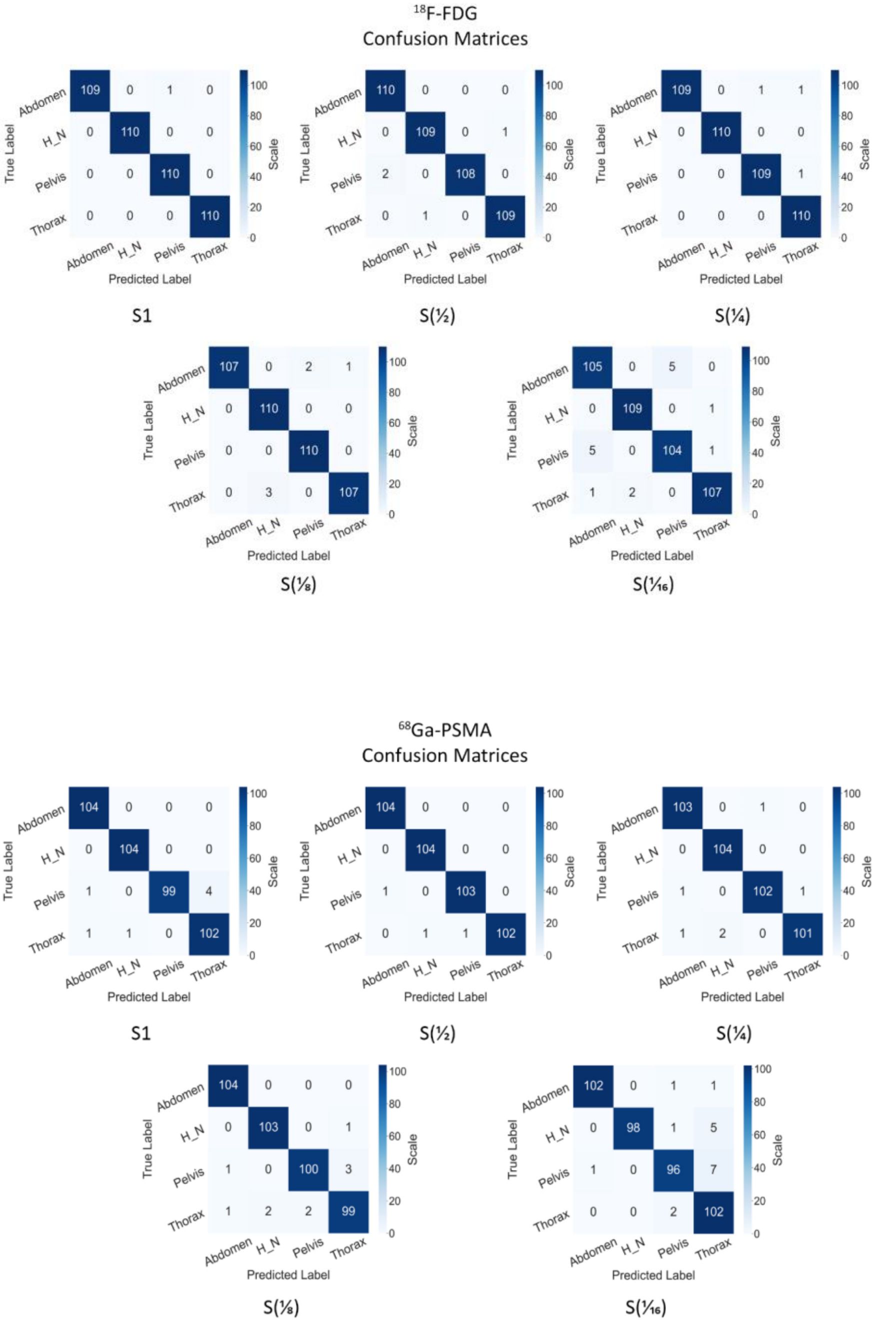

Fig 5. *(Above):* Confusion Matrices of models with the best results in [^{18}F]FDG radiopharmaceutical. S_1, $S_{1/2}$, $S_{1/4}$, $S_{1/8}$ and $S_{1/16}$ denote different scan duration fractions at which [^{18}F]FDG 4-class classification task was explored *(Below):* Confusion Matrices of models with the best results in [^{68}Ga]Ga-PSMA radiopharmaceutical. S_1, $S_{1/2}$, $S_{1/4}$, $S_{1/8}$ and $S_{1/16}$ denote different scan duration fractions at which [^{68}Ga]Ga-PSMA 4-class classification task was explored. As expected, $S_{1/16}$ in both [^{18}F]FDG and [^{68}Ga]Ga-PSMA had the higher incidence rate of misclassification. It is also notable that sum of all misclassification cases for all five levels was higher in [^{68}Ga]Ga-PSMA than [^{18}F]FDG which indicates the fact that [^{18}F]FDG is a general-purpose radiopharmaceutical compared with [^{68}Ga]Ga-PSMA

