



King's Research Portal

DOI:

[10.1109/TRPMS.2022.3231702](https://doi.org/10.1109/TRPMS.2022.3231702)

Document Version

Peer reviewed version

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Hopson, J. B., Neji, R., Dunn, J. T., McGinnity, C. J., Flaus, A., Reader, A., & Hammers, A. (2022). Pre-training via Transfer Learning and Pretext Learning a Convolutional Neural Network for Automated Assessments of Clinical PET Image Quality. *Transactions on Radiation and Plasma Medical Sciences*. Advance online publication. <https://doi.org/10.1109/TRPMS.2022.3231702>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Pre-training via Transfer Learning and Pretext Learning a Convolutional Neural Network for Automated Assessments of Clinical PET Image Quality

Jessica B. Hopson, Radhouene Neji, Joel T. Dunn, Colm J. McGinnity, Anthime Flaus, Andrew J. Reader*, and Alexander Hammers*

Abstract— Positron emission tomography (PET) using a fraction of the usual injected dose would reduce the amount of radioligand needed, as well as the radiation dose to patients and staff, but would compromise reconstructed image quality. For performing the same clinical tasks with such images, a clinical (rather than numerical) image quality assessment is essential. This process can be automated with convolutional neural networks (CNNs). However, the scarcity of clinical quality readings is a challenge. We hypothesise that exploiting easily available quantitative information in pretext learning tasks or using established pre-trained networks could improve CNN performance for predicting clinical assessments with limited data. CNNs were pre-trained to predict injected dose from image patches extracted from eight real patient datasets, reconstructed using between 0.5%-100% of the available data. Transfer learning with seven different patients was used to predict three clinically-scored quality metrics ranging from 0-3: global quality rating, pattern recognition and diagnostic confidence. This was compared to pre-training via a VGG16 network at varying pre-training levels. Pre-training improved test performance for this task: the mean absolute error of 0.53 (compared to 0.87 without pre-training), was within clinical scoring uncertainty. Future work may include using the CNN for novel reconstruction methods performance assessment.

Index Terms— Convolutional neural networks, Deep learning, Image quality, Image reconstruction, Transfer learning

This work was funded by the EPSRC Centre for Doctoral Training in Smart Medical Imaging [#EP/S032789/1] and Siemens Healthcare, Erlangen, Germany. This work was supported by the Wellcome/EPSRC Centre for Medical Engineering at King’s College London (WT 203148/Z/16/Z) and the Department of Health via the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy’s & St Thomas’ NHS Foundation Trust in partnership with King’s College London and King’s College Hospital NHS Foundation Trust. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the institution’s Research and Development department, the North East – York research ethics committee (reference 15/NE/0203), and the radiation protection committee (ARSAC).

J. B. Hopson and A. J. Reader are with the Department of Biomedical Engineering, King’s College London.

J. T. Dunn, C. J. McGinnity, A. Flaus and A. Hammers are with the King’s College London & Guy’s and St Thomas’ PET Centre, King’s College London.

R. Neji is with Siemens Healthcare Limited.

* Denotes equal contribution to this work. Corresponding author: Jessica Hopson, email: jessica.hopson@kcl.ac.uk

I. INTRODUCTION

POSITRON emission tomography (PET) is an important modality in the management of several brain diseases, notably memory problems, dementia, epilepsy [1], and brain tumours [2]. Simultaneous positron emission tomography – magnetic resonance imaging (PET-MR) has potential to become an important tool in the diagnosis of memory clinic patients [3], by simultaneously providing both functional and structural information of the brain [4]. However, a drawback to PET imaging is the need for administration of the radioligand: the more radioligand injected, the higher the associated radiation dose, and radioligand production to Good Manufacturing Practice standards is very expensive [5]. Both can be targeted by reducing the injected dose [6], but at the expense of the reconstructed image quality, caused by increased noise and a loss of resolution [7]. Traditionally, image quality assessment has relied on human raters, but this is not scalable and impractical for large-scale investigations or repeating assessments while optimising reconstruction algorithms. Thus, it is beneficial to determine automatically whether a reconstructed image is of satisfactory diagnostic quality. One approach is to use convolutional neural networks (CNNs), as they are data-driven [8] and versatile, being used for medical imaging tasks including image segmentation, lesion detection and image de-noising [9]–[14]. CNNs have also shown to outperform radiomic analysis in certain tasks [15].

A drawback of using CNNs is that generally a large dataset is required for training [16]. However, this is challenging in medical imaging, as clinically-annotated data are scarce and not scalable [17]. Associated acquisition and processing costs contribute to the paucity of clinician-labelled data. Pre-training (an architecture trained on a separate image database) and transfer learning (taking features learned on one problem and using them for a new problem) could be exploited to overcome this lack of training data [18] [19]. Transfer learning can be approached with either fine tuning or fixed representations [20]. Fine tuning involves unfreezing some or all of the network parameters in order to update the weights with a very low learning rate and a new dataset [18].

Pre-training with natural image databases has been used in the medical imaging field. ImageNet [21] is a dataset of more than 15 million high-resolution natural images annotated for 22,000 classes. A subset of ImageNet used for deep learning tasks consists of 1.2 million images categorised into 1000 classes [22]. Pre-trained networks via ImageNet have been used for the binary classification of papillary thyroid carcinoma [23], where exploiting a VGG16 [24] network achieved a 97.66% accuracy when distinguishing between carcinomas and benign thyroid nodules in two-dimensional cytological image patches. Dunmon *et al.* [25] showed that CNNs could be pre-trained on ImageNet to classify chest radiographs into “normal” or “abnormal” with an area under the receiver operating characteristic curve (AUC) of 0.96. This was built upon by Tang *et al.* [26], who investigated a greater number of established architectures, including VGGNets [24] and Inception-v3 networks [27], and compared pre-trained networks against training from scratch. Their work showed that when using a moderately sized dataset (8500 two-dimensional chest radiographs), pre-training on ImageNet outperformed training the model from scratch. It has also been shown that a transfer-learned VGG16 network pre-trained via ImageNet outperformed an Inception-v3 network for the task of classifying into ‘Normal’ or ‘Pneumonia present’, and matching the complexity of the model to the dataset size along with data augmentation, further improved CNN accuracy [28].

However, there is debate on the usefulness of transfer learning from a natural image domain to a medical imaging domain [29], [30]. Recent work on transfer learning questions the validity of using natural images for pre-training in medical imaging tasks due to the high dissimilarity between the domains [31]. Heker *et al.* [32] also showed transfer learning within domain (from one medical image dataset to another) outperformed inter-domain pre-training (from a natural image database to a medical image database) for segmentation and classification tasks.

Pre-training can be used to overcome extensive training data needs. Mustafa *et al.* [19] showed network performance could be improved by increasing size of both the model architecture, and the natural image pre-training dataset. ImageNet-21k (the full ImageNet dataset classified into 21000 classes) obtained superior performance over ImageNet (the subset of ImageNet classified into 1000 classes). However, the full dataset is more computationally expensive and labels are not mutually exclusive: a single image may have multiple labels, making classification more difficult [33].

As pre-trained networks act as feature extractors, layers need to be frozen to prevent the network from retraining too extensively, potentially negating any pre-training benefit. As shown in [34], inclusion of multiple layers in the fine tuning process can improve model performance. In an Alzheimer’s disease diagnosis classification task, all convolutional layers of pre-trained networks were frozen, except for the last fully connected layer which was trained on the new dataset [35]. Kieffer *et al.* [36] retrained the last convolutional block and fully connected layer of a VGG16 network for their dissimilar domains.

Whilst there have been many image quality assessment investigations in the computer vision field [37]–[39], similar experiments within the medical imaging field have emphasised binary classifications. One such investigation by Sujit *et al.* [40] used a CNN to predict image quality of brain MRIs compared to visual evaluation by two experts by scoring the images as ‘0 (acceptable)’ or ‘1 (unacceptable)’, achieving an accuracy of 84%. Oksuz *et al.* [41] used cardiac MR images to detect motion-related artefacts, classifying the image as “good” or “poor”. Similarly, Ma *et al.* [42] used a CNN to carry out both a binary classification (‘0 – non-diagnostic’ or ‘1 – diagnostic’) and three-class classification (‘0 – poor/non-diagnostic’, ‘1 – diagnostic’ or ‘2 – excellent’) of MR images, achieving an accuracy of 84% and 65%, respectively. In the field of retinopathy, a similar approach was taken by Coyner *et al.* [43] to determine whether clinically-scored 2D images were of “acceptable”, “possibly acceptable” or “not acceptable” quality, and they found a strong correlation (Spearman’s rank correlation coefficient of 0.84 – 0.92) between the CNN and the expert scoring. Another retinopathy study showed that pre-training via ImageNet outperformed training from scratch [44].

Building on previous work [45], the aim of this work was to assess the influence of transfer learning as a function of the number of fixed parameters during training, as well as comparing different pre-trained architectures for the prediction of clinical PET image quality assessments. Most of the literature relates to CT and MR images, and there is limited research on PET image quality assessment. Varying image qualities were generated by reconstructing images with different simulated injected doses such that all other factors that may affect image quality, including reconstruction algorithm and system design, were fixed, and only the dose level determined the differences in image quality. The performance of an established VGG16 network [24] pre-trained on the ImageNet-1K database [21] was assessed and compared to a pre-text learning task (using an easily extracted label, the injected dose) as a form of pre-training to predict three clinically scored metrics. These metrics are: global quality rating (GQR), which evaluates the aesthetic component of the reconstruction, and takes into consideration qualities such as noise level and resolution; pattern recognition (PR), which determines whether the reconstruction allows for the clinician to determine the presence of any pathological patterns suggesting a particular diagnosis; and diagnostic confidence (DC), which concerns the certainty with which the clinician can use the reconstruction to observe these patterns and make a diagnosis. As low count reconstruction causes the degradation of PET image quality, these metrics were designed by the authors to encompass the necessary considerations involved in using these low count clinical PET images for making a confident and accurate diagnosis. These three metrics are scored as: 0 = “unacceptable”, 1 = “poor but usable”, 2 = “acceptable” and 3 = “good/excellent”, with 0.5 ratings accepted. The aim was to reduce the need for clinically-assessed training data by the use of pre-trained networks and transfer learning.

II. METHODS AND MATERIALS

A. Real Patient Data

Data from PET-MR scans of 21 memory clinic patients with suspected dementia were used. Each patient underwent an [^{18}F]FDG-PET and a simultaneous 3D T1-magnetisation-prepared gradient-echo (MP-RAGE) MRI scan on a 3T PET-MR scanner (Biograph mMR, Siemens Healthcare, Erlangen, Germany). Siemens e7 tools was used for image reconstruction, using the clinical standard of the ordered subset expectation maximisation (OSEM) algorithm with 2 iterations and 21 subsets to reconstruct 3D images of $344 \times 344 \times 127$ voxels. It is possible for the list-mode PET data to be resampled [45], [46], and such methods can be used to simulate seven lower injected doses (Fig. 1). To simulate the lower injected doses, the prompts and random coincidences in the corresponding list-mode PET data for the 100% count reconstruction were randomly sampled. Subsequently, the corresponding emission, normalisation, randoms and sinograms were re-calculated along with the attenuation sinograms using Siemens e7 tools. The list-mode PET data from all datasets were resampled [46], simulating seven lower injected doses (Fig. 1). Images were normalised between 0-1 to reduce standard deviations, suppressing the impact of outliers, then blinded and randomised prior to clinician scoring. The simulated dose levels were grouped into “low quality” (0.5% and 1% of counts), “medium quality” (5% and 10% of counts), and “high quality” (25%, 50% and 100% of counts) categories to decrease clinical assessment times. In total, ten patients at three different count levels (one per category) were assessed by an experienced clinician. For calibration, nine images were scored twice by the same clinician in different scoring sessions (Fig. 2). Another experienced clinician also scored the same images. Pearson’s correlation coefficients of 0.91, 0.91, 0.90 for GQR, PR and DC, respectively, were found between the clinicians’ scores. It was also found that 91% of the scores were within 0.5 of that given by clinician 1, which is within the calibration range (Fig. 2) showing the robustness of the scores given by each of the clinicians. After consensus readings, only one score was greater than 0.5 from clinician one’s original scoring, thus clinician one’s scores were used for this study. Fig. 3 shows the 20-point discrete colour scale used during the scoring process which was optimised for reading brain PET images at the Guy’s and St Thomas’ PET Centre. This colour scale allows for the specific pathological patterns to be more easily highlighted, such that with a lower quality image, a diagnosis could still be made, helping to determine pattern recognition and diagnostic confidence. The image used in Fig. 3 corresponds to the same example in the top right subplot in Fig. 1. Table I shows a Pearson’s pairwise correlation coefficient for each of the metric scores given by both clinicians.

B. Extraction and Thresholding of Patches

For each patient, 1000 patches of size 80×80 pixels were extracted at random from transverse, sagittal and coronal planes to cover the whole 3D volume, whilst not being too computationally exhaustive. Patches instead of whole images

were also used as a form of data augmentation, as has been shown to improve model performance [28]. To eliminate background-only patches, a thresholding algorithm was applied to each individual patch. If the average pixel value of the patch was below one-eighth of the average pixel value of the whole image volume, then the patch was rejected as background, ensuring that only patches with meaningful brain information were used, as is used when defining brain masks in FDG PET imaging [47]. Of these thresholded patches, 100 each from the transverse, sagittal and coronal planes, were randomly sampled, such that 300 patches were used for each reconstructed image volume. This number of patches was chosen as it is proportional to the brain:background ratio in the whole image volume.

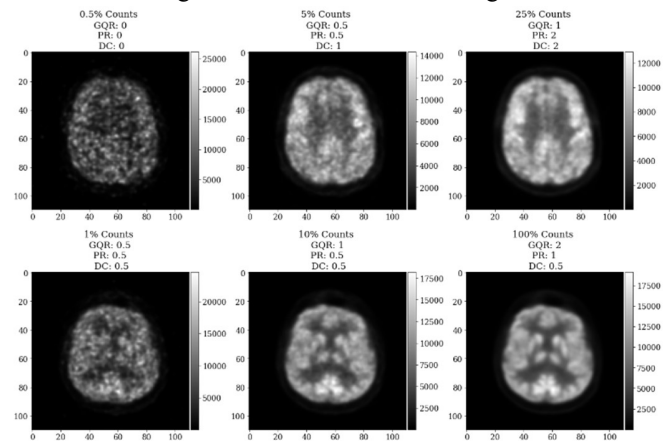


Fig. 1. Cropped example transverse slices from “low”, “medium” and “high” quality reconstructions for two independent datasets.