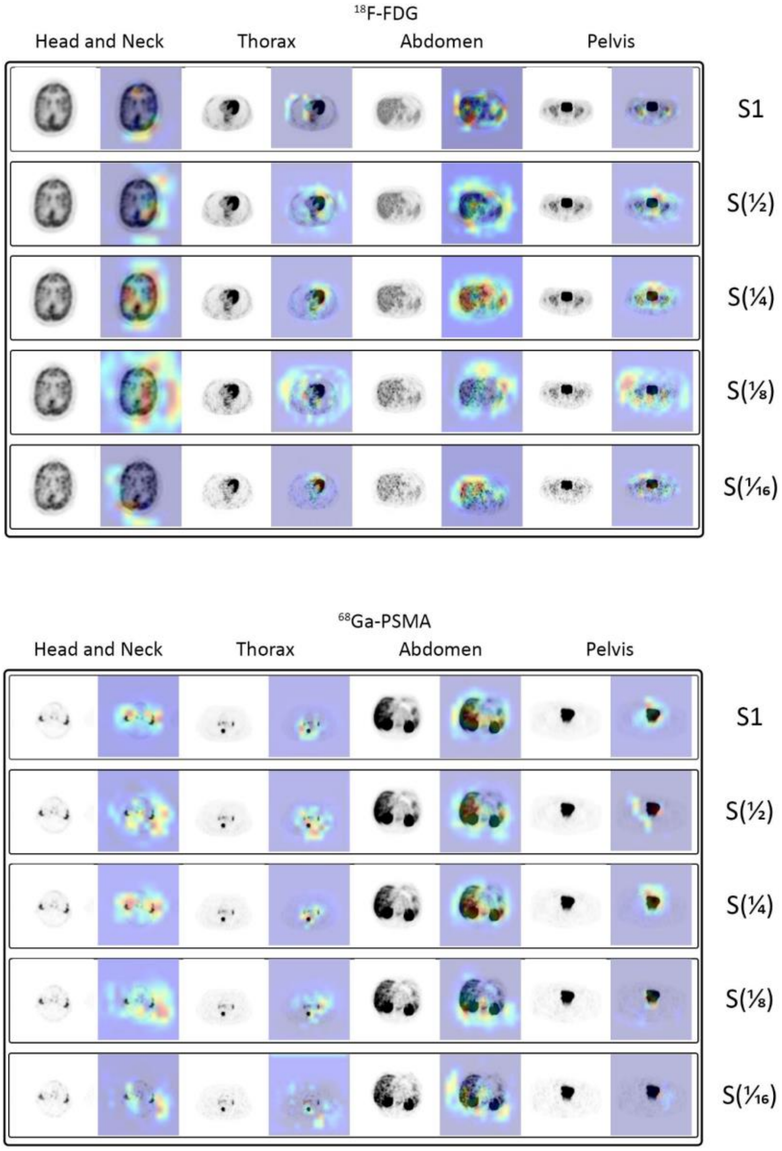

Fig 6. *(Above)*: GradCam illustrations for 4-class Classification of [^{18}F]FDG Radiopharmaceutical. The CNN successfully classifies images based on the structures specific to a particular region (e.g., Gray and White matter of the brain or mediastinum of the thoracic region). It is noticeable that with decreasing PSNR (from S_1 to $S_{1/16}$) the pinpointing accuracy of structures decreases in some regions (abdomen and pelvis)
(Below): GradCam illustrations for 4-class Classification of [^{68}Ga]Ga-PSMA Radiopharmaceutical. Just like [^{18}F]FDG, the CNN again successfully classifies images based on the structures specific to that special region (e.g., liver and spleen in abdominal region and bladder in Pelvic area). Since [^{68}Ga]Ga-PSMA is a specific radiopharmaceutical used in prostate cancer imaging; thus, its uptake in other body regions is not as prominent as [^{18}F]FDG, a multi-purpose radiopharmaceutical