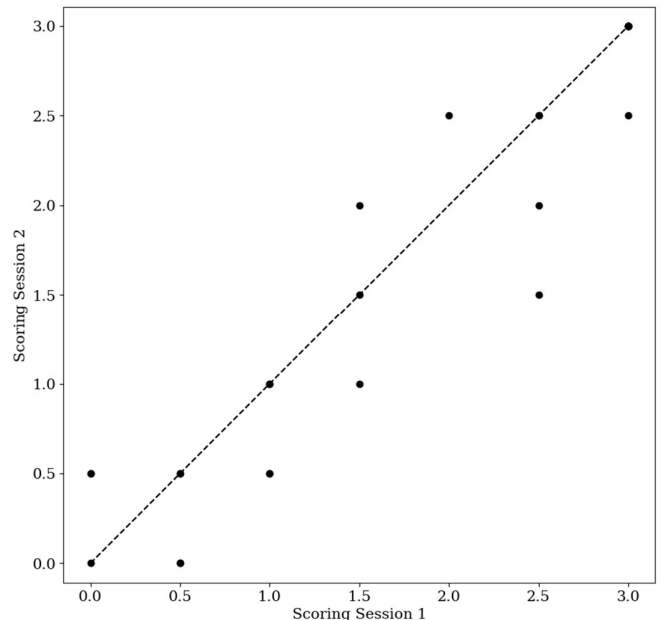


Fig. 2. Inter-session scoring difference for all clinical metrics. The dotted line is the identity line for scoring session 1.

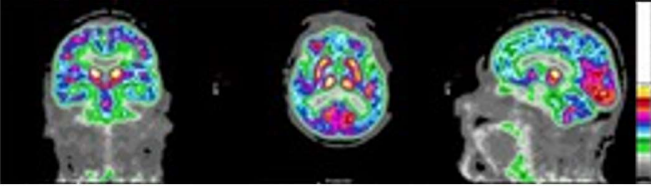


Fig. 3. A clinically assessed image with a 20-point colour scale with a lower global quality rating (1), but higher pattern recognition (2) and diagnostic confidence (2) scores (corresponding to the top-right subplot in Fig. 1). The colour-scale is the technique by which the clinician determines the pattern recognition and diagnostic confidence ratings.

TABLE I
CORRELATION BETWEEN CLINICAL QUALITY METRICS

Metric Comparison	Pearson's Correlation Coefficient	
	Clinician 1	Clinician 2
<i>GQR vs. PR</i>	0.90	0.93
<i>GQR vs. DC</i>	0.84	0.93
<i>PR vs. DC</i>	0.96	0.99

C. Model Architecture

A VGG16 architecture [23] with random weight initialisation (Section II.D) or ImageNet [20] weight initialisation (Sections II.D and II.E) acted as a feature extractor. In addition to the VGG16 backbone being used for other medical imaging tasks [28], [36] and its block structure providing a demonstrable method of pre-training levels, the VGG16 backbone was chosen as it is representative of similar CNN architectures (Fig. 4). A total of 26 different established backbones were trained on the same dataset, with the minimum validation loss monitored. The eleven models with the lowest validation losses, were trained another two times, and were then tested on the same three test patients (Fig. 4). There are five convolutional blocks consisting of convolution layers and a max pooling layer (Fig. 5). The three original fully connected layers of the VGG16 architecture were replaced with a single fully connected layer with three outputs relating to each of the clinical metrics (GQR, PR and DC) with ReLU activation (Fig. 5) in Section II.D. This is because the original VGG16 network is used for classification of the ImageNet database into 1000 different classes. However, this is a regression task, with the output a prediction of three clinical quality scores. Therefore, the original projection head was replaced by another fully connected layer to achieve this task. In Section II.E, the replacement fully connected layer had a single output corresponding to the simulated dose. The same seven patients were used in the transfer learning dataset, reconstructed at three simulated dose levels each. Greyscale input patches were replicated along one axis to obtain three channels as VGG16 inputs are RGB images.

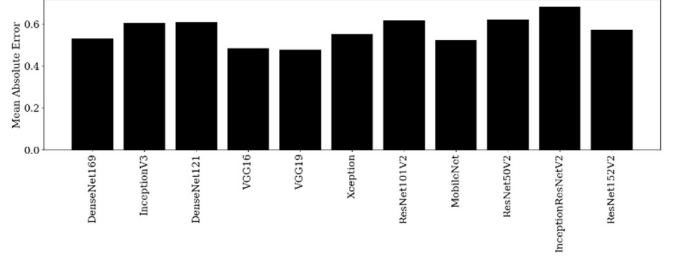


Fig. 4. Bar chart of the mean absolute error based on the same three test patients for 11 established CNN backbones for three training runs. A VGG16 backbone is shown to be representative of other established CNN backbones.

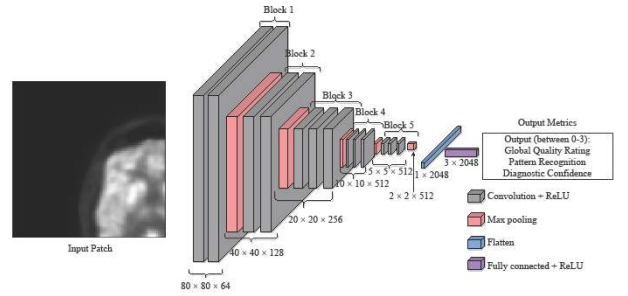


Fig. 5. Transfer learned VGG16 architecture. A patch of size 80×80 is used as the input with the three quality metrics as output. Each of the five convolutional blocks is labelled.

D. Transfer Learning: Pre-training a VGG16 Backbone

The VGG16 network (Fig.5) was initialised with random weights, with a learning rate of either 10^{-3} (default) or 10^{-4} as a baseline comparison model.

The VGG16 network was then also initialised using ImageNet weights, and trained by sequentially unfreezing weights from consecutive blocks starting from training zero convolutional blocks (only the fully connected layer) up to and including all five convolutional blocks. Each model was trained seven times and seven patients were used in the transfer learning dataset. This model was also tested as a function of the number of patients in the transfer learning dataset; a maximum of 6300 patches were used for transfer learning.

E. Pretext Learning: Injected Dose Inference

Dose inference was used as a pre-text learning task due to the positive correlation between the number of counts in the reconstruction and the clinician-assigned score (Fig. 6). Patient weight was used to standardise between patients. In total, eight patients (16,800 patches) were used for training this network, with another patient used for validation. Both pre-trained (via ImageNet and re-training of the last two convolutional blocks) and not pre-trained networks (via random weight initialisation and re-training either the last two or all five convolutional blocks) were compared. For additional comparison, the dose standardised by patient weight was used as input into a model consisting of a fully connected layer with the three clinical metrics as the output. All pipelines are shown in Fig. 7.

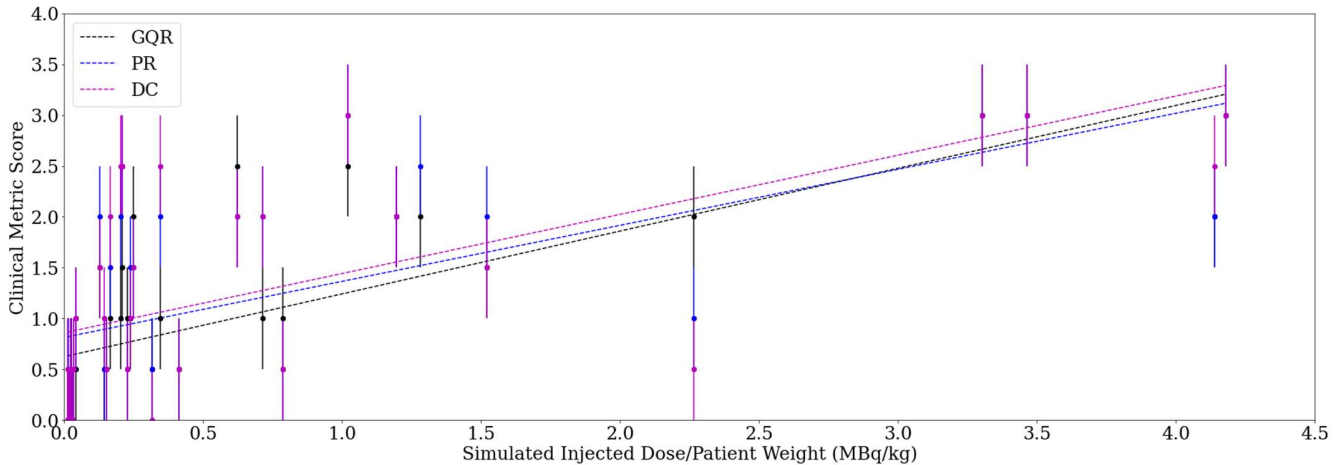


Fig. 6. Correlation between clinically-scored metrics and injected dose standardised by patient weight. Dotted line = line of best fit. Error bars = ± 0.5 , informed by Fig. 2.

F. Evaluation Metrics

Three evaluation metrics were used: mean absolute error (MAE) between the predicted and true clinical scores; percentage exact agreement (predicted score = true score); and percentage close agreement (predicted score = true score ± 0.5), in line with Fig. 2. All models were tested on the same three test patients at three dose levels each.

G. Implementation

Reconstructions via Siemens e7 tools were carried out in MATLAB (The Mathworks, Inc.). All models were trained using the Keras [18] application programming interface implemented in TensorFlow [48]. Training was accelerated using either an NVIDIA Tesla K40 12GB GPU or NVIDIA Quadro RTX6000 24GB GPU. All models used the Adam optimiser with a learning rate of 10^{-4} (unless otherwise stated), a mean-squared-error loss function and batch size of 10. Training occurred for 1000 epochs, and models were saved at the lowest validation loss which was used for analysis.

III. RESULTS

A. Influence of the Number of Retrained Blocks

Based on the three test patients, Fig. 8 shows a series of boxplots describing the evaluation metrics as a function of the number of pre-fully connected layer blocks when seven patients at three dose levels each were used in the transfer learning dataset. Re-training the last one or two blocks outperformed training the fully connected layer only, but there was no clear benefit to retraining the last 3-5 blocks, as percentage close agreement decreased. Re-training a higher number of convolutional blocks, generally increased the variation across separate training runs, without further decreases in the mean or median absolute error. There is a plateau in the MAE around 0.5, in line with Fig. 2.

B. Influence of the Number of Patients in the Transfer Learning Dataset

Fig. 9 shows a series of boxplots describing the same metrics as in Fig. 8. Based on Section III.A, the last two convolutional

blocks of the VGG16 backbone were re-trained. Generally, the MAE and the associated variation decreased with increasing numbers of independent training datasets, but plateaued around 0.5, again in line with Fig. 2. Using one patient in the transfer learning dataset was consistently outperformed across all evaluation metrics by more independent patients in the transfer learning. There was clear improvement in the percentage exact agreement with increasing numbers of independent training patients.

C. Dose Inference Performance as a Function of Pre-Training and Retraining

Table II shows the MAE and percentage exact agreement comparing the different models investigated for analysing the pretext dose inference task (before transfer learning to clinical metric prediction), assessing the validation patient reconstructed at seven different dose levels. The green boxes indicate the best value across the three clinical metrics. Pre-training VGG16 with ImageNet and retraining the last two blocks vastly outperformed the VGG16 network with random weight initialisation and retraining all five convolutional blocks. Table III also shows the impact of using only the easily accessible dose label as input into a single fully connected layer, for a learning rate of both 10^{-3} and 10^{-4} , with the output as the prediction of the three clinical metrics. The green boxes indicate the best value across the three clinical metrics; lighter green boxes indicate a draw.

D. Comparison of all Pipelines

Fig. 10 compares all pipelines; from left to right, the amount of pre-training increases, starting from no pre-training (Fig. 7a), to pre-training via ImageNet (Fig. 7b), to transfer learning from the pretext learning task with random weight initialisation (Fig. 7c), and subsequently with ImageNet weight initialisation (Fig. 7d). No pre-training with a learning rate of 10^{-3} was outperformed by all other models. Whilst reducing the learning rate improved MAE 0.87 to 0.58, including ImageNet weights and freezing the first three convolutional blocks further reduced MAE to 0.53. Adding another layer of pre-training via the pretext learning task with random weight initialisation, reduced

variation over the seven separate training runs. Initialising this model with ImageNet weights, slightly reduced the MAE compared to when random weight initialisation was used; however, MAE between all pre-trained models were comparable. Table IV compares the same models as in Fig. 10, but evaluated across individual clinical metrics for each evaluation metric and the Spearman’s rank correlation coefficient. Bright green boxes indicate the best value for each evaluation metric, whilst lighter green boxes indicate the second best performing network. The majority of the best values for each metric were achieved by a pre-trained model. Generally, GQR is predicted more accurately than PR and DC.

TABLE II
PRETEXT LEARNING TASK: DOSE INFERENCE EVALUATION

VGG16 Network	Pre-training Weights	Mean Absolute Error	% Exact Agreement
Retrain last 2 blocks	ImageNet	0.09	71
Retrain all 5 blocks	Random	0.27	29

IV. DISCUSSION

There is a strong correlation between the metrics (Table I), especially between PR and DC, but these metrics were designed to specifically incorporate what the clinician is observing in order to determine if a decreasing image quality may still lead to an image that is clinically useful for making a diagnosis. The VGG16 backbone was chosen as it representative of other established CNNs (Fig. 4). Testing the VGG16 backbone as a function of the number of trainable convolutional blocks shows the necessary level of pre-training required for a better performance on unseen patients. The expected trend in Fig. 8 is that of a “U-shape” for MAE and an “inverted U-shape” for the percentage agreement metrics. At either extreme (either 0 or all 5 of the convolutional blocks trained), there should be a decreased performance (higher MAE and lower percentage exact and close agreements). This is because when none of the convolutional blocks are trained, there remain high-level features specific to the ImageNet dataset and not to medical images, and similarly, when all parameters are trainable, this negates any pre-training benefit [49]. Unfreezing the final two convolutional blocks’ weights resulted in the best percentage exact and close agreements, suggesting a balance between the number of high level features and pre-training benefit, agreeing with the work of Yosinski *et al.* [49].