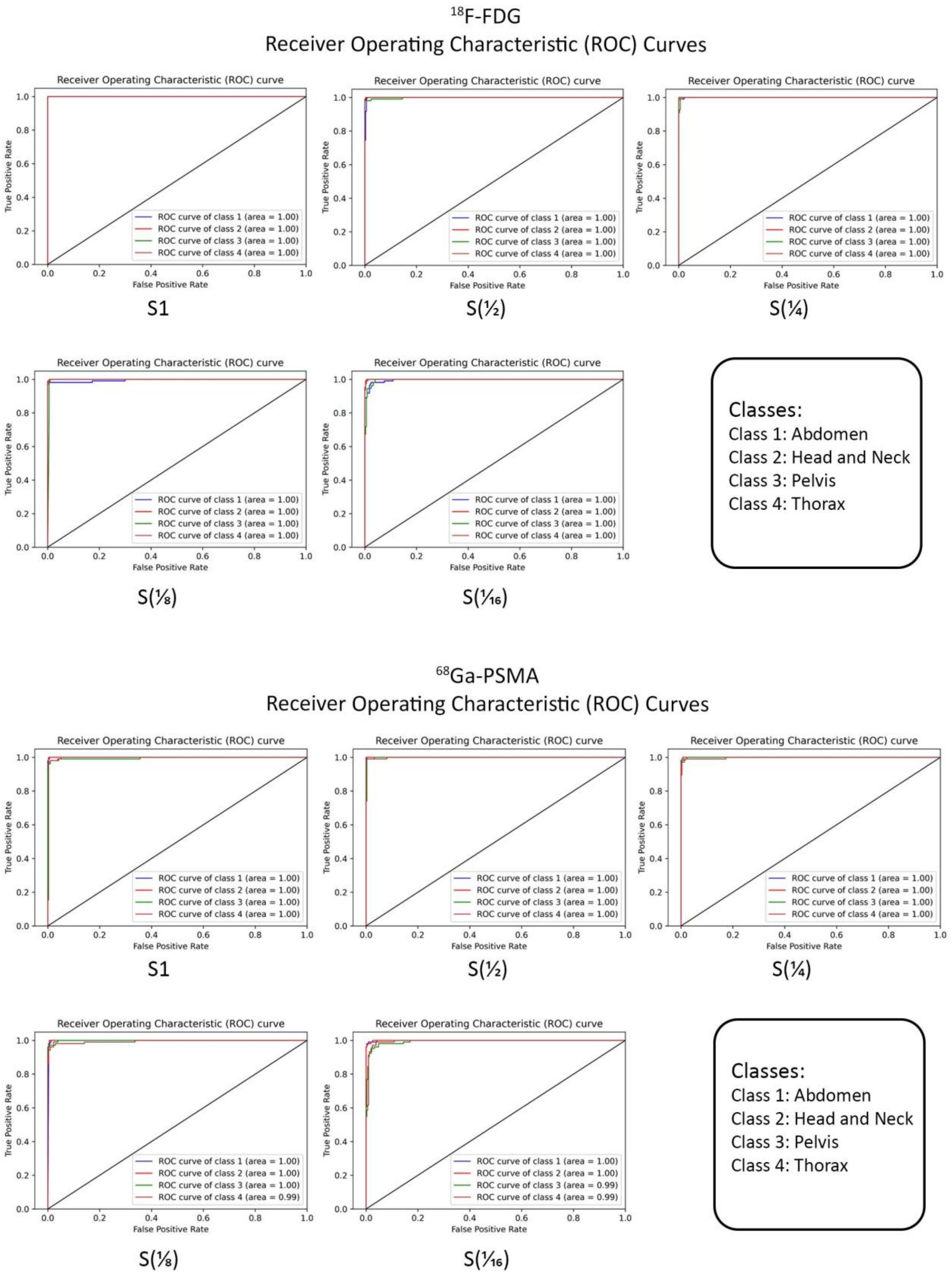

Fig 7. *(Above)*: Receiver Operating Characteristic curves (ROC) of models with the best results in [^{18}F]FDG radiopharmaceutical. *(Below)*: Receiver Operating Characteristic Curves (ROC) of models with the best results in [^{68}Ga]Ga-PSMA radiopharmaceutical

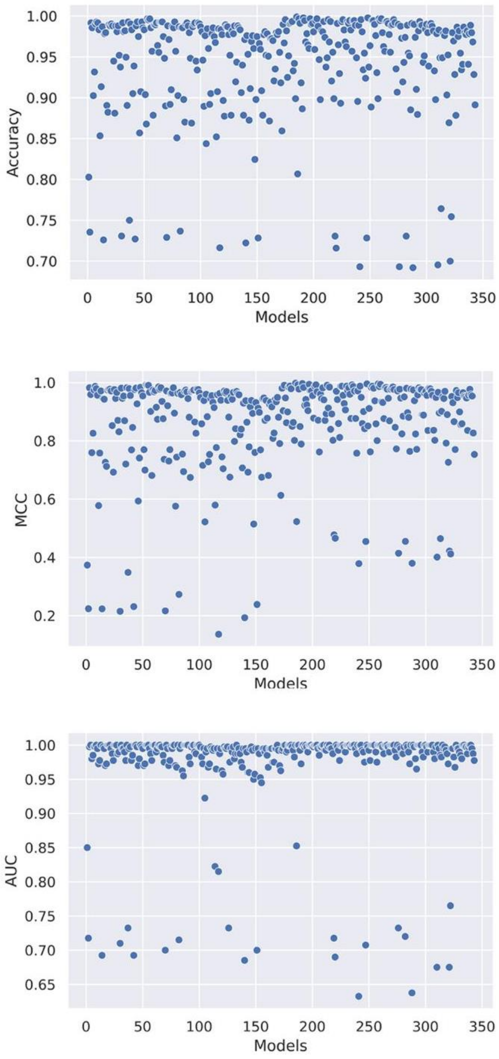

Fig 8. Scatter plots of 4-class classification models (x-axis) and classification metrics