Fig. 8 informed the investigation into the number of patients in the transfer learning dataset where the last two convolutional blocks of the VGG16 backbone were re-trained. The expected trend was that with more patients in the transfer learning dataset, the better the network performance on unseen datasets as there are more examples from which the model can learn [50]. Fig. 9 indeed shows that increasing the number of patients in the training dataset improved the percentage exact agreement

and decreased the MAE, supporting the idea that a larger dataset can improve generalisability to unseen datasets. Dunnmon *et al.* [25] showed that increasing the number of 2D chest X-ray images from 2000 to 20,000 improved the average area under the receiver operating characteristic curve (AUC) from 0.84 to 0.95, but improvements plateaued afterwards, as using 200,000 images, only achieved a non-significant increase in AUC to 0.96. Results from the use of just one patient in the transfer learning dataset were consistently outperformed by using ≥ 2 patients, suggesting that use of one patient in the dataset was insufficient to generalise to unseen data. The percentage close agreement started to plateau at $\sim 70\%$ beyond ≥ 3 independent training patients, but the percentage exact agreement continually improved. This suggests that, whilst three patients are sufficient for predictions within 0.5 of the clinician-scores, a larger dataset (i.e. >7) may be required to achieve exact agreements between the true and predicted scores.

For both Fig. 8 and Fig. 9, the MAE plateaued at ~ 0.5 corresponding to the inter-session scoring uncertainty within the clinical assessments of the images (Fig. 2). A total of 27 images were re-assessed by the same clinician, with 44% of the readings scored at least 0.5 different to the original assessments. Thus, obtaining a MAE of approximately 0.5 may be the optimum achievable for this task.

Table II shows that using pre-training via ImageNet outperformed no pre-training for dose inference. Informed by Fig. 8, pre-training VGG16 with ImageNet weights and retraining the last two blocks was used as the pre-trained network. The pre-trained network vastly outperformed VGG16 with random weight initialisation and training all five convolutional blocks, confirming the benefit of using pre-training instead of training from scratch for this task. The MAE was substantially lower than in the clinical score prediction task, possibly because 56 individual images were available compared to 21 clinically-assessed images.

Comparing all pipelines, Fig. 10 and Table IV show that pre-training provided a benefit to the overall network performance on unseen test data. Fig. 10 shows the MAE for no pre-training was ~ 0.92 when the default learning rate of 10^{-3} was used, suggesting that the network did not generalise well to unseen data. By simply decreasing the learning rate to 10^{-4} , the MAE was substantially reduced to 0.58. Subsequently changing this model to include ImageNet weight initialisation and only re-training the last two convolutional blocks together with the new fully connected layer resulted in decreases in the MAE across all predicted clinical metrics and test patients, i.e. pre-training via ImageNet provided an improvement to overall model performance at test-time. Introducing the dose inference pretext learning task with no pre-training (all parameters trainable and random weight initialisation), reduced the variation between runs, suggesting the dose inference pretext learning task was the source of uncertainty reduction. Using pre-training via ImageNet weight initialisation, again reduced the MAE compared to the same model with random weight initialisation, indicating that by adding another level of pre-training, model performance could be further improved (if minimally). Table IV shows that using a pretext learning task had an inferior performance than transfer learning straight from ImageNet. This is perhaps due to dose inference being more difficult when the images are normalised, as the dose information contained in

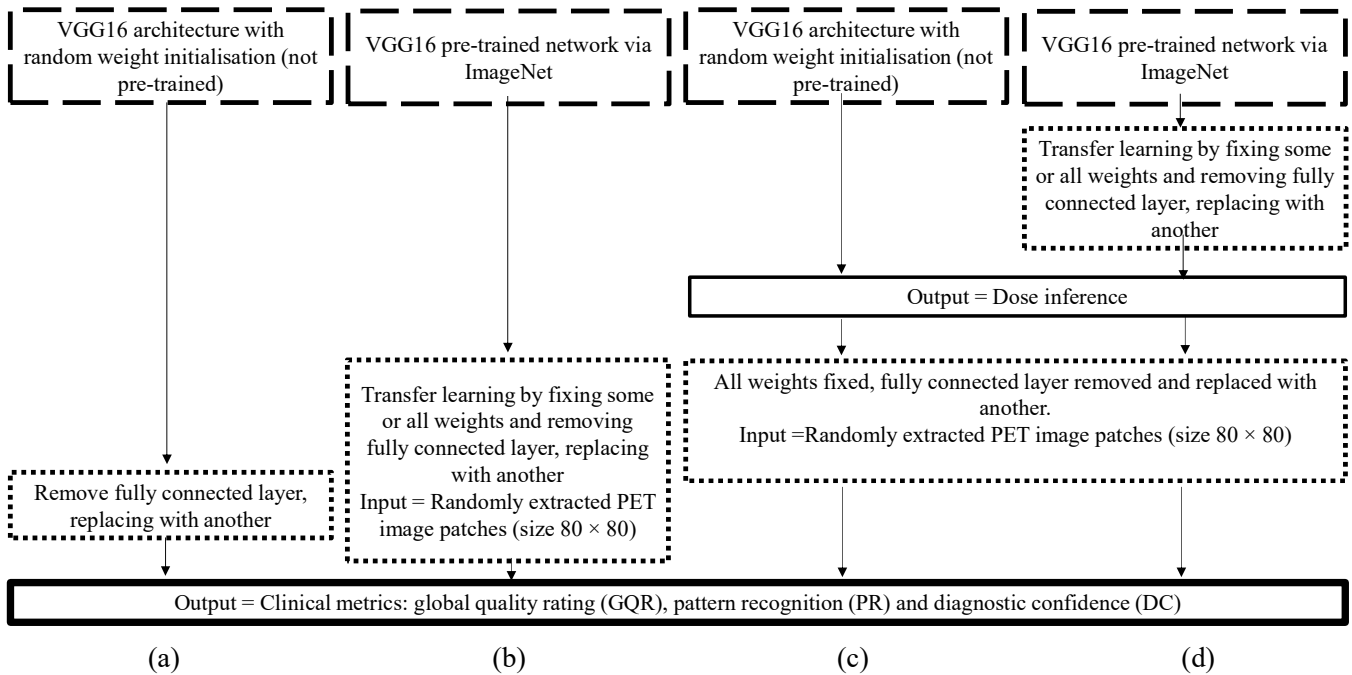


Fig. 7. Pipelines of pre-training. (a) VGG16 architecture without pre-training (random weight initialisation). (b) Transfer-learned VGG16 backbone pre-trained on ImageNet to predict the same three clinical metrics. (c) VGG16 backbone with random weight initialisation to initially infer dose as a pretext learning task, with weights from this model fixed and the fully connected layer removed and replaced with another to output the prediction of the three clinical quality metrics. (d) VGG16 backbone with ImageNet weights to initially infer dose as a pretext learning task, with weights from this model fixed and the fully connected layer removed and replaced with another to output the prediction of the three clinical quality metrics.

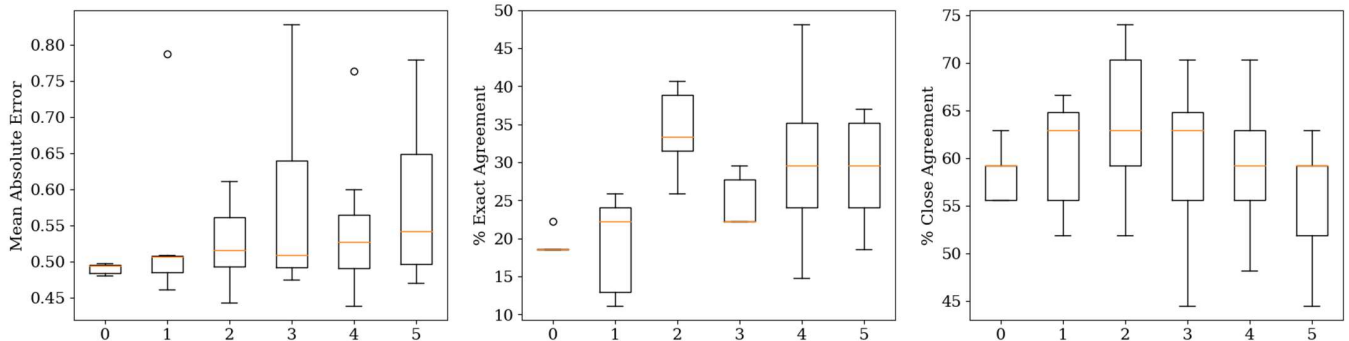


Fig. 8. Boxplots of MAE, percentage exact agreement and percentage close agreement as a function of the number of pre-fully connected layer convolutional blocks of the VGG16 backbone retrained (0 to 5). Left: 0 - only the fully connected layer trained; Right: 5 - all weights updated. Plotted across all clinical metrics for seven training runs. The middle line is the median.

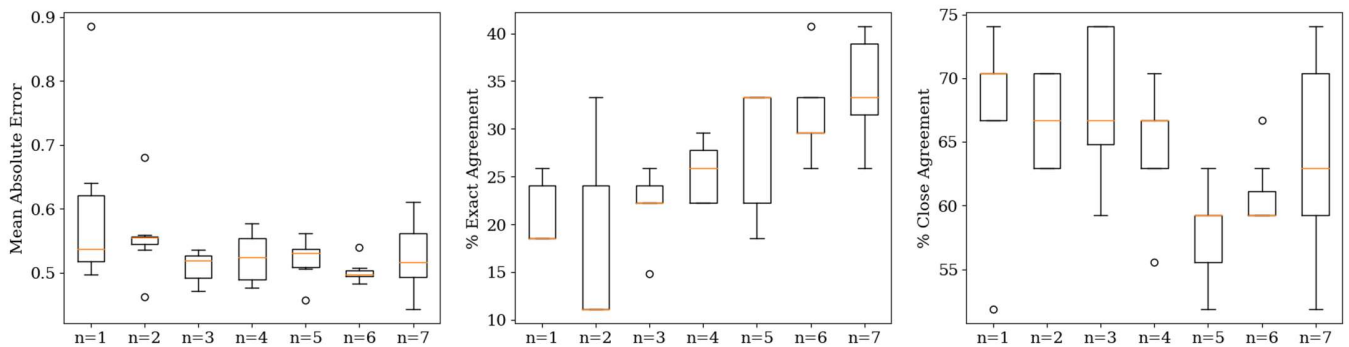


Fig. 9. Boxplots of the mean absolute error, percentage exact agreement and percentage close agreement as a function of the number of separate patients in the transfer learning datasets ($n=1$ to $n=7$). Plotted across all clinical metrics for seven training runs. The middle line is the median. The model used in for this experiment is that in Fig. 7b with the last 2 convolutional blocks trained.

TABLE III
COMPARISON TO EASILY ACCESSIBLE LABELS AS INPUT

Metric	Pre-trained Network			Dose/Weight to Clinical Scores (LR = 10 ⁻³)			Dose/Weight to Clinical Scores (LR = 10 ⁻⁴)		
	GQR	PR	DC	GQR	PR	DC	GQR	PR	DC
% Exact Agreement	56	22	11	0	0	0	0	0	0
% Close Agreement	89	44	33	89	56	33	78	56	33
MAE	0.20	0.64	0.71	0.29	0.65	0.89	0.28	0.64	0.88
Spearman's Rank	0.97	0.64	0.52	0.39	0.66	0.64	0.39	0.66	0.64

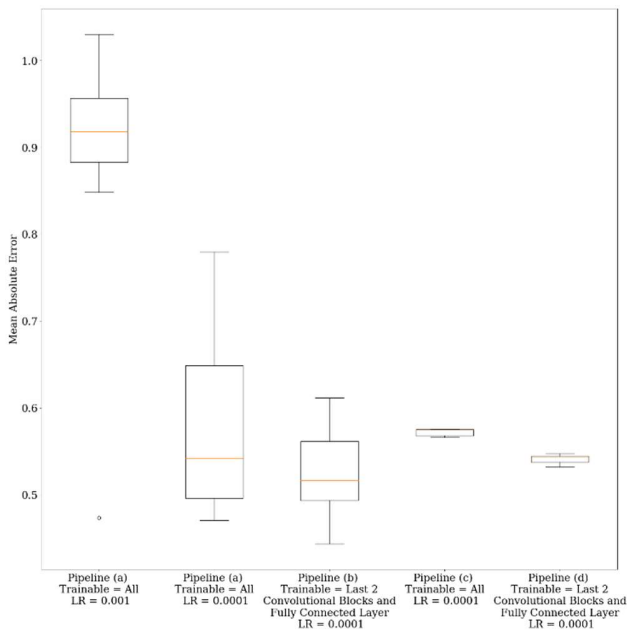


Fig. 10. Comparison boxplots of the MAE across all clinical metrics for each methodology over seven separate training runs. LR = Learning Rate. Pipeline (a) corresponds to training the VGG16 backbone with random weight initialisation; pipeline (b) involves training the VGG16 backbone with ImageNet weight initialisation, in this case, training last two convolutional blocks as the fully connected layer; pipeline (c) corresponds to using a pretext learning task of dose inference with random weight initialisation; and pipeline (d) is the dose inference pretext learning task with ImageNet weight initialisation.

the overall activity levels is eliminated, and ImageNet provides more complexity as a feature extractor, whereas when using the dose inference network, there is only one output from which the transfer-learned model is trying to learn. Whilst there was a comparable MAE across all three clinical metrics (Fig. 10) for all pipelines, it is arguable that transfer learning directly from a VGG16 backbone, as opposed to carrying out a pretext learning task, is more favourable, as additional patient datasets are not required.

The VGG16 backbone with random initialisation achieved the best performance for two of the 12 evaluation metrics (Table IV), but only narrowly so. However, when ImageNet weights

were added, this resulted in best performance for 7 out of 12 evaluation metrics, often by a large margin.

TABLE IV
COMPARISON OF CLINICAL METRICS FOR EACH PIPELINE

Model	Weight Initialisation	Trainable Layers	Mean Absolute Error \pm Standard Deviation			Percentage Exact Agreement \pm Standard Deviation			Percentage Close Agreement \pm Standard Deviation			Spearman's Rank Correlation Coefficient		
			GQR	PR	DC	GQR	PR	DC	GQR	PR	DC	GQR	PR	DC
			VGG16 Backbone (No pre-training, LR = 0.001)	Random	All	0.72 \pm 0.24	0.90 \pm 0.17	0.99 \pm 0.17	11 \pm 11	3 \pm 8	3 \pm 8	52 \pm 22	23 \pm 8	16 \pm 13
VGG16 Backbone (No pre-training, LR = 0.0001)	Random	All	0.51 \pm 0.42	0.61 \pm 0.07	0.63 \pm 0.05	32 \pm 18	30 \pm 12	25 \pm 11	75 \pm 22	44 \pm 6	51 \pm 6	0.60	0.62	0.57
VGG16 Backbone (Pre-training)	ImageNet	Last 2 Blocks and Fully Connected Layer	0.22 \pm 0.06	0.66 \pm 0.09	0.69 \pm 0.05	52 \pm 8	24 \pm 8	27 \pm 11	97 \pm 5	49 \pm 11	46 \pm 10	0.96	0.67	0.59
Pretext Learning	Random Dose Inference	All	0.36 \pm 0.01	0.64 \pm 0.01	0.72 \pm 0.01	11 \pm 2	4 \pm 10	13 \pm 5	72 \pm 10	64 \pm 5	33 \pm 0	0.93	0.54	0.59
Pretext Learning	ImageNet Dose Inference	Last 2 Blocks and Fully Connected Layer	0.32 \pm 0.02	0.58 \pm 0.02	0.68 \pm 0.05	13 \pm 4	31 \pm 5	11 \pm 0	71 \pm 11	44 \pm 0	44 \pm 0	0.93	0.59	0.56

Reducing the learning rate to 10⁻⁴ from 10⁻³ improved model performance across all evaluation metrics. This suggests that at the higher learning rate, training loss converges too early, such that when early stopping is employed, the model did not generalise well to unseen data. Using no pre-training with a reduced learning rate performed best for 2 out of 12 evaluation metrics (Table IV), namely for PR and DC. This may indicate that the VGG16 architecture was sufficient as a feature extractor for these metrics. However, the metrics' standard deviations overlapped between this model and using ImageNet for pre-training, and GQR prediction was inferior. The Spearman's rank correlation coefficients were also consistently highest for pre-trained networks, suggesting an overall better performance for predicting clinician scores.

Table IV supports Fig. 10 and shows that generally, GQR is more closely predicted to the true score than PR and DC. This may be because the global quality of an image is much easier to predict, as the model can estimate noise level. It is well-known that with a decreasing injected dose, the noise level in PET images increases, thus at lower doses, images become noisier [51]. This is also supported in Fig. 6, showing that the correlation between GQR and injected dose is stronger than for PR and DC. However, PR and DC, are more subjective metrics that involve the clinician using the images to make a diagnosis and judging their clinical plausibility based on the presence of specific pathological patterns, which is not provided as prior information to the network. In this case, the clinician estimated being able to make a reasonably confident diagnosis based on some of the noisier images. It has also been shown in Table III that it is not possible to only use injected dose standardised by patient weight as input into a machine learning network. Despite this model being much smaller, and using an easily available label, the best results were achieved when using PET

images in the transfer learning dataset, evident by the number of green boxes in the table. Whilst most prior research is on binary classification and not in PET imaging, our maximum Spearman's rank correlation coefficients of 0.96 for GQR, 0.67 for PR and 0.59 for DC are comparable to that achieved by Coyner *et al.* [43] (0.86 – 0.92).

Future work may include training and testing with a larger number of clinically-scored images, to observe the trend beyond seven separate transfer learning datasets. This work could also be applied to other low-dose reconstruction algorithms such as KEM [7] and MR-guided methods [52] or novel regularisation strategies [53]. As these methods are superior to standard OSEM, they should give higher quality predictions, even at lower simulated doses. However, similarly to [31], this may need to involve the inclusion of reconstructions generated via these algorithms into the training dataset. Further work could also investigate different established CNN backbones. Additionally, the model developed here may be used in different applications, including dynamic PET (where shorter frames are desirable for better time resolution, but also lead to noisier data) and different radioligands. Other recent advances, such as self-supervised learning for pre-training, as used by [54], [55], or vision-and-language pre-training for the similar task of determining clinical evaluations of medical images [56] could be investigated in relation to this clinical task.

V. CONCLUSION

This study shows that pre-training can help improve network performance in the task of automatically predicting clinical image quality assessments of PET images. By transfer-learning using ImageNet weight initialisation and retraining the last two convolutional blocks of a VGG16 backbone, a superior model performance is achieved for this clinical task. Additionally, this method does not require further independent patient datasets, unlike for using a pretext learning task. Overall, this work shows that pre-training on natural images for application to medical images has the potential to provide benefit for regressing from images to clinically-scored metrics.

ACKNOWLEDGMENT

The authors would like to thank Dr Veerle Kersemans for her guidance and support, Dr Zuzana Balazova for her work on data acquisition and processing, and Dr Abolfazl Mehranian for his work on low-count PET reconstruction. All authors declare that they have no known conflicts of interest in terms of competing financial interests or personal relationships that could have an influence or are relevant to the work reported in this paper.

REFERENCES

- [1] I. Nasrallah and J. Dubroff, "An overview of PET neuroimaging," *Semin. Nucl. Med.*, vol. 43, no. 6, pp. 449–461, 2013.
- [2] N. Galldiks, P. Lohmann, N. L. Albert, J. C. Tonn, and K. J. Langen, "Current status of PET imaging in neuro-oncology," *Neuro-Oncology Adv.*, vol. 1, no. 1, pp. 1–11, May 2019.
- [3] N. S. Kaltoft, L. Marnar, V. A. Larsen, S. G. Hasselbalch, I. Law, and O. M. Henriksen, "Hybrid FDG PET/MRI vs. FDG PET and CT in patients with suspected dementia – A comparison of diagnostic yield and propagated influence on clinical diagnosis and patient management," *PLoS One*, vol. 14, no. 5, p. e0216409, May 2019.
- [4] M. S. Judenhofer *et al.*, "Simultaneous PET-MRI: a new approach for functional and morphological imaging," *Nat. Med.*, vol. 14, no. 4, pp. 459–465, Apr. 2008.
- [5] L. Gaspar *et al.*, "Long-term quality assurance of [18F]-fluorodeoxyglucose (FDG) manufacturing," *Am. J. Nucl. Med. Mol. Imaging*, vol. 6, no. 3, p. 154, 2016.
- [6] S. Kaplan and Y. M. Zhu, "Full-Dose PET Image Estimation from Low-Dose PET Image Using Deep Learning: a Pilot Study," *Journal of Digital Imaging*, 2018.
- [7] J. Bland *et al.*, "MR-guided kernel em reconstruction for reduced dose PET imaging," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 2, no. 3, pp. 235–243, May 2018.
- [8] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights Imaging*, vol. 9, no. 4, pp. 611–629, 2018.
- [9] A. Kumar, J. Kim, D. Lyndon, M. Fulham, and D. Feng, "An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification," *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 1, pp. 31–40, Jan. 2017.
- [10] M. H. Yap *et al.*, "Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks," *IEEE J. Biomed. Heal. Informatics*, vol. 22, no. 4, pp. 1218–1226, 2018.
- [11] L. Gondara, "Medical Image Denoising Using Convolutional Denoising Autoencoders," *IEEE Int. Conf. Data Min. Work. ICDMW*, vol. 0, pp. 241–246, 2016.
- [12] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," *Proc. - 2016 4th Int. Conf. 3D Vision, 3DV 2016*, pp. 565–571, 2016.
- [13] B. Kayalibay, G. Jensen, and P. van der Smagt, "CNN-based Segmentation of Medical Imaging Data," *arXiv*, 2017.
- [14] P. C. Tripathi and S. Bag, "CNN-DMRI: A Convolutional Neural Network for Denoising of Magnetic Resonance Images R," *Pattern Recognit. Lett.*, vol. 135, pp. 57–63, 2020.
- [15] D. Truhn, S. Schrading, C. Haarbuerger, H. Schneider, D. Merhof, and C. Kuhl, "Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI," *Radiology*, vol. 290, no. 3, pp. 290–297, 2019.
- [16] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J. Big Data 2021 81*, vol. 8, no. 1, pp. 1–74, Mar. 2021.
- [17] M. J. Willeminck *et al.*, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, Feb. 2020.
- [18] F. Chollet, "Keras." 2015.
- [19] B. Mustafa *et al.*, "Supervised Transfer Learning at Scale for Medical Imaging," 2021.
- [20] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow Twins: Self-Supervised Learning via Redundancy Reduction," 2021.
- [21] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.
- [22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, 2017. [Online]. Available: <https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html>. [Accessed: 14-Apr-2022].
- [23] Q. Guan *et al.*, "Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study," *J. Cancer*, vol. 10, no. 20, p. 4876, 2019.
- [24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.
- [25] J. A. Dunnmon, D. Yi, C. P. Langlotz, C. Ré, D. L. Rubin, and M. P. Lungren, "Assessment of convolutional neural networks for automated classification of chest radiographs," *Radiology*, vol. 290, no. 3, pp. 537–544, 2019.
- [26] Y. X. Tang *et al.*, "Automated abnormality classification of chest radiographs using deep convolutional neural networks," *npj Digit. Med.*, vol. 3, no. 1, 2020.
- [27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 2818–2826, Dec. 2015.

- [28] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *J. Big Data*, vol. 6, no. 1, 2019.
- [29] B. Neyshabur, H. Sedghi, G. Brain, and C. Zhang, "What is being transferred in transfer learning?"
- [30] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, 2019.
- [31] Y. Wen, L. Chen, Y. Deng, and C. Zhou, "Rethinking pre-training on medical imaging," *J. Vis. Commun. Image Represent.*, vol. 78, p. 103145, Jul. 2021.
- [32] M. Heker and H. Greenspan, "Joint Liver Lesion Segmentation and Classification via Transfer Learning," no. 1, pp. 1–5, 2020.
- [33] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K Pretraining for the Masses," pp. 1–20, 2021.
- [34] N. Tajbakhsh *et al.*, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.
- [35] S. Naz, A. Ashraf, and Ahmad Zaib, "Transfer learning using freeze features for Alzheimer neurological disorder detection using ADNI dataset," *Multimed. Syst.*, vol. 28, pp. 85–94, 2022.
- [36] B. Kieffer, M. Babaie, S. Kalra, and H. R. Tizhoosh, "Convolutional neural networks for histopathology image classification: Training vs. Using pre-trained networks," *Proc. 7th Int. Conf. Image Process. Theory, Tools Appl. IPTA 2017*, vol. 2018-Janua, pp. 1–6, 2018.
- [37] S. Bosse, D. Maniry, K. R. Müller, T. Wiegand, and W. Samek, "Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment," 2018.
- [38] X. Yao, Q. Cao, X. Feng, and G. Cheng, "Learning to Assess Image Quality Like an Observer," pp. 1–13, 2022.
- [39] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," 2014.
- [40] S. J. Sujit, I. Coronado, A. Kamali, P. A. Narayana, and R. E. Gabr, "Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks," *J. Magn. Reson. Imaging*, vol. 50, no. 4, pp. 1260–1267, 2019.
- [41] I. Oksuz *et al.*, "Automatic CNN-based detection of cardiac MR motion artefacts using k-space data augmentation and curriculum learning," *Med. Image Anal.*, vol. 55, pp. 136–147, Jul. 2019.
- [42] J. J. Ma *et al.*, "Diagnostic Image Quality Assessment and Classification in Medical Imaging: Opportunities and Challenges," *Proc. - Int. Symp. Biomed. Imaging*, vol. 2020-April, no. 1, pp. 337–340, 2020.
- [43] A. S. Coyner *et al.*, "Automated Fundus Image Quality Assessment in Retinopathy of Prematurity Using Deep Convolutional Neural Networks," *Ophthalmol. Retin.*, vol. 3, no. 5, pp. 444–450, May 2019.
- [44] G. T. Zago, R. V. Andreão, B. Dorizzi, and E. O. Teatini Salles, "Retinal image quality assessment using deep learning," *Comput. Biol. Med.*, vol. 103, pp. 64–70, Dec. 2018.
- [45] J. B. Hopson *et al.*, "Pre-training and Transfer Learning for Training Set Reduction and Improving Automated Assessments of Clinical PET Image Quality," in *2021 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2021.
- [46] P. J. Markiewicz *et al.*, "Rapid processing of PET list-mode data for efficient uncertainty estimation and data analysis," *Phys. Med. Biol.*, vol. 61, no. 13, pp. N322–N336, 2016.
- [47] J. Ashburner *et al.*, "SPM8 Manual," 2013. [Online]. Available: https://www.fil.ion.ucl.ac.uk/spm/doc/spm8_manual.pdf. [Accessed: 11-Apr-2022].
- [48] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," 2016.
- [49] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?"
- [50] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, May 2017.
- [51] C. N. Ladefoged, P. Hasbak, C. Hornes, L. Højgaard, and F. L. Andersen, "Low-dose PET image noise reduction using deep learning: Application to cardiac viability FDG imaging in patients with ischemic heart disease," *Phys. Med. Biol.*, vol. 66, no. 5, 2021.
- [52] A. Mehranian *et al.*, "PET image reconstruction using multi-parametric anato-functional priors," *Phys. Med. Biol.*, vol. 62, no. 15, pp. 5975–6007, 2017.
- [53] A. J. Reader and S. Ellis, "Bootstrap-Optimised Regularised Image Reconstruction for Emission Tomography," *IEEE Trans. Med. Imaging*, vol. 39, no. 6, pp. 2163–2175, Jun. 2020.
- [54] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," pp. 15979–15988, 2022.
- [55] C. Niu and G. Wang, "Unsupervised contrastive learning based transformer for lung nodule detection," *Phys. Med. Biol.*, 2022.
- [56] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi, "Multi-modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training," *IEEE J. Biomed. Heal. Informatics*, vol. XX, no. Xx, pp. 1–10, 2022.